大连理工大学
DALIAN UNIVERSITY OF TECHNOLOGY

硕士学位论文
MASTERAL DISSERTATION

特定领域中文术语抽取

学科专业 计算机应用技术

作者姓名 李 丹

指导教师 李丽双 副教授

答辩日期 2011 年 12 月

硕 士 学 位 论 文

特定领域中文术语抽取

Chinese Term Extraction in Specific Domain

作 者 姓 名：李丹

学科、专业：计算机应用技术

学　　号：20909306

指 导 教 师：李丽双副教授

完 成 日 期：2011-11-09

大连理工大学

Dalian University of Technology

大连理工大学学位论文独创性声明

作者郑重声明：所呈交的学位论文，是本人在导师的指导下进行研究工作所取得的成果。尽我所知，除文中已经注明引用内容和致谢的地方外，本论文不包含其他个人或集体已经发表的研究成果，也不包含其他已申请学位或其他用途使用过的成果。与我一同工作的同志对本研究所做的贡献均已在论文中做了明确的说明并表示了谢意。

若有不实之处，本人愿意承担相关法律责任。

学位论文题目：特定领域中文术语抽取

作者签名：李丹　　日期：2011 年 12 月 15 日

摘　要

术语是代表特定学科领域基本概念的语言单元，体现了该领域的核心知识，方便人们快速的获取专业知识。随着技术的进步，各种新知识不断涌现，相应的术语也在不断发展，由于信息爆炸时代大量的数字信息资料产生，传统的由人力获取术语的方法变的不再可行。如何自动获取术语自然成为人们研究的热点。术语自动抽取是信息处理领域中一项重要的研究任务，在词典编撰、领域本体构建、机器翻译等领域都有重要的应用。

目前常用的术语抽取方法有基于规则的方法、基于统计的方法、统计和规则结合的方法。其中基于统计的方法又可以根据有无已标注的语料分为有监督的统计机器学习方法和无监督的基于统计量的方法。由于缺少已标注语料，前人对基于统计机器学习的术语抽取方法的研究不多，本文研究特定领域的术语抽取方法，分析了领域术语的特点，比较了其和命名实体的区别。针对汽车领域制定了标注规则，对语料进行标注。采用基于条件随机场的机器学习方法对领域术语进行抽取，得到精确率、召回率、F-值分别为86.41%，80.50%，82.50%。

针对人工标注领域语料代价大的情况，本文将主动学习策略引入基于条件随机场的术语抽取方法中。使用主动学习的不确定性样本选择策略，结合 CRFs 模型给出的边缘概率计算置信度，实验结果证明使用主动学习的方法选择样本比随机选择样本规模所得到的结果要好，使用较少的已标语料即可获得预期的抽取效果。

基于有监督统计机器学习的方法可以获得较好的结果，但其对于已标语料的规模和质量都有不小的依赖性，本文研究了无监督的基于统计量的领域术语抽取方法。分别分析了信息熵、互信息、C-value 对不同长度的领域术语抽取上的性能，对于由 1~3 个词语组成的术语使用词性组成规则进行过滤，提高了术语抽取的精确率，最终的 F-值为15.41%。

本文研究了特定领域术语抽取的方法，基于统计量的方法使用的资源和代价最小，但结果最差。基于条件随机场的方法的最终结果最好，主动学习方法结果与其相比相差不多，但是使用了较少的训练语料。

关键词：领域术语；术语抽取；条件随机场；主动学习；统计量

Chinese Term Extraction in Specific Domain

Abstract

Term is a language unit that represents the basic concepts of specific subject areas, reflecting the core knowledge in the field, convenient for people to get professional knowledge rapidly. With the development of technology, all kinds of new knowledge constantly emerging, the corresponding term is also in the continuous development, as a lot of digital information material having been produced in the era of information explosion, the traditional method of accessing terms by human becomes no longer feasible. How to automatically get term becomes a hot research naturally. Automatic terminology extraction is one of important research tasks in the information processing field and has important applications in fields of the dictionary compilation, domain ontology construction, machine translation, etc.

Current term extraction methods are commonly used rule-based method, based on statistical methods, the method of combining statistical and rule. Based on statistical methods, according to having the labeled corpus or not, can be divided into supervised statistical machine learning methods and unsupervised methods based on statistics. Due to the lack of labeled corpus, predecessors did a little research on term extraction method that basing on the statistical machine learning. In this paper, we study the term extraction method in specific areas, analyzing the characteristics of domain terms, comparing the difference between its and the named entity. For the automotive sector has developed labeling rules, tagging of corpus. The precision, recall and F-measure of term extraction based on CRFs are 86.41%, 80.50%, 82.50% respectively.

Against its trouble to label domain corpus by manual, this article will introduce active learning strategies into term extraction methods based on Conditional Random Field. Use the uncertainty sample selection strategy of active learning, combining with the conditional probability calculated confidence that come from CRFs module, Experimental results show that results obtained from using of active learning methods to increase sample are better than increasing sample size randomly, with less tagged corpus to get the desired effect.

Based on supervised statistical machine learning methods can obtain better results, but has a lot of dependence on the scale and quality of the tagged corpus, this paper studies unsupervised statistics based domain term extraction method. The paper analyzes the performance of information entropy, mutual information, C-value in domain term extraction,

combined with the formed part of speech rules of terms to filter and improve the accuracy of term extraction. The final F-score is 15.41%.

This paper gives the approaches of domain term extraction. The statistic method needs the minimum resource but its result is not good. The CRFs method achieves the best result, while method based on active learning and CRFs gets the similar result but needs less tagged corpus compared with approach not use active learning.

Key Words: Domain Term; Term Extraction; Conditional Random Fields; Active Learning; Statistic method

目　录

摘　要 I
Abstract II
1　绪论 1
1.1　研究背景与意义 1
1.2　研究现状 1
1.3　本文主要研究内容和组织结构 3
2　术语的定义及术语抽取的特点 5
2.1　术语的定义 5
2.2　术语的特点 5
2.3　汽车领域术语 6
2.4　汽车领域术语抽取的特点 7
3　基于条件随机场的术语抽取方法 8
3.1　条件随机场的基本原理 8
3.1.1　CRFs 的图结构 8
3.1.2　CRFs 的势函数 8
3.1.3　CRFs 特征函数生成 9
3.2　语料预处理 11
3.3　有效特征选取 11
3.4　实验与分析 15
3.4.1　实验数据 16
3.4.2　实验结果与分析 16
3.5　本章小结 18
4　结合条件随机场与主动学习策略的术语抽取 19
4.1　主动学习 19
4.2　主动学习与 CRFs 结合 20
4.3　实验与结果分析 22
4.4　本章小结及改进方向 23
5　基于统计量的术语抽取方法 24
5.1　统计模型 24
5.2　统计量 26
5.3　基于统计量的术语抽取实验 29

5.3.1 实验数据预处理 29
5.3.2 基于统计量方法的术语抽取 30
5.3.3 统计量与规则相结合 34
5.4 三类术语抽取方法的分析 37
5.5 本章小结 39
结 论 40
参 考 文 献 42
攻读硕士学位期间发表学术论文情况 44
致 谢 45
大连理工大学学位论文版权使用授权书 46

1 绪论

1.1 研究背景与意义

科技不断发展，社会不断进步，人们的要求也在提高，希望高效率的获取高质量的信息来充实自身。然而在信息化社会，新知识、新技术通过各种途径以人们始料未及的速度和规模迅速传播发展，直接表现就是海量的涉及各个领域的知识以电子形式存储在互联网上，如何快速高效便捷的获取领域知识就成为了亟待解决的一个问题。学科之间不是独立的，经常会用到互相的理论或知识，这要求人们能够快速的掌握一个领域的核心知识。学科的基本知识通常是通过概念解释来传播的，概念则以一个更加精炼的形式表述，也就是术语。即领域术语代表了特定学科的核心知识，术语的变化在一定程度上反映了该学科的发展变化。掌握领域术语对于快速理解整个学科有很重要的意义。传统的方式是由领域专家以人工的方式制定出相应的术语库供人们使用，但是在海量数据面前人工方式显得无能为力，而且技术不断发展，术语也在不断更新，手工修订术语库已不现实，如何利用计算机自动获取术语自然成为了人们研究的热点。

目前术语自动抽取已是信息处理领域中一项重要的研究任务，在词典编撰、领域本体构建[1]、文本分类[2]等领域都有重要的应用。

1.2 研究现状

目前已有的术语自动抽取方法主要有基于语言学规则的方法，基于统计的方法，规则与统计相结合的方法。

（1） 规则方法

基于规则的方法主要是根据语言学知识及相应的领域知识制定规则模板，与规则模板匹配的视为术语。文献[3]将科技领域的术语限定为名词短语，利用词性模式来获得候选词串。文献[4]利用语法分析，通过语法结构等信息来判定名词短语是否是术语。基于规则的方法是专门针对相应领域的术语构词特点从文本中获取术语，可以得到较高质量的术语串，但其受限于所利用的语言学知识的质量，人工制定的规则通常不能涵盖全部的术语组成特点，导致术语识别不全面，而且如果规则越多造成冲突的可能性越大。

（2） 统计方法

基于统计的方法是建立在大规模语料库的基础上的，根据有无已标注的语料又可分为无监督的统计方法和有监督统计机器学习方法。

无监督的基于统计学的方法是通过计算字符串的某种或某些概率意义上的统计量度来判断其是否是领域术语。总结前人的研究，主要用到的统计量有频率、假设检验、

似然比、信息熵、互信息、C/NC-value、TF-IDF 等。对于中文术语抽来说，由于汉字之间没有没类似于西文那样的天然分隔符，如何获得有意义的字符串成为一个难点，所以目前的研究者对于术语单元性的研究相对较多，文献[5]利用互信息衡量词串内部的结合强度进行术语抽取。何婷婷[6]等人提出了质子串分解的方法，使用改进的互信息参数 F-MI 进行术语抽取，文献[7]在的基于质子串分解方法的基础上加上了卡方检验参数提高了复杂串的抽取精度。在获取了符合语言学规则的正确词语后，衡量候选词的术语性来确定是否是正确的术语。TF-IDF 及其变形是比较常用的衡量术语性的统计量，它是利用术语在领域内流通性高而在其它领域的使用很少或没有这个特点，能过滤掉高频的通用领域词汇。文献[8]指出 TF-IDF 的方法比较依赖于使用的背景语料，如果两个领域有一定交叉会削弱术语识别能力，提出了一种基于词频分布变化程度的术语性计算方法。基于统计的方法不需要过多的领域知识，且一般不受语种的限制，可移植性较好，但是此方法依赖于语料库的规模，如果语料库过小则不足以获取较准确的统计信息，通常会得到领域无关的词串，并且会过滤掉一些低频术语。

由于缺少已标注的语料库，所以使用有监督的机器学习方法进行术语抽取的研究比较少。有监督机器学习的方法是从已标数据中学习知识，训练一个模型，然后用这个模型去预测未知数据的标签。文献[9]利用隐马尔科夫模型对计算机术语进行识别。文献[10]到文献[12]均使用条件随机场进行术语抽取，不同之处在于选取的特征不同，文献[10]中利用词及词性作为特征，文献[11]除了使用词、词性作为特征外还考虑了互信息、信息熵和 TF-IDF 作为特征，章承志[12]使用一体化的术语抽取方法，将语言学特征和统计特征融合到一起，而且考虑了词语所在句子的术语度，其实验证明了所选特征的是有效的。

有监督的机器学习方法的结果比较好，但是需要一定规模的人工标注语料，而且对于术语抽取任务来说语料的标注是需要领域知识的，费时费力。有研究者在命名实体识别任务上引入了主动学习的策略，使用较少规模的训练语料即可达到预期的结果，减少了人工标注语料的代价。

（3） 规则和统计结合的方法

单纯的制定规则受到多方面的限制，而基于统计的方法又需要以大规模的语料为支撑，而且对于语料的质量要求也比较高，对于数据稀疏严重的语料来说，统计方法很难取得比较理想的结果。所以目前的研究基本上是综合利用规则和统计两种方法的优点，二者结合来进行术语抽取。一般将术语抽取任务分为两个步骤，首先进行候选术语的抽取，再对候选串进行过滤得到最终的术语列表。统计方法和规则的方法没有固定的使用顺序，有些研究者首先使用语言规则得到候选词串，然后再利用统计学方法进行过滤。

文献[13]使用三条词性组合模式(Noun + Noun),(Adj|Noun) + Noun 和(Adj|Noun)+|((Adj|Noun)*(NounPrep)?)(Adj|Noun)*来抽取候选术语，然后结合使用T检验和C/NC-value得到最终的术语。周浪等人[14]首先分析了已有的计算机领域术语库的特点，总结出计算机领域术语在长度、词法模式等方面的规则，使用规则抽取候选术语，使用子串归并、搭配检验和TF-IDF来进行过滤。也有人先利用统计学方法获取候选术语，再将不符合语言学规则或术语组成特点的词剔除。文献[15]使用pattree索引结构，利用SEF和C-value两种统计量获取候选术语，然后利用候选词及其窗口内上下文在词典中的信息和词法模式进行过滤，得到最终的术语。

1.3 本文主要研究内容和组织结构

本文主要是针对特定领域的中文术语抽取的相关技术进行研究，包括以下几个方面：

（1） 本文选取汽车领域作为研究对象，分析研究了汽车术语的组成特点，制定了相应的标注准则，对从网页上获取的领域语料进行预处理后进行人工标注，将术语抽取任务转化为序列标注问题，采用条件随机场作为机器学习模型，采用语言学特征和统计特征，分析了各个特征对识别结果的所起的作用。

（2） 将主动学习策略和条件随机场模型结合引入到术语抽取任务中，使用不确定性样本选择策略，利用CRFs给出的条件概率计算置信度，和随机选择样本的方法进行了对比，实验结果证明主动学习结合CRFs模型的方法是有效的。

（3） 研究了无监督的统计方法，分析各种统计量在领域抽取上的作用，结合词性组成规则进行候选串过滤，得到最终的术语列表。

本文的组织结构如下：

第一章，介绍了术语抽取的研究背景和意义，总结分析了目前术语抽取的研究现状以及本文的主要工作。

第二章，主要介绍术语的定义及其特征，并针对汽车领域术语的特点进行了分析，制定了相应的标注标准，分析了领域术语和命名实体识别的不同点。

第三章，采用基于条件随机场的汽车领域术语抽取方法，研究了语言学特征和统计特征对抽取结果的影响。

第四章，将条件随机场与主动学习策略相结合，利用CRFs给出的边缘概率计算置信度，使用最不确定样本选择的主动学习样本选择策略，分析了使用主动学习方法和不采用主动学习方法的对比实验。

第五章，研究了无监督的统计方法，分析信息熵、互信息、$C-value$各自对不同长度的术语抽取性能上的影响，在得到候选串后经过领域相减去除通用领域词汇，然后使用词性组合规则进行过滤，得到最终的术语列表。

2 术语的定义及术语抽取的特点

2.1 术语的定义

术语是代表特定学科领域基本概念的语言单元，可以是词也可以是词组。在我国又称为名词或科技名词。目前没有明确的关于“术语”的定义，我们引用冯志伟在其《现代术语学引论》[16]中的定义为，术语是“通过语音或文字来表达或限定专业概念的约定性符号，可以是词也可以是词组”。

2.2 术语的特点

术语的定义只是给出了一个抽象的表述，进行术语抽取工作，要首先弄清楚术语的特点。冯志伟在《现代术语学引论》[16]中总结了术语的8种构成原则，即术语的特点：准确性，即术语要能够确切的反映概念的本质特征，比如如“磁带”就表示一种带状录音或录像的载体，表述很形象；单义性，是指至少在一个领域中，一个术语只表示一个概念，不能一词多指。这一点在不同的领域内可以不成立，如“门”，在不同的领域代表不同的含义，可以表示普通意义上的建筑物上的出入口装置，在生物学中则表示分类类群中的一个等级，现在也有用来描述一个爆炸性的事件，比如“水门事件”；系统性，一个特定领域的所有术语，必须处在一个明确的层次结构之中，共同构成一个系统；语言的正确性，即术语的结构要符合该语种的构词规则和词组构成规则。简明性，术语要简明扼要，易读易记，不易过度冗长；理据性，术语的学术含义不能违反术语的结构所表现出来的理据，应该尽量做到“望文生义”；稳定性，某个学科的术语一经确定，除非特别需要，不宜轻易改动；能产性，术语确定之后，还可以由旧术语出发，通过构词法或词组构成的方法，派生出新的术语[16]。

对于术语抽取来讲，我们主要从术语的准确性、正确性等方面进行分析，也可以说是从术语的组成结构特性和领域特性上分析。

从术语的组成结构来看，可以分为简单术语和复合术语。简单术语即是单词型术语，其是术语组成的基本单元，不能再被分割为更小的术语。比如“悬架”，单独拆为“悬”或“架”都不能组成术语。复合术语是由简单术语和简单术语或简单术语和非术语按照一定的语法规则组成的，代表不同意义的词或短语。比如“电控悬架”、“主动悬架”，是由非术语“电控”、“主动”分别和简单术语“悬架”组成的词，表示不同种类的悬架，再比如表示发动机种类的“汽油发动机”就是由汽车的燃料术语“汽油”和表示部件的术语“发动机”组合而成。

术语的领域特性是指术语表述的是某个特定学科的概念，即只在该学科或该学科的交叉学科中流通，在通用领域或其它专业领域不流通或很少被使用。

很多研究者将术语抽取的工作重点放在两方面研究，一是单元性（Unithood），一是术语性（Termhood）。单元性即是指一个术语必须是能够作为一个具有完整意义的正确字串，用单元性可以衡量字串内部结合的紧密程度。术语性也可以称为领域性，表明一个术语代表其特定领域的程度。

2.3 汽车领域术语

在本文中我们针对汽车领域进行术语抽取，使用有监督的机器学习方法必须有已标注的样本用以训练学习器，目前没有可供使用的已标语料，故需要先进行人工标注。目前缺少一个关于汽车领域术语的统一标准，我们对《汽车行业名词术语汇编》[17]中的和汽车零部件相关的7525个术语进行了学习和分析，统计得到单词型术语占9%，由两个单词组成的复杂术语占35%，三词术语占31%，四、五、六词术语分别占15%、6%、2%，七词及以上术语占2%，即复杂术语一般由2~4个单词组成，占全部术语的81%，符合中文术语的一般性特点。为了方便人工标注，我们分析了汽车领域术语的特点并借助前人对领域术语特点的研究成果，制定了一定的标注标准，凡是符合标注标准的词都被视为汽车领域的术语。标注标准如下：

（1） 描述或表示汽车的词，一般是随着汽车领域的产生和发展而出现的，比如“轿车”、“两厢车”等，由于汽车领域外来词汇比较多，通常情况下人们会用外文直接描述，像类似于 “SUV”（运动型多用途汽车）、“RV”（休闲车）等英文单词或缩略词也归于汽车领域术语。

（2） 表示汽车零部件或组成成分的词，如“底盘”、“后视镜”，另外像“气门”、“活塞”等机械领域的词，虽然不是专属于汽车领域的，但是也是描述汽车结构或功能所必需的，视为领域术语。

（3） 与汽车相关的系统或结构，如“防抱死制动系统”、“高压共轨系统”等，相应的英文缩略词同样作为术语。

（4） 一些词在通用领域也用，但是在汽车领域表示特定的含义，如“抬头”、“塌屁股”描述的是汽车的某种状态，可作为汽车术语。

（5） 要遵循术语要尽可能的详细和完整的原则，如类似“1.6升5缸发动机”、“四行程发动机缸内燃油直喷技术”，要将其作为一个整体。

（6） 描述汽车品牌及其型号的词语在本文中不作为领域术语，可单独作为一类词进行识别。

（7） 文章中若出现英文缩写和中文译文联合使用的情况，按两个术语分别标注。如“ABS（防抱死制动系统）”，标注为“ABS”和“防抱死制动系统”两个术语。

2.4 汽车领域术语抽取的特点

通过对汽车领域术语特点的分析可以看出领域术语在结构上比较复杂，所以与一般的命名实体识别相比，领域术语的自动抽取具有其特殊性，具体表现在：

（1） 没有明确的关于领域术语的定义，不能清晰地界定术语的边界。目前已有的词典或是词表不足以涵盖全部的术语，而且随着技术的进步，新的产品或应用会不断增多，相应的术语表示也会不断丰富。比如“绿色汽车”、“零公里”是近几年提出的概念。

（2） 由于汽车领域引入国外技术比较多，在表述时多采用音译词或是英文缩写，比如“皮卡”（”pick-up”的音译）、“RV”（休闲车），而且由于使用习惯等原因，在表述时使用的不同的名称代表统一事物，比如“皮卡”和“轿卡”就代表同一类型汽车，在使用时比较随意，没有特定的用法。

（3） 汽车领域的术语模式多变，表现在长度、词性、组成模式等方面。例如，“悬架”和“综合电子控制动力转向系统”相差10个字长，还有类似于“可变预行程tics系统”和“D2T式制动器”的中英文混合术语。

（4） 一般的命名实体（人名、地名或组织机构名等）通常会存在比较明显的特征词，上下文环境也相对规律，而就汽车领域术语而言很难找出比较统一的特点，而且中英文混用的现象明显。

（5） 领域术语的一个公共特点就是存在嵌套（网状术语），比如“曲轴箱换气式二行程汽油机”，其中“曲轴箱”、“二行程发动机”、“发动机”又都分别作为术语出现。

3 基于条件随机场的术语抽取方法

条件随机场（Conditional Random Fields, CRFs）是一种判别式图模型，由 Lafferty[18] 等人于 2001 年提出。CRFs 同时具备最大熵模型（ME）和隐马尔科夫模型（HMM）的特点，并且不存在 HMM 那样严格的独立性假设，而且其采用的是全局归一化的方法，克服了最大熵马尔可夫模型的标记偏置问题，是目前处理序列标注问题最好的统计机器学习模型，在分词、命名实体识别等问题上已经得到广泛的应用。虽然领域术语和一般的命名实体在自身结构、所运用的环境等方面有很大的不同，但是就其识别任务而言也有一定的相似性，故我们将领域术语的抽取任务转化为序列标注问题，利用 CRFs 进行汽车领域术语的抽取。

3.1 条件随机场的基本原理

3.1.1 CRFs 的图结构

条件随机场（CRFs）是一种无向图模型，如果给定了待标记的观察序列，则其可以被用来在标记序列上定义一个联合概率分布。设 X 和 Y 都是联合分布随机变量，分别表示待标记的观察序列和待标序列对应的标记序列，那么(X, Y)就是一个以观察序列 X 为全局条件的无向图模型。V 中的每一个节点都有一个随机变量 Y_v 与其对应。定义 $G=(V,E)$ 是一个无向图，V 表示结点集合，E 表示各个结点之间的无向边。$Y=\{Y_v \mid v \in V\}$，即表示 V 中的每个结点对应着一个随机变量中的元素 Y_v。如果每个随机变量 Y 满足关于 G 的马尔可夫属性，给定任何观察序列 X 和除去 Y_v 以外的标记序列 $Y_u(u \neq v, u \in V, v \in V)$，$Y_v$ 的概率可以用公式（3.1）所示的联合概率表示：

$$P(Y_v \mid X, Y_u, u \neq v) = P(Y_v \mid X, Y_u, u \sim v) \qquad (3.1)$$

其中 $u \sim v$ 表示结点 u 和结点 v 在无向图中相邻。

常用的 CRFs 无向图结构是一阶链式的，称为线性链式 CRFs（Linear-chain CRFs）。输入的观察序列的各节点间没有边，即不存在任何独立性假设。只将观察序列作为前提条件，选择联合条件概率最大的序列作为其相应的标记序列。

3.1.2 CRFs 的势函数

在 CRFs 模型中，对于给定的观察序列 X，其对应的标记序列 Y 的概率是势函数（Potential Function）乘积的一个归一化形式，势函数是来自于条件独立的概念，公式（3.2）给出了它的表示方式：

$$\exp(\sum_j \lambda_j t_j(Y_{i-1},Y_i,X,i)+\sum_k \mu_k s_k(Y_i,X,i)) \tag{3.2}$$

其中$t_j(Y_{i-1},Y_i,X,i)$表示观察序列X中位置i-1和位置i的元素对应的标记结果的转移特征函数，$s_k(Y_i,X,i)$表示观察序列X中位置i的元素对应的标记序列和观察序列X的状态特征函数，λ_j和μ_k表示特征权重，可以从已知标记的训练语料中估计得到。

我们将转移特征函数和状态特征函数统一表示成为$f_j(Y_{i-1},Y_i,X,i)$，即给定观察序列X，其对应的标记序列Y的概率可以形式化表示为公式（3.3）：

$$P(Y|X,\lambda)=\frac{1}{Z(X)}\exp(\sum_{i=1}^{n}\sum_j \lambda_j f_j(Y_{i-1},Y_i,X,i)) \tag{3.3}$$

其中$Z(X)$是归一化因子，与Y无关：

$$Z(X)=\sum_Y \exp(\sum_{i=1}^{n}\sum_j \lambda_j f_j(Y_{i-1},Y_i,X,i)) \tag{3.4}$$

CRFs由训练语料训练得到$P(Y|X)$统计模型进行参数估计，对应的解码过程则是求解Y^*使得$P(Y|X,\lambda)$最大，求解过程如公式（3.5）：

$$\begin{aligned} Y^* &= \arg\max_Y P(Y|X,\lambda)=\arg\max_Y \frac{1}{Z(X)}\exp(\sum_{i=1}^{n}\sum_j \lambda_j f_j(Y_{i-1},Y_i,X,i)) \\ &= \arg\max_Y \sum_{i=1}^{n}\sum_j \lambda_j f_j(Y_{i-1},Y_i,X,i) \end{aligned} \tag{3.5}$$

可使用Viterbi动态优化算法求出最优解Y^*。

3.1.3 CRFs特征函数生成

以一个例子来说明CRFs生成特征函数，进行参数估计的过程。给定训练样本中的一个观察序列X={安装/齿轮/变速器}，采用基于词的方法，其相应的标签序列为Y={OBI}，假设当前读入X序列的第二个位置，即i=1（从i=0记），X_i=齿轮，Y_i=B，此时系统生成相应的状态特征函数$s_k(Y_i,X,i)$和状态转移函数$t_j(Y_{i-1},Y_i,X,i)$，若采用BIO标记法的话，Y_i和Y_{i-1}都有三种状态$\{B,I,O\}$，则$Y_{i-1}Y_i$会有九种情况的组合$\{BB,BI,BO,IB,II,IO,OB,OI,OO\}$，此时在$i=1$的位置上生成了3个状态特征函数和9个转移特征函数，都是二值函数，分别为：

$$s_1(Y_i,X,i)=\begin{cases}1 & X_i=\text{齿轮，}\ Y_i=B\\ 0 & \text{其它}\end{cases}$$

$$s_2(Y_i,X,i)=\begin{cases}1 & X_i=\text{齿轮，}\ Y_i=I\\ 0 & \text{其它}\end{cases}$$

$$s_3(Y_i,X,i)=\begin{cases}1 & X_i=\text{齿轮，}\ Y_i=O\\ 0 & \text{其它}\end{cases}$$

$$\mathrm{t}_1(Y_{i-1},Y_i,X,i)=\begin{cases}1 & X_i=\text{齿轮，}\ Y_{i-1}Y_i=BB\\ 0 & \text{其它}\end{cases}$$

$$\mathrm{t}_2(Y_{i-1},Y_i,X,i)=\begin{cases}1 & X_i=\text{齿轮，}\ Y_{i-1}Y_i=BI\\ 0 & \text{其它}\end{cases}$$

$$\mathrm{t}_3(Y_{i-1},Y_i,X,i)=\begin{cases}1 & X_i=\text{齿轮，}\ Y_{i-1}Y_i=BO\\ 0 & \text{其它}\end{cases}$$

$$\mathrm{t}_4(Y_{i-1},Y_i,X,i)=\begin{cases}1 & X_i=\text{齿轮，}\ Y_{i-1}Y_i=IB\\ 0 & \text{其它}\end{cases}$$

$$\mathrm{t}_5(Y_{i-1},Y_i,X,i)=\begin{cases}1 & X_i=\text{齿轮，}\ Y_{i-1}Y_i=II\\ 0 & \text{其它}\end{cases}$$

$$\mathrm{t}_6(Y_{i-1},Y_i,X,i)=\begin{cases}1 & X_i=\text{齿轮，}\ Y_{i-1}Y_i=IO\\ 0 & \text{其它}\end{cases}$$

$$\mathrm{t}_7(Y_{i-1},Y_i,X,i)=\begin{cases}1 & X_i=\text{齿轮，}\ Y_{i-1}Y_i=OB\\ 0 & \text{其它}\end{cases}$$

$$\mathrm{t}_8(Y_{i-1},Y_i,X,i)=\begin{cases}1 & X_i=\text{齿轮，}\ Y_{i-1}Y_i=OI\\ 0 & \text{其它}\end{cases}$$

$$\mathrm{t}_9(Y_{i-1},Y_i,X,i)=\begin{cases}1 & X_i=\text{齿轮，}\ Y_{i-1}Y_i=OO\\ 0 & \text{其它}\end{cases}$$

若当前的 X 和 Y 的情况满足某个二值特征函数（状态特征函数和转移特征函数）的条件，此二值函数的值就为 1，否则为 0。即当 $X_1=$齿轮 时 $Y_1=B$，同时 $Y_0=O$，则 $s_1(Y_i,X,i)$ 和 $\mathrm{t}_7(Y_{i-1},Y_i,X,i)$ 被满足，值为 1，其余为 0。

此时系统已经生成了相应的特征函数，下一步就需要训练得到这些特征函数的权重，一般的 CRFs 模型采用最大似然估计求的权重 λ。

3.2 语料预处理

目前没有统一的用于术语抽取性能测试的语料，我们用爬虫从“太平洋汽车网”上爬取了 500 多篇网页。由于需要纯文本语料，需要对网页进行预处理，包括去除非中文网页和去除 html 标签，因为新旧网页有可能会有重复的内容，还需要对其进行简单的去重操作。预处理完后得到约 1M 大小的纯文本，共 52,9651 字。

在本章中将术语识别当做序列标注问题，就涉及到是采用基于字的方法还是基于词的方法。若基于词，使用通用领域的分词工具对领域文本进行分词会产生一定错误。对于命名实体识别来说，之前的研究结果表明基于字的效果要比基于词的效果好，但是术语和命名实体不同，命名实体一般认为是人名、地名和机构名，除去复杂的机构名外，命名实体的长度一般比较小，而中文术语通常是由二到六个词组成，粒度比较大，而文献[19]也通过实验表明对于人名地名等比较短的命名实体来说，基于字的方法比基于词的方法效果好，而对于复杂的机构名，基于词的方法效果要更好。我们对已有词典中的 7524 条术语进行长度分析，各长度的术语分布情况如表 3.1 所示：

表 3.1 词典中各长度的术语所占比例

Tab. 3.1 Proportion of terms in different length

字数	2	3	4	5~6	7~8	9~10	>10
所占比例（%）	10.26	15.88	28.40	29.84	10.79	3.56	1.27

从表 3.1 可以知道长术语占的比重比较大，基于词的方法可以减少简单词的判断。虽然使用分词工具进行分词处理会产生一定的错误，比如“安装在概念车上”经过分词后是“安装/v 在/p 概念/n 车上/s”，则“概念车”这个词是不会被识别出来的，相比较这些错误，我们认为利用分词工具得到的语义信息所带来的好处更重要，故在本文中使用基于词的方法。

首先对语料进行分词词性标注预处理，分词工具采用实验室开发的分词软件“nihao”。采用目前常用的 BIO 标记法进行标记，即 B 是值一个术语的开头，I 是指一个术语出去开头剩下的部分，O 是指非术语。如“鼓/B 式/I 制动器/I 一般/O 用于/O 后轮/B （/O 前轮/B 用/O 盘/B 式/I 制动器/I ）/O 。/O”。

3.3 有效特征选取

基于 CRFs 的术语抽取，选择合适的特征很关键。文献[10]使用词本身和词性作为特征，文献[11]选取了 6 个特征，即词、词性、左右信息熵、互信息和 *TF/IDF*。文献[12]

将术语的统计信息融合到CRFs模型的特征中，并使用背景语料来强化词语的术语特性，即使用了词的频率、领域频率差、词频的*Rank*值，以及术语所在句子的信息。本文总结了前人的工作，并结合汽车领域术语的特点，选取了九个特征，分别介绍如下：

（1） 词本身 *Word*

根据领域术语的特性可知，有些词只在本领域流通，故词本身包含了术语最大的信息，所以使用词本身作为特征。

（2） 词性 *POS*

通过对已有的汽车术语资源分析可知虽然组成词性模式有很多种，但是大部分是名词性短语，统计得到前三位词性组合模式为“*n+n*”、“*v+n*”、“*n*”，可见词性对于术语的识别是一个重要特征。另外，汽车领域中一些术语是中英文搭配组成，用词性作为特征可以将此种情况考虑在内。

（3） 词的长度 *WordLen*

领域术语中有一部分词是未登录词，通用的分词系统对于未登录词的处理办法通常是分成单个字，比如“排挡杆”被标记为“排/*v* 档/*Ng*　杆/*Ng*”，可以利用这个特性，通过考虑当前词的长度来判断其是否作为术语中的一部分。

（4） 是否在已知词典中 *IsDic*

本文整理的词典中共7525条术语，由3109个词组成，可知一些词不止在一个术语中出现。表3.2给出了分词后的长度分布情况，可以看出，长术语占80%以上，单词在长术语中出现的位置信息可以作为一项特征。经分析统计，词典中的3109个词按在术语中的所处位置可分为以下六种情况：

① 只作为单词型术语，如“外胎”，词典中不存在其出现在复杂术语中的情况。此类词共166个，占5.34%，记为 *OS*；

② 可单独使用也可以作为复杂术语的一部分，占8.11%，记为 *DS*；

③ 只出现在复杂术语的开头，占14.09%，记为 *DB*；

④ 只作为复合词的结尾，占20.75%，记为 *DE*；

⑤ 只出现在复合词的中间位置（针对由两个以上的词组成的术语），占40.59%，记为 *DI*；

⑥ 只出现在复合词中，但其出现的位置不固定，占11.13%，记为 *OD*。

根据以上分析，我们将词典特征分为7个值，分别为 *OS*、*DS*、*DB*、*DE*、*DI*、*OD*、*O*，其中 *O* 为当前词不在词典中。

表 3.2 不同长度的术语所占的比例

Tab. 3.2 Proportion of terms in different length

词数	1	2	3	4	5	6	>=7
所占比例（%）	9	35	31	15	6	2	2

（5）当前词前后窗口大小范围内的词的词典特征 *WinDic*

文献[15]指出，一个候选术语，如果其前后窗口大小范围内的词中，已在词典中存在的词所占的比例大于一定阈值，则此候选术语也被视为术语。文献[20]分析得到一个领域通用词，如“是”，其周围的词通常是领域相关的。本文结合这两个特点，将上下文的词典特征分为三种类型：一是当前词窗口范围内的词在词典中出现的比例大于阈值且当前词也在背景语料中出现，其值为 1；二是比例大于阈值，但是当前词不在背景语料中出现，记值为 2；三是除去一、二外的情况，值记为 3。

文献[12]将术语的统计信息融合到 CRFs 模型的特征中，并使用背景语料来强化词语的术语特性。本文借鉴其中采用的统计特征，在前文介绍的特征的基础上加入和频率有关的特征（6）~（9），详细介绍如下：

（6）当前词在领域语料中的频率 *DomainFreq*

记 *C_word* 为当前词在语料中出现的频次，*C* 为语料中的总词数，则当前词的频率为：*DomainFreq* = f(*w*)/*C*。由于计算出的频率值是浮点数，不能直接用于 CRFs 的特征值，可以把浮点值按大小分为几类，本文按五类划分，即特征值取 1、2、3、4、5。

（7）当前词在背景语料中的频率 *ContrastFreq*

选用计算机语料作为背景领域语料，共 8014 行，20,800 个词。频率的计算方法和特征值的取值方法与汽车领域相同。

（8）当前词在两类语料中的频率差*△Freq*

（9）当前词所在句子中的所有词的语料频率差之和 *Sen_△Freq*

系统是要根据特征模板从训练语料中抽取特征，特征模板科学与否对于抽取效果的影响也很大。实验中要验证各类特征对抽取性能的影响，所以各个特征是在基本特征模板的基础上逐条加入的。基本特征模板为：

$Word(n)\{n=-2,-1,0,1,2\}$

$POS(n)\{n=-2,-1,0,1,2\}$

$WordLen(n)\{n=-2,-1,0,1,2\}$

$Word(n-1)Word(n)\{n=-1,0,1\}$

$Word(n)POS(n)\{n=0\}$

$Word(n)WordLen(n)\{n=0\}$

$Word(n)WordLen(n)POS(n)\{n=0\}$

n 表示窗口大小，即考虑上下文特征。前三条表示一元特征，后四条是组合特征，如 $Word(n-1)Word(n)$ 表示前一词和当前词（例如“汽油/发动机”，$Word(n-1)Word(n)$ 就表示前一词是“汽油”同时当前词是“发动机”）。

在基本特征模板的基础上逐条加上其余的特征，通过实验研究各个特征对抽取效果的影响，各个特征逐条加入，依次是：

（1） *IsDic* 特征：

$IsDic(n)\{n=-2,-1,0,1,2\}$

$IsDic(n-1)IsDic(n)\{n=-1,0,1\}$

$Word(n)WordLen(n)IsDic(n)\{n=0\}$

（2） WinDic, DomainFreq, ContrastFreq, △Freq, Sen_△Freq 都是一元特征且窗口为 1，如下所示：

$WinDic(n)\{n=0\}$；

$DomainFreq(n)\{n=0\}$；

$ContrastFreq(n)\{n=0\}$；

$\Delta Freq(n)\{n=0\}$；

$Sen\_\Delta Freq(n)\{n=0\}$

表 3.3 给出了抽取的汽车领域术语特征的示例：

表 3.3 特征的示例

Tab 3.3 An example of features

Word	POS	WordLen	IsDic	WinDic	Domain	Contrast	△Freq	Sen△Freq
位移	n	2	D-B	2	4	5	3	2
传感器	n	3	D	1	2	3	1	2
安装	v	2	D	1	2	2	1	2
在	p	1	D-B	1	1	1	1	2
油门	n	3	D	2	2	5	1	2
踏板	f	3	D	2	2	5	1	2
内	f	1	D-B	1	2	2	5	2

3.4 实验与分析

基于 CRFs 的术语抽取流程如下：

（1） 从网页上获取语料，去除网页标签、纯英文文本、简单去重得到纯文本语料。根据指定的标注规则进行人工标注。将所得语料分为训练语料和测试语料，实验中为了减少数据不平衡带来的影响，采用五倍交叉的方法，将语料分为 5 份，取其中的 4 份作为训练语料，剩下的 1 份作为测试语料，分别做五组实验，最后结果去五组的平均值。

（2） 使用分词工具对语料进行分词词性标注处理，从语料中抽取制定好的特征，按照 CRFs 规定的格式建立好训练集和测试集。

（3） 训练 CRFs 模型。训练模型的过程主要是生成特征函数并进行权重估计。

（4） 使用生成 CRFs 模型在测试语料上进行测试，测试过程是根据 Viterbi 动态算法求解最优序列的过程。

整个流程如图 3.1 所示：

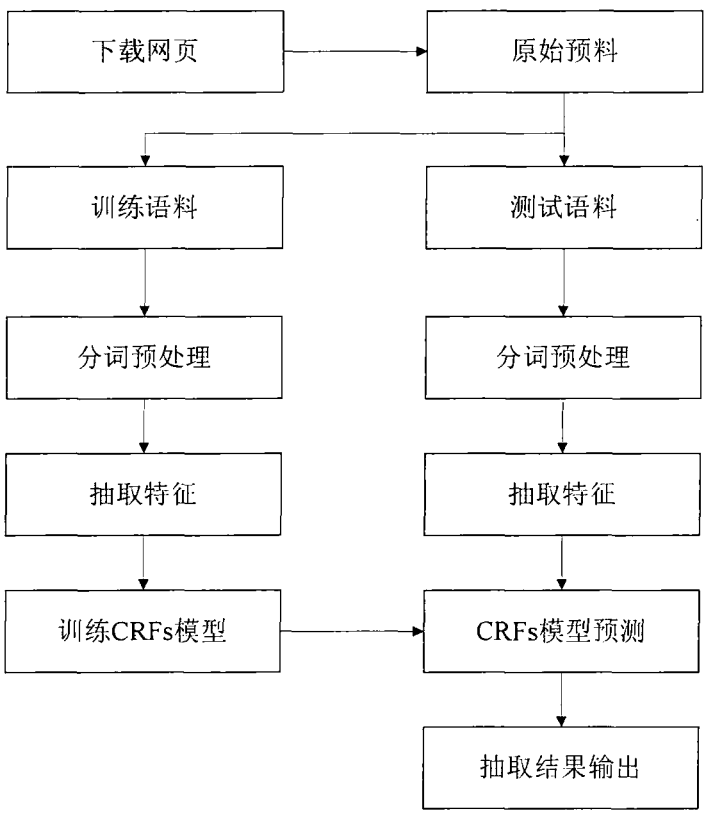


图 3.1 基于 CRFs 的术语抽取流程

Fig. 3.1 Procedure of term extraction based on CRFs

3.4.1 实验数据

使用 Heritrix 从“太平洋汽车网”的“汽车知识”版块爬取约 500 篇网页，去除 HTML 标签等噪音得到纯文本文档，进行去重处理，得到约 1M 的领域语料，共 52,9651 字。由于没有统一汽车术语的标准，我们在研究了相关汽车知识后制定了相应的标注标注对语料进行标注。人工标注后得到 28,674 个术语（包含重复）为了减少数据不平衡对实验结果的影响，对语料分成 5 组，进行 5 倍交叉测试。背景语料选取的计算机领域的，共 8014 个句子。

以第一组数据为例，测试语料中共 2069 条术语（不包含重复），按分词后的组成成分的个数作为计算词长的标准，如“汽车发动机”分词后为“汽车/t 发动机”，计其词长为 2。经过分析可以看出本语料包含的术语在长度上基本符合一般领域术语的分布规律。各长度所占比例如图 3.2 所示：

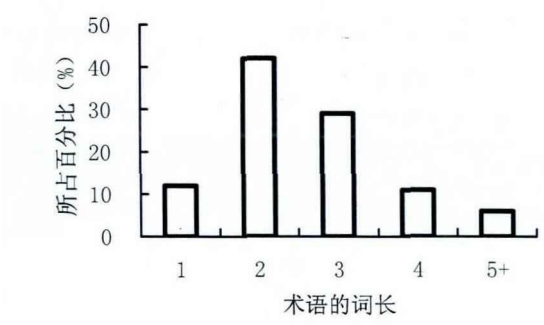


图 3.2 测试语料中各长度的术语所占比例

Fig. 3.2 The proportion of each length

3.4.2 实验结果与分析

采用精确率（P）、召回率（R）以及 F-值（F）作为评价指标，计算方法如下：

$$P=\frac{num_r}{num_s}\times 100\%,\quad R=\frac{num_r}{num_c}\times 100\%,\quad F=\frac{2\times P\times R}{P+R}\times 100\%$$

其中，num_s 是系统识别出的术语的总数，num_r 是 num_s 中的正确术语个数，num_c 是语料中的术语总数。

（1） 实验结果

本文采用了九个特征进行术语抽取，为了验证特征的有效性，我们将各组特征分别加入到特征集中，实验结果如表 3.4 所示，其中各组结果均为交叉测试得到的平均值。

表 3.4　不同特征的识别结果

Tab. 3.4　The results based on different features

所用特征	P(%)	R(%)	F(%)
Word,POS,WordLen	85.49	78.40	81.79
Word,POS,WordLen,IsDic,WinDic	85.18	79.63	82.31
Word, POS,WordLen,IsDic,WinDic, DomainFreq, ContrastFreq	84.61	80.50	**82.50**
Word,POS,WordLen,IsDic,WinDic, DomainFreq, ContrastFreq,△Freq	84.36	80.56	82.41
Word,POS,WordLen,IsDic,WinDic, DomainFreq,ContrastFreq,△Freq,Sen_△Freq	84.34	**80.63**	82.44

由表 3.4 可以看出，使用词本身、词性、词长时正确率最高，加入词典特征正确率后有所降低，召回率提高。加入词典特征正确率反而降低，分析原因可能是有些字在有些词中属于术语的一部分，而在有些词中则不是，比如词典中的“差速器”分词后位“差/速/器”，而在语料中，“差”这个字多用在“之/差”、“较/差”等词中，从而干扰了正确率。在前六个特征的基础上加入词在领域语料和背景语料的频率特征后召回率增加，正确率略有降低，F-值达到最高 82.50%。加入词所在句子的累积频率特征后召回率达到最高的 80.63%，但同时也导致正确率降低，F-值略有降低。

（2） 不同长度的词的识别结果

统计各个长度术语的识别的情况，见表 3.5：

表 3.5　各个长度的术语的识别情况

Tab.3.5　The results of different lengths of terms

词长	1	2	3	4	>=5
百分比(%)	78.60	72.68	74.00	75.22	66.41

其中百分比是指各长度正确识别的术语（不包含重复词）占测试语料中该长度的术语数的比例。从表中可以看出，简单术语识别效果最好，5 词以上长术语的识别效果最差。

（3） 识别结果分析

以第一组为例分析实验结果，我们发现错误主要集中在以下几个方面：

① 识别词语不全，如“多连杆悬架横梁”识别成了“连杆悬架横梁”、“双重防震悬架横梁”识别成了“悬架横梁”。

② 由于分词错误导致的错误，如“定钳式盘式制动器”被识别成“下定钳式盘式制动器”，因为分词的结果是“装/v 下定/v 钳/Ng 式/k 盘/qr 式/k 制动器/n”，CRFs 模型共识别出 1437 个术语（不包含重复），其中错误的占 323 个，有 17 个词是因为分词错误导致。

③ 识别出的词比正确的术语多出一部分，除去因为分词错误的情况外，还有比如“车载 gps”识别成“车载 gps 价格”、“减速器”识别成“带有减速器”的情况。

④ 由于没有统一的标准，在标注上有一些歧义，比如根据标注规则，“3.2 升 fsi 发动机”被判定为一个术语，但是识别结果是“fsi 发动机”，类似的还有“车蜡”被识别成“高档车蜡”，这类词不能断定其错误，和术语判定标准有关。

⑤ 一些词不被认为是汽车领域的术语，但因其自身特点或其所处上下文环境和术语类似也可能被识别出来，比如“激光”、“超声波”等。

⑥ 由于人工标注上不可避免的错误导致识别结果不正确。

由表 3.5 可知单词型术语识别效果最好，长术语较差。其中，单词型术语中诸如“SUV”、“RV”等英文缩写词识别效果较差，分析原因可能是由于这类词所处的语言环境相对不固定，再加上语料稀疏导致。长术语识别效果较差可能是由于出现频次少，组成词串的各个词之间的联系不紧密造成的。

3.5 本章小结

本章主要是针对汽车领域进行术语抽取，将其转化为序列标注问题，使用 CRFs 模型将词、词性、词典、领域频率等多个有效特征整合，采用交叉验证的方法，最终的 F-值达到 82.50%。通过对识别结果进行分析发现模型给出错误标签的词语的边缘概率大部分比较低，可以针对边缘概率低的部分进行后处理提高抽取效果。

4 结合条件随机场与主动学习策略的术语抽取

有监督的机器学习方法受限于已标注语料的规模和质量，人工标注语料费时费力，对标注人员的要求比较高，而且人工操作有很多主观因素影响，标注错误在所难免。就术语抽取来说，技术在不断发展，术语也在不断变化，人工标注语料做到同步更新是不现实的。如何减少人工操作，尽可能多的利用机器进行判断是自动术语抽取的主要解决问题。主动学习的引入可以在一定程度上解决有监督机器学习过度依赖已标语料的问题。我们很容易可以获得大量未标语料，主动学习方法可以从未标语料中选出含有信息量丰富的样本，交由领域专家进行人工标注，减少了语料标注的冗余，提高了工作质量，在机器学习上已经有了比较广泛的应用。目前已经有研究者在词性标注[21]、命名实体识别[22]上应用了主动学习策略，验证了方法的可行性，在本章中我们将主动学习策略引入到术语抽取上，和条件随机场相结合，降低有监督学习对已标样本的过度依赖。

4.1 主动学习

主动学习的基本思想是根据算法迭代的从未标样本中自动找出那些含有丰富信息的样本，将其交由人工标注，加入到原有的训练样本中重新训练模型，这样那些对训练学习器作用不大的样本就不用人工标注了，用较少的有用的已标样本就可以得到精度和之前差不多或更高的学习器，减少了人工标注语料的代价。

主动学习的理论依据就是通过降低学习器的期望错误率来对样本进行优化选择[23]，未标样本的期望错误率表示为 $\int_X E[(\hat{y}(x;D)-y(x))^2 \mid x]P(x)dx$，其中 D 表示标注已标样本集，X 是未标样本集，x 表示其中的一个样本，$y(x)$表示 x 的标签，$\hat{y}(x;D)$ 表示在给定 D 情况下对 x 的预测值，进一步分解期望错误率为三部分，如公式（4.1）所示：

$$\begin{aligned} E[(\hat{y}(x;D)-y(x))^2 \mid x] = {} & E[(y(\mathrm{x})-E[y \mid x])^2]+ \\ & (E_D[\hat{y}(x;D)]-E[y \mid x])^2 + \\ & E_D[(\hat{y}(x;D)-E_D[\hat{y}(x;D)])^2] \end{aligned} \tag{4.1}$$

公式（4.1）中第一部分表示标签和期望标签的差，表示学习器的固有噪声，中间部分是学习器预测值和期望标签的差，表示学习器偏置，第三部分是预测值和期望预测值之差，表示学习器的方差，减少这三部分中的任何一项均能降低学习器的期望错误率，训练样本中的噪声是独立于学习器的，所以不能通过主动学习降低，但是可以通过选择样本降低学习器的偏置和方差，从而降低期望错误率[24]。

基于不确定性的样本选择策略是目前比较常用的主动学习策略之一，该方法是选择当前学习器最不能确定标签的样本，将其交由人工标注，加入到已有的训练集中重新训练新的学习器。基本流程是使用少量已标语料训练得到基本学习器，使用该学习器对未标样本进行预测，同时给出学习器对样本标注的置信度，根据置信度的大小进行选择，再将选择出的样本加入到已有训练集重新训练，这是一个迭代的过程，直到学习器达到某一预计标准或未标语料为空。这种策略的主要依据就是学习器最不确定的样本含有更多训练集中所没有的信息，将这些样本加入到训练集中与随机选取相同规模的样本加入训练集中相比，更能提升学习器的性能。

由以上分析可知如何获得学习器给出的样本标记置信度是基于不确定样本的主动学习方法的主要问题，由第三章的分析可知，CRFs 模型可以求得全局最优序列使得序列的条件概率最大，我们可以利用 CRFs 模型给出的概率值计算模型标记样本的置信度来选择样本，即将主动学习策略和条件随机场结合使用进行术语抽取，以期减少对人工标注语料的过度依赖。

4.2 主动学习与 CRFs 结合

由分析上一章的实验结果可知 CRFs 模型预测错误的标签通常其边缘概率较低，也可以说边缘概率是 CRFs 模型对其给出的预测值正确性的信心，边缘概率越高，则模型给出的标签的正确性越大，反之，边缘概率较低时，模型给出的标签很可能是错误的。给出一个例子说明边缘概率的含义。

给定一个观测序列“片式 熔 断 器”，采用 BIO 标注，则其可能的状态值序列有 81 种，每一个状态序列都有一个相应的概率值，这 81 中序列的概率值的和为 1，因为 CRFs 模型是求出使条件概率值最大的最优序列，每个状态标签的边缘概率是基于序列的概率计算的，如“片式”的标签为“B”的边缘概率是所有 81 个序列中第一个 token 为“B”的序列的概率和，共 27 种序列。经过计算，CRFs 模型各处的状态值序列为“OBII”，第一个 token 的边缘概率为 0.022，值很低，而正确的状态序列应为“BIII”，可见边缘概率低的 token 的模型预测值是错的，原因可能是现有的训练语料中不含有类似的信息，如果将这样的序列进行人工标注并加入到训练语料中重新训练模型，则有可能丰富模型所含有的信息，从而提高了模型的精度。

所以我们将边缘概率值作为样本选择策略的置信度，选取包含较低边缘概率的序列作为模型最不确定的样本，进行人工标注，再重新训练学习器。

算法描述如下：

（1） 初始的已标语料作为训练集 B 训练 CRFs 模型；

（2） 预测未标样本集 U 的标记，计算 U 中样本的置信度，选取最不确定的前 N 个序列，即置信度最低的前 N 个，记为 S；

（3） 将选出的序列加入到训练集 B 中，更新训练集（$B+S$）和未标样本集（$U-S$），重新训练模型；

（4） 如果未标样本集 U 为空或模型预测结果达到预定要求则结束迭代过程，否则重复步骤（2）。

主动学习与 CRFs 结合进行术语抽取的流程如图 4.1 所示：

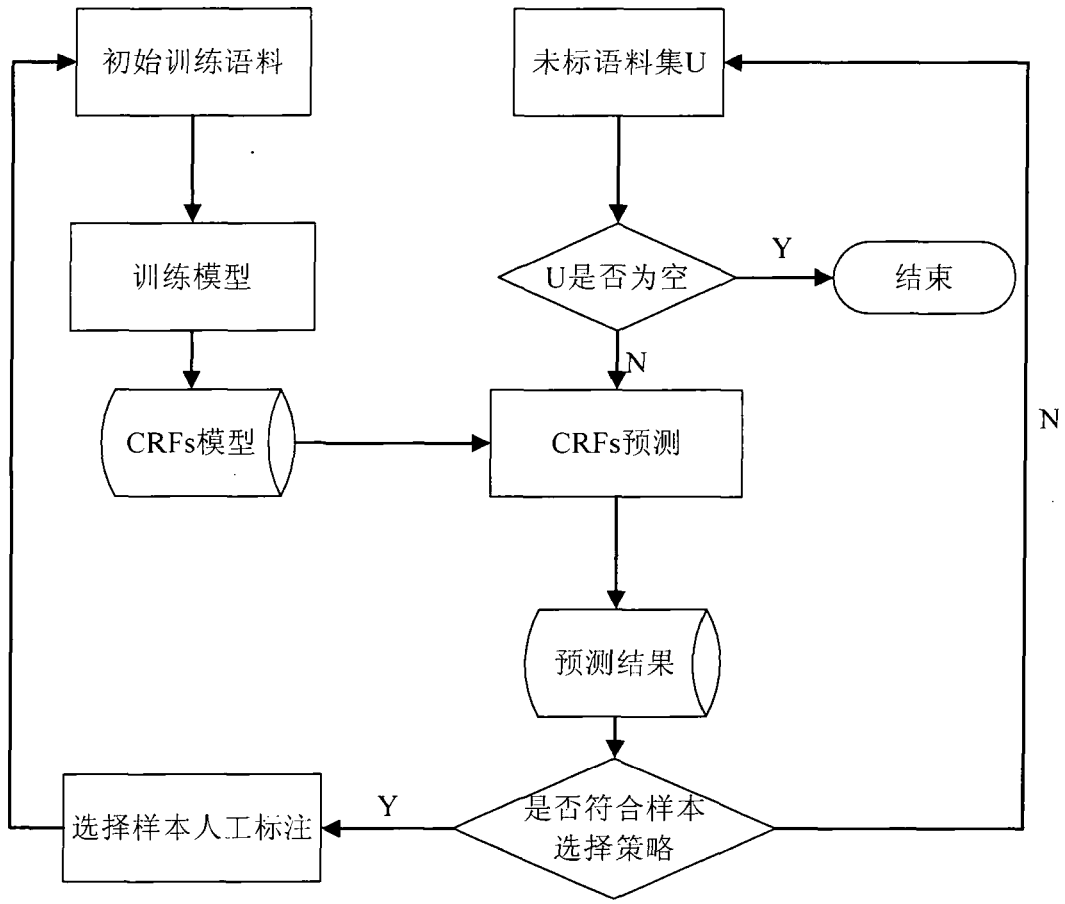


图 4.1 主动学习结合 CRFs 的术语抽取流程图

Fig. 4.1 Active learning combined with CRFs to extract terms

4.3 实验与结果分析

我们使用第三章所介绍的特征，并选择使结果最高的特征模板作为本实验的特征模板。采用图 4.1 所示的算法进行实验。

实验所用语料和基于条件随机场的抽取实验所用语料相同。进行五倍交叉测试，每组的测试语料保持不变，将训练语料分为 10 份，取 10%作为初始训练集，剩下的作为未标语料。迭代过程中根据置信度的值每次选取 N（实验中 N 为全部训练语料的 10%）个样本加入到训练集中重新训练。为了验证上一节分析的算法的有效性，本文做了对比实验，即不是采用主动学习策略选择样本而是从未标语料中随机选取 N 个样本进行人工标注加入到初始训练集中重新训练模型。对比结果如表 4.1 所示：

表 4.1 主动学习和非主动学习的实验结果对比

Tab. 4.1 The results of active learning and non-active learning method

Proportion(%)	主动学习策略+CRFs			随机选择样本+CRFs		
	P(%)	R(%)	F-值(%)	P(%)	R(%)	F-值(%)
10	74.82	64.41	69.22	74.74	64.43	69.19
20	80.37	74.64	77.39	78.22	71.96	74.95
30	82.37	77.60	79.91	80.65	74.74	77.55
40	83.52	79.10	81.24	81.46	76.20	78.73
50	84.13	79.93	81.97	82.45	77.63	79.96
60	84.42	80.31	82.31	83.02	78.41	80.64
70	84.53	80.40	82.41	83.56	79.07	81.25
80	84.59	80.43	82.45	84.00	79.88	81.88
90	84.61	80.47	82.48	84.32	80.18	82.19
100	84.61	80.48	82.49	84.61	80.46	82.48

其中，Proportion 表示训练语料占全部语料的百分比。

用折线图的形式描述正确率、召回率、F-值随主动学习训练语料的规模变化而变化的规律，见图 4.2。图 4.3 表示的是采用主动学习策略和不采用主动学习策略的方法的对比结果。

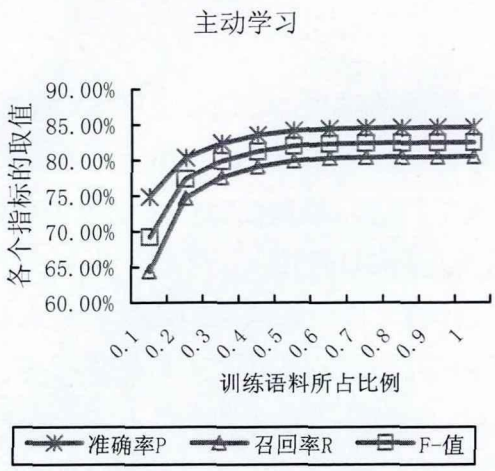
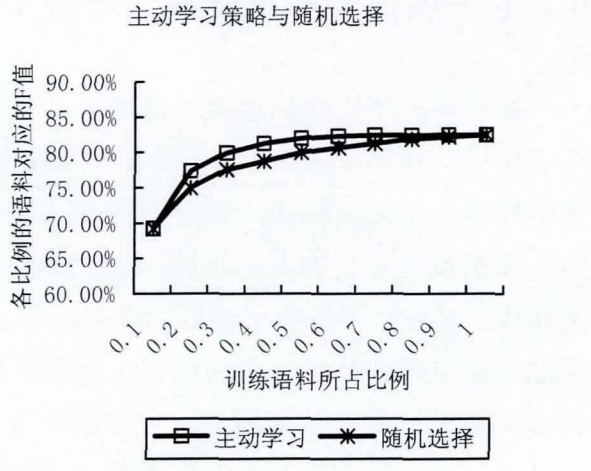


图 4.2 主动学习结果图

Fig. 4.2 Results using active learning



4.3 主动学习和随机方法的对比结果

Fig. 4.3 Results using active learning and non-active learning method

由图 4.2 可以看出，随着主动学习选择的训练样本的逐渐增加，CRFs 模型的性能也随着逐渐提高，当语料增加到 70%的时候其 F-值达到了 82.41%，已经接近最高的 82.49%，只相差 0.08 个百分点，可见如果引入主动学习策略，使用原来训练语料的 70%即可达到和使用全部语料相近的结果，也即是剩下的语料中含有的信息量不多，不足以提高学习器的性能，这部分语料可以不用人工标注即可获得较理想的结果。同时结合图 4.3 显示的数据，加入同样规模的训练语料，使用主动学习策略比随机选取样本所得到的结果要有明显的优势。可见使用主动学习策略可以帮助学习器选择更有用的样本，提高学习器性能，一定程度上减少了人工标注量。

4.4 本章小结及改进方向

本章主要使用了主动学习策略和条件随机场相结合的方法，以减少对已标样本的过度依赖。实验结果表明方法是有效的，使用较少的训练语料可以得到性能较高的学习器。在本章中我们认为边缘概率值的大小反映了 CRFs 模型对标签的信心，用此作为主动学习样本选择策略的置信度，选择最不确定的样本交由人工标注，下一步工作可以启发式学习的方法，进一步减少对标注语料的依赖，减小人工标注的代价。

5 基于统计量的术语抽取方法

基于条件随机场的领域术语抽取已取得比较理想的结果，可是其过度依赖于已标语料的规模与质量。在条件随机场模型中引入主动学习策略可以帮助模型选择含有信息丰富的样本交由人工标注，从一定程度上减少了人工代价，但是已标样本依然不可或缺，对于非领域专家外的人员来说，标注领域语料是很困难的而且人工劳作不可避免的会产生错误。如何尽可能的不需要领域知识和已标语料进行术语抽取成为越来越多人的研究重点。基于统计量的术语抽取方法不需要已标语料作指导训练模型，根据以概率为基础的某种或某些统计量来衡量一个字串成为术语的可能性。术语有单元性和术语性两个主要特点，单元性主要是考察字串成为一个符合语言学规则的有完整意义的词的可能性，目前常用的统计量为频率（Freq）、互信息（MI）、对数似然比（Log-likelihood ratio）、假设检验（卡方检验、T 检验）、信息熵（Entropy）等，术语性是考察一个词或词组和领域的相关程度，常用的统计量为 C/NC-Value、TF/IDF 等。文献[25]分别考察了九种统计量在二字字串结合紧密度上的性能，比较结果如表 5.1 所示：

表 5.1 单个统计量的抽词性能比较

Tab. 5.1 Comparation of results based on different statistic arguments

Top 17,333	Freq	MI	SA	SCP	Dice	LogL	Chi	ZS	TS
F-值(%)	26.28	54.77	42.98	51.77	49.37	43.13	52.97	53.20	39.12
性能比较	MI > ZS > Chi > SCP > Dice > LogL > SA > TS > Freq								

表 5.1 显示互信息在字串结合度的衡量上有优势。判定一个字串是不是一个完整意义的词，除了考虑它的内部结合度外，边界的判定也很重要。一个字串左右边界搭配的不确定性的大小对于判定其是否是一个独立的词有比较明确的意义，如果一个字串的左右边界搭配很不固定，说明此字串经常独立的出现，而反之，如果这个字串的左右边界比较固定，说明它有可能是作为一个子串出现的。信息熵能很好的表述字串搭配的不确定性。以前人的研究为基础，我们在本章中使用 Pat-tree 作为索引结构，主要研究互信息、信息熵、C-value 这三个统计量对领域术语抽取性能的影响，并结合语言学规则对候选串进行过滤，得到最终的术语列表。

5.1 统计模型

基于统计量的术语抽取方法需要计算候选字串的统计量值，需要反复扫描语料获取字串及其频率信息，因此是否采用一个合适的语言模型对于整个术语抽取系统的性能很

重要的影响。由前文的分析可知，领域术语的长度基本集中在2~8个字，如果采用n元统计模型需要依次扫描2~8字的字串，效率很低。我们使用pattree作为索引结构，可以快速的访问任意长度的字串及其频次，提高了整体效率。

Pat-tree是由Gonnet[26]于1992年提出来的，是由PATRICIA算法发展而来。Pat-tree本质上是一种压缩二叉查询树，采用半无限长字符串（semi-infinite string, sistring）作为索引结构，是信息检索领域常用的数据结构，简立峰[27]将其应用到中文信息检索的关键词抽取上，取得了不错的效果。

半无限长字串是一种子串存储方式，其特点是从起点开始向一方无限延伸。一个长度为n的字串，它的全部子串的个数为$\frac{n(n+1)}{2}$，而sistring的个数为n，即存储n个sistring即可遍历全部子串。比如“汽油发动机”，全部子串为“汽”、“油”、“发”、“动”、“机”、“汽油”、“油发”、“发动”……、“汽油发动机”共15个，它的全部sistring见下表：

表5.2 半无限长字串的示例

Tab. 5.2 An example of sistring

“汽油发动机”的半无限长字串
汽油发动机 00000……
油发动机 0000000……
发动机 000000000……
动机 00000000000……
机 0000000000000……

Pat-tree数据结构的主要要素有在字串中的位置、比较位、左孩子指针、右孩子指针、频率。位置记录的sistring在整个字串流中的起始位置；比较位是记录左子树根节点和右子树根节点记录的二进制序列中第一个不同的位置，如果是0，转向左子树，如果是1，转向右子树；频率记录当前节点代表的字串在语料中出现的次数。

以“发动汽油发动机”为例，其二进制序列如表5.3所示，其Pat-tree结构为图5.1所示：

表 5.3　二进制编码序列

Tab. 5.3　Binary coding sequence

Sistring	二进制序列
动机 00	1011011010101111 1011101111111010……
机 00	1011101111111010 0000000000000000……
汽油发动机 00	1100011011111011 1101001111001101……
油发动机 00	1101001111001101 1011011110100010……
发动机 00	1011011110100010 1011011010101111……
动机 00	1011011010101111 1011101111111010……
机 00	1011101111111010 0000000000000000……

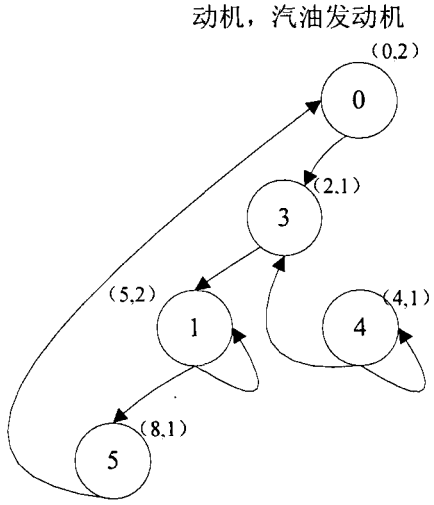


图 5.1　Pat-tree 示例

Fig. 5.1　An example of Pat-tree

5.2　统计量

（1）　信息熵

熵（Entropy）指的是信息的不确定性，一个事件越不确定，其熵值越高。在术语抽取领域，可以使用信息熵来衡量一个字串对其上下文的依赖程度，本文计算字串的左信息熵（Left Entropy, LE）和右信息熵（Right Entropy, RE）来判断这种字符串边界的不确定性。例如汽车领域文本中的“实现离合器油缸活塞的移动，从而完成汽车起动、换挡时的离合器动作”，“离合器”一词的左邻接字是“现”和“的”，在整个语料中进行统计即可知其左邻接字一共 37 种，搭配很不确定，则认为“离合器”是可以作为左边界词，同样对于其右邻接字是“油”和“动”，整个语料中右邻接字一共 41 种，

则认为“离合器”可以作为右边界词。而对于“合器”这个词的左邻接字在整个语料只有“混”、“偶”和“离”三种情况，则认为“合器”通常情况下是和这三个左邻接字同时出现，不适合作为一个左边界词。

左、右信息熵可以表述如下，见公式（5.1）和（5.2）：

$$LE(s) = -\sum_{l \in L} p(ls \mid s) \log_2 p(ls \mid s) \tag{5.1}$$

$$RE(s) = -\sum_{r \in R} p(sr \mid s) \log_2 p(sr \mid s) \tag{5.2}$$

其中 s 是候选字串，l 和 r 分别是 s 的左邻接串和右邻接串，L 表示 s 的所有左邻接串的集合，R 表示 s 的所有右邻接串的集合。$p(ls \mid s)$ 表示在 s 出现的情况下，l 和 s 共现的条件概率，$p(sr \mid s)$ 类似。*LE(s)*和 *RE(s)*越大,则说明 s 左或右的邻接串越不固定，即说明 s 通常不会依赖于其左或右边的词出现，独立成词的可能性越大。将左信息熵和右信息熵结合起来考虑得到公式（5.3）：

$$Entropy(s) = eLE(s) + (1-e)RE(s) \tag{5.3}$$

在本文中使用 e 为 0.5，即左右信息熵的权重相同。

（2） 互信息

互信息 $I(X,Y)$ 是信息论中的一个重要的概念，表示事件 X 和 Y 间的相关性。应用在计算语言学中则表示两个对象（词）间的相互性，互信息[28]的公式为：

$$MI(w_1, w_2) = \log \frac{p(w_1 w_2)}{p(w_1) p(w_2)} \tag{5.4}$$

其中 w_1, w_2 表示两个字串，$p(w_1)$ 和 $p(w_2)$ 分别表示 w_1, w_2 的频率，$p(w_1 w_2)$ 表示 w_1, w_2 的共现概率。如果 w_1 和 w_2 通常是相邻共现的，$w_1 w_2$ 的概率大于 w_1, w_2 随机出现的概率，即 $p(w_1 w_2) >> p(w_1) p(w_2)$，此时 MI 值很大；如果二者是随机分布的，$p(w_1 w_2) \approx p(w_1) p(w_2)$，则 $MI \approx 0$；如果二者是互补分布的，即 w_1 出现的地方 w_2 不出现，同样 w_2 出现的地方 w_1 不出现，$p(w_1 w_2) << p(w_1) p(w_2)$，则有 $MI << 0$。可见互信息值越大，两个字串越经常以一个整体出现，即 $w_1 w_2$ 越可能是一个完整意义的词，反之，互信息值越小，w_1 和 w_2 相邻共现的可能性越小，即二者作为一个整体的可能性越小。在语料规模足够大的情况下，可以用字串在语料中出现的次数（用 $f(w)$ 表示）来估计概率值，即互信息的公式可以表示为公式（5.5）：

$$MI(w_1,w_2)=\log\frac{f(w_1w_2)}{f(w_1)f(w_2)} \tag{5.5}$$

公式（5.5）只能计算由两部分组成的字串的互信息，不适用于三字及以上的字串，文献[27]使用了互信息的一种变形形式 SE（Significance Estimation），将其推广到 n 字词串，见公式（5.6）：

$$SE_s=MI_{ab}=\frac{p(s)}{p(a)+p(b)-p(s)}=\frac{f(s)}{f(a)+f(b)-f(s)} \tag{5.6}$$

其中 $s=w_1w_2\cdots w_{n-1}w_n$ 为一个字串，a 是 s 的最长左子串，即 $a=w_1\cdots w_{n-1}$，b 是 s 的最长右子串，即 $b=w_2\cdots w_n$，$f(x)$ 为对应 x 在语料中出现的次数。可以使用 SE 统计量来判断一个字符串在句子中比它的子串更适合作为一个完整的独立体。由公式（5.6）可分析出 SE 的值在 0 和 1 之间，越接近 1，越说明 a 和 b 更适合作为 s 的子串出现；越接近 0，说明 a 和 b 经常独立于 s 出现。如果这个值大于给定的阈值则认为字串 s 可以作为一个独立的词存在。

互信息或其变形形式在字串的内部结合度的衡量上比较有效，但是其对于低频词的情况不适用，当字串的频率很低时，互信息的值也比较小，对于低频术语的识别不利。

（3） C-value

术语的嵌套问题一直是术语抽取中的一个难点，嵌套即是长术语中包含了简单术语，比如“电控机械式自动变速器”这个术语，其子串“自动变速器”和“变速器”也分别是术语。根据术语的组成特点，长术语通常是由简单术语和简单术语或简单术语和非术语组合而成，类似“电控机械式自动变速器”这样的环状术语有很多，给抽取工作带来了难题。国内外的研究者对嵌套术语的研究比较多，Katerina 和 Sophia[29]就提出了 C-value 参数用于解决嵌套术语的抽取问题。经过一系列的参数改进，C-value 统计量可以表示为公式（5.7）[30]：

$$C-value(s)=\begin{cases}\log_2|s|f(s) & s\text{不被嵌套}\\ \log_2|s|(f(s)-\frac{1}{P(T_s)}\sum_{b\in T_s}f(b)) & \text{s被嵌套在其他候选术语中}\end{cases} \tag{5.7}$$

其中 s 是一个候选字串，$|s|$ 是 s 中字的个数，$f(s)$ 是字串 s 在语料中的出现次数，T_s 是包含 s 的所有候选字串的集合，$P(T_s)$ 是包含 s 的字串的数目。

当 s 的父串是其本身时，C-value 的值取决于 s 的频次和长度，即频次越高，是术语的可能性越大，并且对于频次相同的两个候选术语来说，长度大的候选是术语的可能性

越大。当 *s* 存在父串时，其 C-value 的值是字串 s 在语料中出现的频次与 s 的所有父串的频次和的差，也即是字串 s 独立出现，而不是作为其它串的子串中出现的频次。

5.3 基于统计量的术语抽取实验

在本章中我们考察信息熵、互信息、C-value 对不同长度的领域术语抽取性能的独立影响，然后将参数组合，考察组合参数在抽取中的作用。利用统计方法得到的术语中通常会包含一些词不符合本领域术语的组成规则，针对汽车领域术语，我们根据已知的汽车术语总结出词性组合规律对统计方法得到的候选列表进行过滤，进一步得到精确率更高的术语列表。

基于统计量度的术语抽取流程如下：

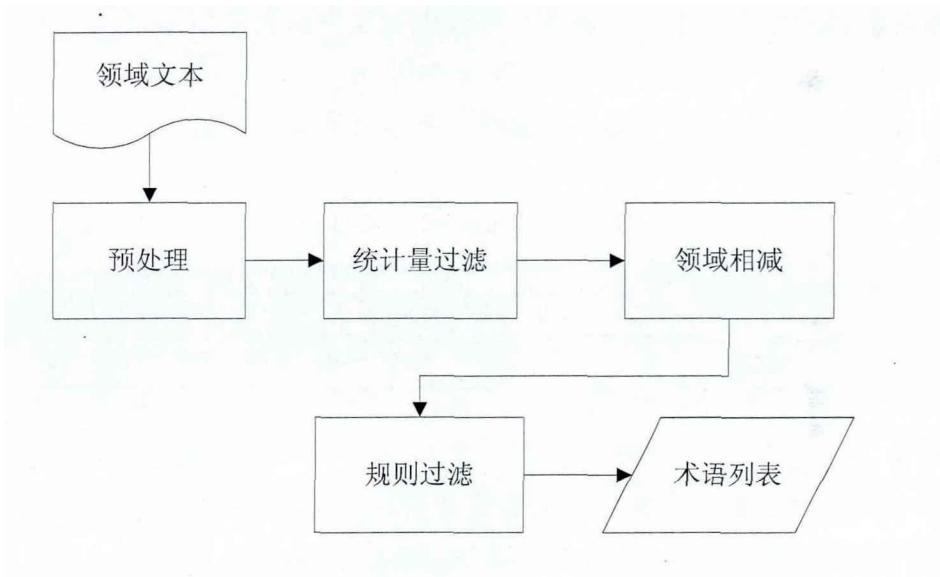


图 5.2 基于统计量的术语抽取流程

Fig.5.2 Procedure of term extraction based statistic method

5.3.1 实验数据预处理

实验所用语料依然是在网页上获取的，预处理得到纯文本的方法和上一章相同。全部语料共 12,058 个句子，我们取其中的 9638 句作为训练语料，剩下的作为已知标签的语料用来统计规则。

由于使用 Pat-tree 数据结构，已知半无限长字串的复杂度和字串的长度有关，为了减小复杂度，需要输入尽可能短的字串。我们已标点符号和停用词作为断句符来分割句

子，将整个文本切割成较短的字串，一个一行，作为 Pat-tree 的输入。停用词是一些常用的冠词、代词等，本身并无特殊的含义，经常和其它词一起使用构成词或短语。这些词一般不会在术语中出现，可以用此作为标识切分句子。经过人工统计，我们选取了 469 个停用词和符号。停用词示例："嘿"、"哼"、"唷"、"呼"、"乎"等。

由于基于统计量术语抽取方法的基本思想计算字符串的某种基于频率的统计量值，符合阈值即被认为是一个正确的术语，这其中会包含那些通用领域的高频词，故使用背景语料进行领域相减，以期将通用领域的高频词过滤掉。实验中采用计算机领域的语料，预处理方法和汽车领域语料相同。

5.3.2 基于统计量方法的术语抽取

根据术语在领域内流通性比较高的特点，一般术语的频率都比较大，在遍历 Pat-tree 得到字串时选择频率不小于 2 的字串。再根据前文的分析可以知道，汽车领域术语绝大部分都在 10 字以内，故遍历时选择长度为 2 到 10 的字串。

经统计训练语料中共有 5638 个术语（不包含重复），按长度统计，各长度的术语数如表 5.4 所示。

表 5.4 不同长度的术语所占的比例

Tab. 5.4 Proportion of terms in different length

术语的字数	2	3	4	5	6	7	8	9	10
个数	519	850	1507	924	849	420	280	120	78

只统计了长度在 10 字以内的术语，大于 10 的术语一共有 91 个，为了保证一定的精确率，忽略这些术语，在抽取字串是只考虑长度为 2 到 10 之间的。

我们考察信息熵、互信息、C-value 三个统计量对汽车领域术语的抽取效果，分别进行三组实验，算法如图 5.3 所示：

（a） 对切分好的文本建立 Pat-tree 索引结构

（b） 遍历长度为 2~10 且词频大于等于 2 的字串，计算其统计量，若其值大于阈值，将此字串加入到候选列表中，否则，继续遍历

（c） 对计算机领域语料进行同样的流程，得到候选列表

（d） 遍历汽车领域候选列表，将不在计算机领域术语集中出现的字串作为最终的汽车领域术语输出

图 5.3 基于统计方法的流程

Fig.5.3 Procedure of the statistical method

三种统计量分别进行上述步骤，使用精确率（P），召回率（R）和F-值（F）作为其评价标准，计算方式如下：

$$P=\frac{num_r}{num_s}\times 100\%,\quad R=\frac{num_r}{num_c}\times 100\%,\quad F=\frac{2\times P\times R}{P+R}\times 100\%$$

其中 num_s 是抽取出的术语总数，num_r 是抽取出的术语中正确的个数，num_c 是语料中的术语总数，这三个量都不包含重复术语的数目。

三组统计量的识别结果如下：

（1） 信息熵

选取阈值为1.5，经过过滤后的总词数为8354个，其中正确的共838个，计算得到P、R、F的值分别为10.03%、14.86%、11.98%。

正确的个数为按长度划分得各长度数目为表5.5示：

表5.5 信息熵结果中不同长度的术语的数目

Tab. 5.5 Number of terms in different length based on entropy

术语的字数	2	3	4	5	6	7	8	9	10
抽取总个数	4988	2230	868	195	59	8	3	2	1
正确的个数	276	219	243	72	23	3	2	0	0

（2）互信息

选取阈值为0.8，共抽取出23668个词，正确的有1085个，计算P、R、F的值分别为4.58%、19.24%、7.40%，各长度的数目分别如表5.6所示：

表5.6 互信息结果中不同长度的术语的数目

Tab. 5.6 Number of terms in different length based on MI

术语的字数	2	3	4	5	6	7	8	9	10
抽取总个数	17	623	5621	4882	4283	3097	2278	1639	1228
正确的个数	1	35	438	216	198	97	63	18	19

（3）C-value

选取阈值为5.0，共抽取出17881个术语，共1456个正确术语，计算P、R、F的值分别为8.41%、25.82%、12.69%，各长度情况详见表5.7：

表 5.7　C-值结果中不同长度的术语的数目

Tab. 5.7　Number of correct terms in different length based on C-value

术语的字数	2	3	4	5	6	7	8	9	10
抽取的总个数	5685	5194	3242	1188	681	341	164	83	1303
正确的个数	278	314	439	171	116	40	34	8	19

将三种统计量的抽取结果作对比，见图 5.4：

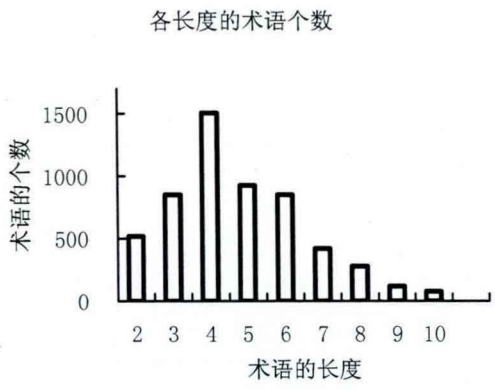
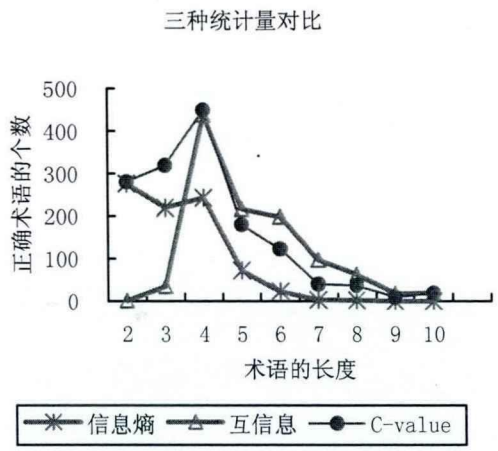


图 5.4　各长度的术语数目

Fig. 5.4　Number of per length



5.5　三种统计量对比

Fig. 5.5　Comparision of three statistic parameters

由图 5.4 和图 5.5 可以看出，对于长度为 2 和 3 的术语，信息熵识别出的正确个数比较多，对于长术语，互信息和 C-value 的抽取效果比较好。由表 5.5 可以看出，信息熵对于较长术语的识别效果很差，分析原因是较长术语的频率通常比较低，信息熵对于低频词不适用，当一个词出现足够的次数才能分析其左右边界的不确定性。对于互信息这个统计量，其精确率很低，因为其是衡量字串的内部结合度的，对于“发动机”这类的词，其子串“发动”和“动机”的内部结合度也很高，这样抽取时也会将其不是术语的子串抽取出来。对于我们将这三种统计量结合，对于长度为 2 和 3 的较短候选串使用信息熵，对于较长候选串使用互信息，以其提高召回率，同时结合 C-value 参数，可以提高精确率。

结合方法的实现流程如图 5.6 所示：

（1） 对切分好的文本建立 Pat-tree 索引结构

（2） 遍历长度为 2~10 词频大于等于 2 的字串，记字串的长度为 l。

若 $2 \le l \le 3$，计算字串的左右平均信息熵值，若大于阈值，计算 C-值，若 C-值大于阈值，将此字串加入到输出列表中，否则，继续遍历；

若 $l \ge 4$，计算字串的互信息值，若大于阈值，计算 C-值，C-值大于阈值，将此字串加入到候选列表 C_A 中，否则继续遍历。

（3） 对计算机领域语料进行同样的流程，得到候选列表 C_C。

（4） 遍历 C_A 中的内容，将不在计算机领域术语集 C_C 中出现的字串作为最终的汽车领域术语输出。

图 5.6 结合方法的流程

Fig.5.6 Procedure of the combined method

根据上述流程得到组合方法的实验结果如下：

共抽取出 16,837 个字串，其中正确的有 1183 个，各长度的字串中正确的个数如表 5.8 所示：

表 5.8 组合方法结果中不同长度的术语正确的术语数

Tab. 5.8 Number of correct terms in different length based on combined method

术语的字数	2	3	4	5	6	7	8	9	10
正确术语的个数	253	208	249	100	96	36	34	9	19
抽取出的总个数	3507	1618	1142	515	481	266	146	85	1228

为了比较组合方法的有效性，我们做了两组对比实验，一组是将信息熵与 C-value 组合，应用在全部字串上；一组是将互信息与 C-value 组合，应用在全部字串上。

实验结果如表 5.9 和表 5.10 所示：

表 5.9 互信息组合 C-值结果中不同长度的术语正确的术语数

Tab. 5.9 Number of correct terms in different length based on combined MI and C-value

术语的字数	2	3	4	5	6	7	8	9	10
正确术语的个数	1	25	249	100	96	36	34	9	19
抽取出的总个数	8	127	1142	515	481	266	146	85	1228

表 5.10 信息熵组合 C-值结果中不同长度的术语正确的术语数

Tab. 5.10 Number of correct terms in different length based on combined entropy and C-value

术语的字数	2	3	4	5	6	7	8	9	10
正确术语的个数	253	208	241	72	23	3	2	0	0
抽取出的总个数	3507	1618	768	175	59	8	3	2	1

三种方法的评测结果对比如下，用 Para1 表示信息熵、互信息、C-值三种参数结合的方法，Para2 表示信息熵结合 C-值的方法，Para3 表示互信息结合 C-值的方法：

表 5.11 三种方法的对比

Tab. 5.11 Comparison of different statistical parameters

	P(%)	R(%)	F(%)
Para1	11.70	17.81	14.12
Para2	13.05	14.22	13.60
Para3	14.23	10.09	11.81

由表 5.11 可以看出，三种参数相结合的方法得到的 F-值最高，且召回率也是最高的。针对精确率不好的情况，可以采用其他策略比如语言学规则进行过滤，提高精确率，对于互信息参数的情况，若短候选串也用互信息进行判断的话很多串会被过滤掉，之后无论采用什么方法也无法召回。可见三种参数组合的方法在提高术语抽取效果上是有效的。

5.3.3 统计量与规则相结合

尽管术语的组成形式很多样化，但是从词性组成模式上看还是遵循一定的规则的。实验中采用了 9638 条句子作为语料进行基于统计量方法的实验，将剩下的 2420 个句子进行人工标注，并用分词工具进行词性标注，作为统计规则的依据，

对人工标注的术语和词典中的术语进行分析，记录每个词性组合模式的频率，得到 top20 如表所示：

表 5.12　词性组合模式的 top20

Tab. 5.12　The top20 pos patterns

词性组合模式 top1~10	频率	词性组合模式 top11~20	频率
n+n	617	b+n	103
v+n	380	n+Ng	98
n	299	a+n	86
vn+n	186	v+v+n	77
n+n+n	166	n+v+Ng	76
n+vn+n	139	n+vi	58
n+v+n	132	vi+v	57
v+n+n	117	n+vn	53
v+Ng	113	v+Ng	48
n+v	113	v+v	45

词典中的 7525 条术语共有 2130 种词性组合模式，由表中显示的情况可以看出出现频次排在前 20 的多是由 2~3 个词组成的，由 3 个词以上组合而成的术语的词性结构比较不稳定，不适合总结词性组合模式的规律，周浪[14]在计算机领域的术语抽取中也利用了这一点，词性组合模式只适用于由 2~3 个词组成术语。故在本实验中只针对由 3 个词以内的词组成的术语进行词性组合模式的过滤。

算法描述如图 5.7 所示：

（1） 扫描已知的术语集中的每一项，记录其词性的组合模式，统计各种词性组合模式的频率。如果频率大于给定的阈值（实验中采用 2），就将此词性组合模式加入到集合 T_w 中。

（2） 对候选术语列表中的词用分词工具进行分词和词性标注，获取其词性组合模式，与 T_w 中的词性模式进行匹配，符合的保留在列表中。

（3） 最终输出术语列表。

图 5.7　规则过滤的流程

Fig.5.7　Procedure of the method based on pos rules

即统计和规则相结合的术语抽取方法的流程图如图 5.8 所示：

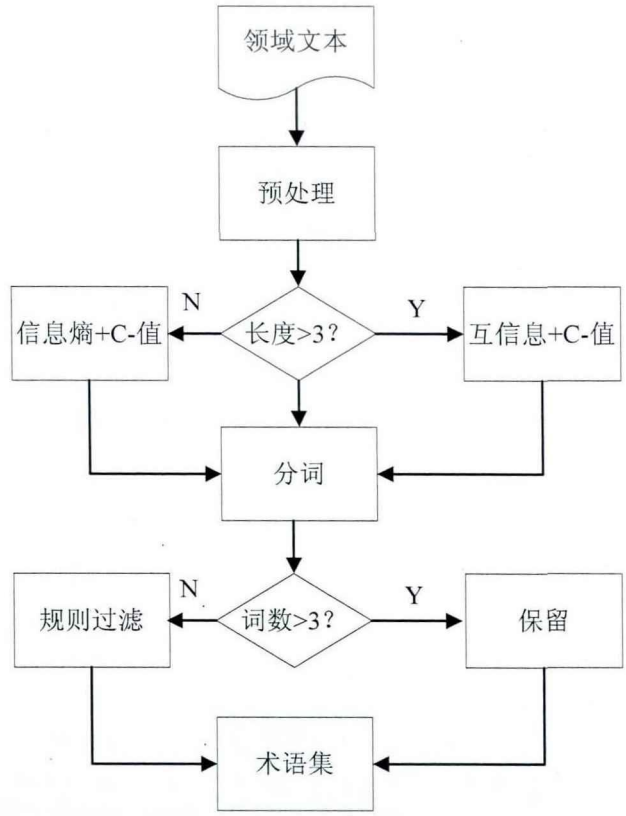


图 5.8　统计和规则相结合的流程图

Fig 5.8　The flow of method based on statistic and rule

经过过滤得到的结果如下表 5.13 所示：

表 5.13　各长度的术语的抽取结果

Tab. 5.13　Results of term extraction for per length

术语的字数	2	3	4	5	6	7	8	9	10
正确术语的个数	238	201	225	78	60	36	34	9	19
抽取出的总个数	2128	874	892	257	179	266	146	85	1228

即一共抽取出 6055 个术语，其中 900 个是正确的，P、R、F 的值分别为 14.86%、15.96%、15.41%。

由 3 个以内个词组成的术语的字数分布范围为 2~6 个字，利用词性组合模式过滤前后各长度术语抽取结果的精确率对比结果如图 5.8 所示：

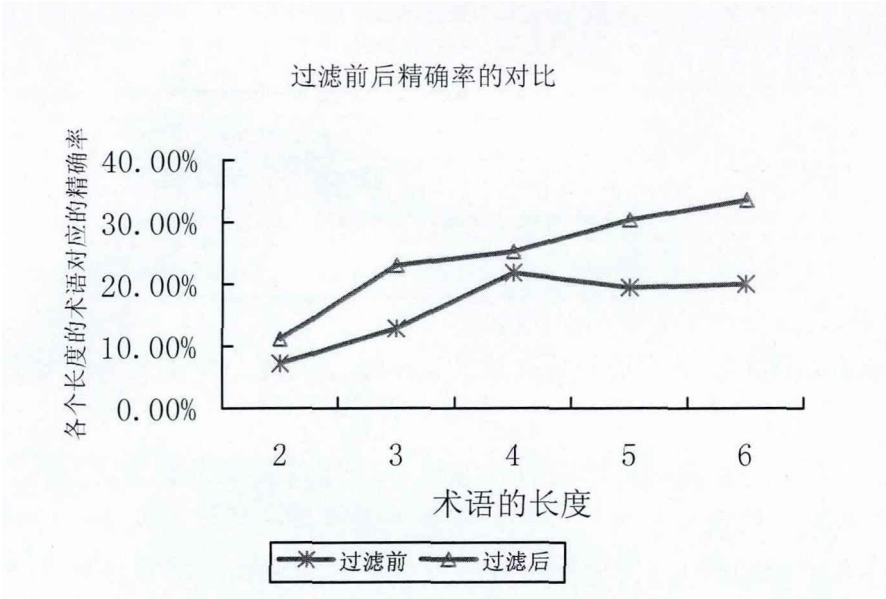


图 5.8　过滤前后精确率的对比

Fig. 5.8　Comparison of precision ratio

有图 5.8 可知，长度为 2~6 的术语经过规则过滤后的精确率均有提高。

总结各个统计量参数的结果如下表 5.14 所示：

表 5.14　几种方法的对比结果

Tab. 5.14　Comparison of different methods

	P(%)	R(%)	F-值(%)
信息熵	10.03	14.86	11.98
互信息	4.58	19.24	7.40
C-value	8.41	25.82	12.61
信息熵+C-value	13.05	14.22	13.60
互信息+C-value	14.23	10.09	11.81
三种结合	11.70	17.81	14.12
三种结合+规则	14.86	15.96	15.41

由表 5.14 可知将三种统计量结合进行抽取的效果要比单独使用各个统计量的结果要好，在此基础上利用语言学规则进行过滤，最终 F-值提高了 1.3 个百分点。

5.4　三类术语抽取方法的分析

本文中采用了三类方法进行汽车领域术语抽取，识别结果如下表 5.15 所示：

表 5.15 三种方法的对比结果

Tab. 5.15 Results based on three methods

	P(%)	R(%)	F(%)
CRFs	84.61	80.50	82.50
主动学习结合 CRFs	84.59	80.43	82.45
基于统计量的方法	14.86	15.96	15.41

从结果上看无监督的基于统计量的方法结果要远小于基于有监督的方法。分析原因如下：

（1） 基于统计量的方法在进行候选串过滤时选取的是词频不小于 2 的词，但是从网站上获取的语料数据稀疏很严重，理论上术语是在领域内流通性较高的词，可实际上有一部分术语的频率为 1，比如“高压共轨式电控燃油喷射系统”、“霍尔效应式曲轴位置传感器”等，这些词候选串过滤时已被过滤，严重影响了召回率。C-value 考虑了词串的长度，优化了低频长术语的抽取，但是对于词长较小的低频术语仍是不利的，比如“中控锁”的词频为 3，C-value 为 3.2，小于我们给定的阈值 5.0。如果一个术语的左右子串是一般领域的高频词，其互信息值会比较低，比如“传动机”的词频为 15，而它的最长左右子串“传动”和“动机”的词频分别为 203 和 1555，这种情况下正确的术语也会被过滤。互信息参数对于短术语的抽取效果不好，使用信息熵代替互信息进行短术语的抽取，可以提高其抽取效果。

（2） 无监督的方法的抽取过程是先利用统计量对候选字串进行排序，大于阈值的则认为是一个正确的字串，然后再利用语言学规则对候选串进行过滤，统计和语言学规则信息是分别在两个步骤起作用的，而有监督的机器学习方法（CRFs）将分词词性标注和频率等多种特征融合到一起，通过选取最适合的特征模板训练得到统计模型，所得结果远好于无监督方法。虽然语料的规模和数据稀疏问题导致统计信息不足，在一定程度上影响了无监督的统计方法在汽车领域语料上的术语抽取效果，但由于 CRFs 模型整合了词、词长、词性和频率等多种有效特征，在同样存在数据稀疏问题的语料上所得结果要好。

将主动学习策略引入 CRFs 模型中，由表 5.15 可以看出引入主动学习策略后，当训练语料是全部已标语料的 70%时的 F-值比使用全部训练语料时差 0.09 个百分点，而不适用主动学习策略时，使用 70%的训练语料得到的 F-值比最高的 F-值差 1.2 个百分点，可见在 CRFs 模型中引入主动学习策略是有效果的，可以部分减少模型对已标样本的依赖，减少人工标注的工作量。

5.5 本章小结

本章研究了基于统计量的术语抽取方法，分析了信息熵、互信息、C-value 三种参数对于术语抽取结果的影响，本文将三种参数结合，针对长度为 2~3 的术语采用信息熵结合 C-value 的方法，对长度大于 3 的术语使用互信息结合 C-value 的方法，结果取两个集合的并集。最后总结词性组合模式对由 1~3 个词组成的候选术语进行过滤，精确率比只用统计量时提高了 3.16%，最终的精确率、召回率、F-值分别为 14.86%，15.96%，15.41%。比较有监督的机器学习方法，无监督的基于统计量方法的效果很差，除了实验所用语料数据稀疏问题严重的原因外，统计量本身还需要进行进一步改进。

结　论

本文主要研究了特定领域的中文术语抽取方法，根据有无已标注的领域语料，从有监督和无监督两种方法上进行研究。有监督的方法我们将术语抽取问题转化为序列标注问题，利用条件随机场综合多种有效特征进行抽取，又进一步引入主动学习策略以减少对已标语料的依赖。无监督的方法主要采用基于统计和规则结合的方法。本文对领域术语抽取的研究主要有以下几个方面：

（1） 基于条件随机场的领域术语抽取

选取汽车领域为目标领域，分析汽车术语的特点，制定了相应的术语标注准则，使用条件随机场训练学习器。详细介绍了训练学习器所用的九个特征，分析了每类特征对抽取效果的影响。在词、词性等特征的基础上加入词典特征后 F-值提高了 0.5 个百分点，在此基础上又加入了领域词频和背景语料词频特征，F-值提高了 0.2 个百分点，最终达到 82.50%。

（2） 引入主动学习策略

为了减少有监督学习对已标注训练语料的过度依赖，本文将主动学习策略引入基于条件随机场术语抽取中。使用最不确定样本选择策略，利用 CRFs 模型给出的边缘概率作为样本选择策略的置信度，即对于含有低边缘概率的序列，模型不能确定其预测值，找出学习器最不确定的 N 个样本，交由人工标注，实验证明在领域术语抽取任务上，主动学习也能起到良好的效果。

（3）基于统计量的术语抽取

为了不受限于已标语料的规模和质量，本文还研究了基于统计量的术语抽取方法，分别分析了用来判断边界的信息熵、用来计算字串内部结合紧密度的互信息和用来解决嵌套问题的 C-值这三种统计量对不同长度的术语抽取效果的影响。实验表明，信息熵对于短术语的识别效果比较好，互信息对于长术语的识别比利用信息熵的结果要好。利用这两个参数的各自优点，在抽取过程中对于较短字串使用信息熵和 C-值结合进行过滤，对于较长字串使用互信息结合 C-值进行过滤。最终的 F-值为 14.12%。为了进一步提高所抽取的术语的精确率，对于由 1~3 个词组成的候选术语集使用词性组合规则进行过滤，总的 F-值提高了 1.3 个百分点，达到 15.41%。

基于无监督的统计量方法与有监督的方法相比要差很多，一方面原因是实验所用的语料是从网页上爬取的，数据稀疏比较严重，很多正确的术语的出现频率只有 1，在语料预处理时就被过滤掉了，这样在后面的处理中召回率肯定比较低，而且对于低频词各统计量的效果也都比较差，从而影响了整个抽取效果。而基于 CRFs 的方法融合语言学

和统计特征，利用各个特征的优点最终结果比较高。另一方面原因是在统计量的使用上仍有不足，不同形式的统计量最适用的条件有所不同，下一步工作要继续研究各种统计量参数，选取合适的参数或合适的参数组合进行实验。

参 考 文 献

[1] 温春，王晓斌，石昭祥. 中文领域本体学习中术语的自动抽取[J]. 计算机应用研究, 2009, 26(7):2652-2655.

[2] 刘桃，刘秉权，徐志明等. 领域术语自动抽取及其在文本分类中的应用[J]. 电子学报, 2007, 35(2):328-332.

[3] Beatrice D, Eric G, Jean M L. Towards automatic extraction of monolingual and bilingual terminology[C]. In Proceedings of the 15th conference on Computational Linguistics. Japan, 1994:515-521.

[4] Didier B. Surface grammatical analysis for the extraction of terminological noun phrases[C]. Proceedings of the 14th conference on Computational Linguistics. Stroudsburg, 1992:977-981.

[5] 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, 2005:72-73.

[6] 何婷婷，张勇. 基于质子串分解的中文术语自动抽取[J]. 计算机工程, 2006(32):188-189.

[7] 胡文敏，何婷婷，张勇. 基于卡方检验的汉语术语抽取[J]. 计算机应用, 2007, 27(12):3019-3025.

[8] 周浪，张亮，冯冲等. 基于词频分布变化统计的术语抽取方法[J]. 计算机科学, 2009, 36(5):177-180

[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研究[J]. 现代图书情报技术, 2008(12):54-48.

[10] 刘豹，张桂平，蔡东风. 基于统计和规则相结合的科技术语自动抽取研究[J]. 计算机工程与应用, 2008, 44(23):147-150.

[11] Zheng D Q, Zhao T J, Yang J. Research on Domain Term Extraction Based on Conditional Random Fields[C]. ICCPOL 2009, LNAI 5459, 2009:290-296.

[12] 章承志. 基于多层术语度的一体化术语抽取研究[J]. 情报学报, 2011, 28(3):275-285.

[13] VU T, AW A T, Zhang M. Term extraction through unithood and termhood unification[C]. In Proceedings of the Third International Joint Conference on Natural Language Processing, 2008:631-636.

[14] 周浪，史树敏，冯冲等. 基于多策略融合的中文术语抽取方法[J]. 情报学报, 2010, 29(3):460-467.

[15] Ji L, Sum M, Lu Q, etal. Chinese terminology extraction using window-based contextual information[C]. CICLing 2007, LNCS 4394, 2007:62-74.

[16] 冯志伟. 现代术语学引论[M]. 北京:语文出版社, 1996.

[17] 《中国汽车工程手册》编辑办公室. 汽车行业名词术语汇编[M]. 北京:人民交通出版社, 1996.

[18] Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]. Proceedings of the International Conference on Machine Learning, Williams, 2001:282-289.

[19] 刘章勋.中文命名实体识别粒度和特征选择研究[D].哈尔滨:哈尔滨工业大学,2010.

[20] Yang Y H, Lu Q, Zhao T J. Chinese term extraction using minimal resources[C]. Proceedings of the 22nd International Conference on Computational Linguistics, Manchester, 2007:1033-1040.

[21] Engelson S P, Dagan I. Minimizing manual annotation cost in supervised training from corpora[C]. In Proceedings of the 34th Annual Meeting of the ACL, 1996:319 - 326.

[22] Saito K, Imamura K. Tag confidence measure for semi-automatically updating named entity recognition[C]. ACL-IJCNLP, Suntec, Singapore, 2009:168-176.

[23] Cohn D A, Chahramani Z M. Active learning with statistical models[J]. Journal of Aritificial Intelligence Research, 1996, 4:129-145.

[24] 冯冲,陈肇雄,黄河燕.采用主动学习策略的组织机构名识别[J].小型微型计算机系统,2006,27(4):710-714.

[25] 罗盛芬,孙茂松.基于字串内部结合紧密度的汉语自动抽词实验研究[J].中文信息学报,2003,17(3):9-14.

[26] Gaston H G, Ricardo A B, Tim S. New indices for text:pat trees and pat arrays[C]. Information Retrieval Data Structures & Algorithms, 1992:66-82.

[27] Chien L F. Pat-tree based keyword extraction for Chinese information retrieval[C]. SIGIR' 97 Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1997:50-58.

[28] Kenneth W C. Word association norms, mutual information, and lexicography[J]. Computational Linguistics, 1990, 16(1):22-29.

[29] Katerina T F, Sophia A. Extracting nested collocations[C]. In Proceeding COLING' 96 Proceedings of the 16th coference on Association Computational Linguistic. Stroudsburg, 1996:41-47.

[30] Mima H, Ananiadou S. An application and evalution of the C/NC-value approach for the automatic term recognition of multi-word units in Japanese[J]. International Journal on Terminology, 2001:175-194.

攻读硕士学位期间发表学术论文情况

1 李丽双, 李丹, 黄德根. 基于条件随机场的汽车领域术语抽取. 大连理工大学网络学刊.

致　谢

衷心感谢导师李丽双副教授对我的悉心教诲，从研究生入学到现在两年多以来，李老师渊博的学识，严谨的学术作风和高尚的人格对我产生了很大的影响，使我受益匪浅。本论文从选题到实现整个过程，李老师都给予我无私的指导，使我的研究之路变得自信和充实。

感谢黄德根教授的悉心指导，他深厚的学术功底和独特的理论视角都对我今后的工作学习产生很大的影响。同时感谢黄老师给我们提供优越的学习环境，让我的学习生活变得更加舒适温馨。

感谢孙静师姐和平金玉师姐对我的无私帮助，使初到实验室时对研究任务不懂的我能够很快入门。

感谢实验室的王敏，佟德琴，刘海霞，高洁，李晓燕，刘钦，王鹏，杨田等对我学习和生活上的帮助，是你们让我时刻感受到家的温暖。

感谢我的父母对我的支持和关心。

最后，再次向帮助过我和支持过我的所有老师和同学表示深深的谢意。

谢谢大家！

大连理工大学学位论文版权使用授权书

本人完全了解学校有关学位论文知识产权的规定，在校攻读学位期间论文工作的知识产权属于大连理工大学，允许论文被查阅和借阅。学校有权保留论文并向国家有关部门或机构送交论文的复印件和电子版，可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印、或扫描等复制手段保存和汇编本学位论文。

学位论文题目：特定领域中文术语抽取

作者签名：李丹　　日期：2011 年 12 月 15 日

导师签名：李丽双　　日期：2011 年 12 月 15 日