中文新词识别技术综述

张海军[1,3] 史树敏[2] 朱朝勇[1] 黄河燕[2]
(中国科学技术大学计算机科学与技术学院 合肥 230027)[1]
(中国科学院计算机语言信息工程研究中心 北京 100097)[2] (新疆师范大学计算机系 乌鲁木齐 830054)[3]

摘 要 新词识别是中文信息处理领域的关键技术。新词识别主要包括候选字串的提取过滤和词性猜测两项任务。中文没有特定符号标志词边界,因此任何相邻字符都有成词的可能性,这给新词提取过滤带来了很大困难;由于没有先验知识和统计数据,新词词性猜测一直是中文词性标注的技术瓶颈。详细分析了中文新词识别技术的研究现状,重点讨论了候选新词提取和词性猜测的研究方法与存在的主要问题,最后对新词识别研究方向进行了展望。

关键词 新词识别,未登录词,候选字串,训练语料,词性猜测
中图法分类号 TP391.1 **文献标识码** A

Survey of Chinese New Words Identification

ZHANG Hai-jun[1,3] SHI Shu-min[2] ZHU Chao-yong[1] HUANG He-yan[2]
(School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China)[1]
(Research Center of Computer and Language Information Engineering, Chinese Academy of Sciences, Beijing 100097, China)[2]
(Department of Computer Science and Technology, Xinjiang Normal University, Urumqi 830054, China)[3]

Abstract New Words Identification (NWI) is a key technology in the field of Chinese information processing. NWI mainly includes two tasks: one is new words candidate extracting and filtering, the other is new words POS guessing. Since there is no specific symbol to mark word boundary for Chinese words, any adjacent characters are possible to compose a word, which brings a lot of obstacles for NWI. Moreover, because the prior knowledge and statistical data are not available, new words POS guessing has become the technological bottleneck of Chinese tagging. The status of the field for Chinese NWI was analyzed in detail, and the research techniques and existing problems for new words candidates extracting and new words POS guessing were discussed emphatically. In the end, the paper presented the prospects of the study for Chinese NWI.

Keywords New words Identification, Unknown words, Candidate string, Training corpus, POS guessing

1 引言

随着时代发展与技术进步,新词大量出现已经成为不可避免的语言现象。在中文信息处理的众多领域,如自动分词、信息检索、词典编纂以及机器翻译等,都需要新词的自动识别。新词识别性能在很大程度上影响着相关信息处理的效果,如中文自动分词技术及新词识别结果已经成为提高分词效果的瓶颈[1-3]。

与印欧语言不同,中文没有特定符号来表示词的界限[4-7],因此任何相邻字符都有构词的可能性[4,8];而且,书面语中没有字符形态变化,这都是新词自动识别的巨大障碍。

在新词识别领域,对"新词"这个概念尚无统一界定,目前的研究包括未登录词识别(Unknown Words Identification UWI)和新词识别(NWI)两方面。其中,未登录词是指未在当前所用词典中出现的词[4,5,9],UWI是中文自动分词过程中的重要阶段,这方面的研究开展得较早,取得了很多成果;而所谓的新词(New Word)是指随着时代的发展而新出现或旧词新用的词[10],如"非典"、"山寨"等。此意义上的新词识别近些年才发展起来。但由于新词也属于未登录词,因此许多研究者对这两个概念不加区别,本文也不做明确区分。

目前,以台湾中研院、哈尔滨工业大学、中科院计算所、北京大学语言研究所、微软亚洲研究院、华中师范大学、山西大学等机构为代表的科研人员在新词识别领域开展了大量工作,取得了丰硕成果,有力地促进了中文信息处理技术的发展。

新词识别主要包括两项具体任务:(1)候选新词的提取以及垃圾字串的过滤;(2)新词的词性猜测。当前,国内外开展的研究主要围绕第一个方面进行,对于新词词性猜测,还有很多工作值得进一步深入[11]。

新词识别的研究方法总体上有两种:基于规则的方法和

到稿日期:2009-04-30 返修日期:2009-07-24 本文受国家自然科学基金项目(60672149),国家863计划重点项目(2006AA010109)资助。
张海军(1973-),男,博士生,主要研究方向为自然语言处理、新词识别技术,E-mail:ustczhj@mail.ustc.edu.cn;史树敏(1978-),女,博士,主要研究方向为信息抽取、自然语言处理、本体论;朱朝勇(1985-),男,博士生,主要研究方向为机器翻译和信息检索;黄河燕(1963-),女,博士,研究员,博士生导师,主要研究方向为机器翻译、自然语言处理、智能系统。

基于统计的方法。前者利用构词学原理、配合语义信息或词性信息来构造模板，然后通过匹配来发现新词；而后者是通过对语料中的词条组成或特征信息进行统计来识别新词[12]。基于规则方法的优点是准确率高，针对性强，但手工编写和维护规则困难[13,14]，且规则一般是领域相关的，所以适应性和移植性比较差；基于统计方法的优点是灵活、适应能力强，可移植性好，但需要大规模语料进行模型训练，由于使用的语言知识较少，一般都存在数据稀疏和准确率低的问题[4,8,13]。目前大部分研究者使用规则和统计相结合的方法，以期发挥各自的优势[15]。

本文第2节论述候选新词的提取和过滤技术，并对比了各种方法的优缺点；第3节分析了新词词性猜测研究现状和存在的问题；最后探讨了新词识别技术的未来研究发展方向。

2 候选新词的提取和过滤技术

理论上，任何相邻的汉字都可能是新词的候选对象。候选新词提取是新词识别的必经阶段。候选字串中必然包含许多非词字串(垃圾串)，对这些垃圾串的过滤相当困难。

获取候选新词，目前主要有两类方法，一是有监督方法，即在大规模训练语料基础上，通过统计方法来确定新词边界，进而获得候选新词；另一种是无监督方法，即不使用大规模训练语料，而是对待处理的文本进行字串频率统计，频率高于阈值的重复串作为候选新词(或使用启发规则来获取候选字串，比如词缀模板)。

2.1 基于监督方法提取候选新词

在大规模训练语料的支持下，研究者将候选新词提取问题转化为分类或标注问题，对各种统计模型进行了尝试和探索。

2.1.1 基于普通统计特征

应用普通统计特征作为候选词串的提取标准。实际上是将候选新词识别问题看作分类问题，使用统计特征，如共现频率、独立词概率等作为分类标准，来区分新词和非新词[3,4,8,16-20]。

Chen等[4,16,20]通过提取新词语素(语素是词的组成部分)的方法来进行新词识别。他们认为新词包含在分词后的散串中，将单字词从散串中剔除后，剩下的单字就是新词语素，由新词语素来组成候选新词。该方法基于大规模语料库，自动获得提取单字词的规则，通过这些规则识别新词语素，使用语素之间的共现频率来确定候选词边界，最后使用形态、语法和统计规则来过滤候选新词。该方法对低频词的识别效果好，准确率可达88%，召回率可达67%。但自动从语料库中获得识别单字词的规则，难以实现和控制；且需大规模训练语料支持。

Wu[8,18]等使用独立词概率IWP(Independent Word Probability)作为判断新词的标准，将中文分词和新词识别系统集成，使二者相互补充。首先用分词系统对句子进行预先分词，在此基础上，将连续的单字串作为新词的候选对象，使用独立词概率(IWP)判断新词，获取的新词可以补充分词词典。对于难以判决的字串，交给句法解析器做最终决断。该方法的准确性很高，可靠性好，但由于使用句法解析器，使得效率过低，不适用于普通的新词识别系统；对于较长的字串判决，数据稀疏问题严重。

2.1.2 基于隐马尔科夫模型(HMM)

HMM能充分利用上下文之间的统计关联，体现统计单元之间的联系，非常适合于标注问题。

很多研究人员将新词识别看作标注问题，具体思路为，先训练HMM模型，然后用Viterbi算法标记字符，最后通过对标记的解码来获得新词[21-23]。张华平[21,22]等提出基于角色标记的新词识别模型。主要思想是：首先使用未登录词角色(role)标记语料训练模型参数，然后在分词基础上，使用viterbi算法对句子进行角色标注，最后使用最大匹配方法，依据标记的组合原则，得到候选新词。Fu[23]等也将新词识别看作标注过程，先使用手工标注的语料(标记了字符位置和词形模式，新词要做特殊标记)训练HMM模型；然后对已分词的句子，使用viterbi算法进行解码，得到由标记组成的序列；最后根据序列的标记模式得到新词。

使用标记，适应性较强，通过角色标记(或词形标记)，可以充分利用词中字符之间的依赖关系来识别任意长度的新词；但标注带有特殊标记训练语料的复杂度很高；且标记不一致问题也会影响识别性能。

还有研究者将新词识别看作分类问题。Fu[24]等将POS看作词的类别，将条件概率模型作为主体框架，用词的连接模型和形成模型对条件概率模型进行简化和训练；使用Viterbi算法来寻找新词边界。实验表明，该方法取得了较好效果，能够将多个特征整合到一起，来提高新词识别效果；但需特殊处理训练语料，预处理较复杂；且整合的特征类别有限。

2.1.3 基于决策树(DT)

DT适合于对多个彼此不相关的特征进行综合，从而整合成统一的模型，进行模式分类。应用此模型，实际上是将新词识别看作判别问题，根据多个特征来判断所抽取成分是否是词。

SORNLERTLAMVANICH[25]等应用DT进行词语抽取，使用串频、左(右)熵、字串长度等信息作为特征来训练DT。对预先抽取的字串，人工标注是词或非词，并计算上述特征数据，通过这些数据来训练决策树，获取新词识别模型。新词识别准确率较高，可达85%。

2.1.4 基于支持向量机模型(SVM)

一些研究中亦使用SVM模型来进行新词和非新词的分类，实现新词识别功能[26,27]。Li[26]等采用独立词概率(IWP)、词频等作为SVM的训练特征，实现对NW11型(两个单字)和NW21型(双字词后跟单字)新词的识别。试验表明，用SVM进行新词识别是有效的，但该模型处理的新词模式有限，需要扩展。Goh[27]等针对不同类型新词所具有的特征集合不同这一特点，使用多个SVM分类器组成层次结构来进行新词识别。虽然针对性较好，但由于分类器过多，时间效率不高。

也有研究者使用SVM模型对字符标注特定标记，然后对标记进行解析，以获取新词。Goh[28]等先用HMM模型来进行分词和预标注，在此基础上，对词标注〈POS-position〉标记，然后使用上述标注过的语料训练SVM模型，最后用SVM对字符标注〈POS-position〉标记，从而获得新词。但基于HMM粗分获得的语料进行模型训练，可能会引入较大的误差。

2.1.5 基于最大熵模型(ME)或条件随机域模型(CRF)

ME 和 CRF 是当前比较流行的两种统计模型，能够综合利用字、词、词性等多层次资源，非常适于标注问题。采用上述模型进行新词识别的基本思路是，用训练好的 ME(或 CRF)标记文本，通过对标记解码，即可获得词边界，实现中文分词[29,30]，进而可以发现新词[30]。

Peng[30]等使用基于字符位置的标记方法训练 CRF 模型，进行中文分词。分词后，计算字串可信度，当可信度高于阈值时，认为是新词。由于 CRF 标记的准确性高，使得新词识别具有很好的效果。

2.1.6 基于监督方法小结

从总体上讲，使用普通统计特征来进行候选新词识别，方法比较简单，系统也容易实现，但这些统计特征一般相对独立，相互关联和全局信息不多，不能充分发挥大规模训练语料的作用。

基于 HMM 的方法比较成熟，能充分利用上下文信息，体现信息之间的关联，但很难加入其它语言信息和特征，使可用信息较单一；从计算上看，对于高元(3 元以上)的 HMM 计算很困难。

DT 模型虽然具有简单高效、计算量不大等优点，但对特征的上下文关联处理较复杂。

SVM 能够整合多个特征来实现新词识别功能，识别准确性较高。但 SVM 模型难以体现特征之间的关联和相互作用。

ME 和 CRF 模型能够很好地应用领域知识和标记之间的依赖，充分利用各种统计信息，提高标注准确性。但在模型训练方面，都需要大规模训练语料的支持，且训练和解码速度较慢。

2.2 基于非监督方法获取候选新词

由于非监督方法没有大规模训练语料的支持，候选字串也是在没有指导的情况下取得的，因此，候选词串数量要比有监督方法的大得多，这对垃圾串过滤提出了更高要求。

2.2.1 基于启发规则

采用词类方法进行新词识别，需要先构造新词模板，然后使用模板匹配来获得候选新词[12,17]。

根据复合词和衍生词的构词原则，梁婷[12]等提出了一种基于规则和统计相结合的三字词识别方法，首先使用构词模板，识别出三字串构成候选新词，然后使用筛检规则进行初步筛选，最后使用神经网络，利用统计特征进行过滤，得到新词。该方法识别新词速度快，但只针对三字词，适应性不强。

Chen[17]根据构词规则来提取词头、词缀以及特殊字符的集合，用来识别专有名词和数字，帮助提高分词效果。但该方法只对简单命名实体识别有效。

2.2.2 基于普通重复串统计方法

这类方法进行新词识别的思路是，直接进行字串的串频统计，频率高于阈值的字串作为候选新词，最后将候选新词中的垃圾串滤除后，剩下的就是新词了。最常用的串频统计方法是用 n 元递增模型[10,13,31,32]。

郑家恒[31]等使用 n 元递增模型($n=2,3,4$)扫描文档提取候选字串，然后使用通用构词规则、特殊构词规则以及互斥字串规则对候选字串进行过滤与召回，来获得新词。虽然使用规则可获得较高的准确率，但由于最长扫描 4 字词，可能漏词；且规则构造难以保证完备。

邹纲[10]等统计按时间排序后的网页中所有的重复字串，频率高于阈值的作为候选词串；接着以某个时间点为界限，把候选字串划分为前景集合与背景集合，取集合差作为新词候选集合。最后使用过滤规则排除垃圾词串。该方法的优点是可获得某个时间点后出现的新词，且集合差运算后，能过滤掉部分垃圾词串。但由于新词出现的时间具有模糊性，集合相减会将部分新词误删；使用基于分词的串频统计，性能会受到分词工具的影响。崔世起[32]等在邹纲的重复字串基础上，先对词的构成模式进行分类，然后采用针对性方法进行过滤。通过统计获得垃圾头(尾)的过滤词典来过滤单字垃圾串；对类似 3+1,2+1 形式的使用词缀词典过滤。该方法针对部分模式字串，过滤效果较好，但处理模式不完善，会造成新词漏召。

有研究者使用滑动窗口来简化串频统计。刘挺[33]使用基于滑动窗口的局部串频统计来提取候选词串，用经验函数来计算候选串权值，权值高于阈值的作为新词。该方法能够有效地提高分词系统的分词效果，但由于使用局部串频统计，会影响新词召回率。

此外，还有研究者使用迭代的二元语法规则，通过反复的迭代与切分来获得候选新词。曹勇刚[34]等先对句子进行二元切分，并统计切分后的整个文档在这一级别上字串的频率，将频率高于阈值的字串加入词频表中；然后合并选取的二元词，作为字进入下一轮迭代；之后再次使用二元语法进行句子切分，再统计这一层面上的串频，扩充词频表，直至二元迭代结束。最后基于停用词表设计过滤规则，去除垃圾字串。试验表明该方法识别效果较好，速度较快。但在迭代时，后续迭代要受到前期迭代影响，可能造成新词漏召。

2.2.3 基于高效的重复串统计算法

罗智勇[35]等使用 PAT-Array 算法提取的重复串作为候选新词；用手工标注的小规模训练语料来训练 SVM 分类模型剔除垃圾串，从而获得新词。由于使用高效的统计算法，重复串的统计效率得到了很大提升。但使用小规模标注语料来训练过滤模型，会影响垃圾串过滤效果。

2.2.4 基于非监督方法小结

使用基于启发规则的方法来获得候选词串，实现起来较简单，识别速度快，准确率较高，垃圾串过滤的压力相对较小；但新词召回率会受到规则完备性的影响；且应用范围会受到限制。

重复串统计方法新词召回率高，但垃圾串过滤任务繁重。使用普通串频统计方法取得重复串，方法简单，容易实现，但效率不高。高效的重复串统计算法，虽然实现起来略显复杂，但效率很高，应该是非监督方法中的主导方法。

2.3 监督方法和非监督方法的总体比较

使用基于监督的候选词获取方法，在技术上相对成熟，产生的垃圾串较少，识别准确率较高，对于低频词有更好的识别效果，适用于在线的新词识别；但需要大规模训练语料或进行复杂的语料处理，前期准备工作复杂。

使用重复串进行新词识别的优点是对新造词识别效果好，而且新词长度不受限制，无需大规模训练语料支持；但低频新词的召回效果较差，特别是只出现一次的新词。由于提取候选串没有监督，垃圾串较多，需要复杂的过滤方法，因此，其效率不高，不适合在线新词抽取。

根据前面的讨论，垃圾串的过滤主要应用于非监督候选词提取模型中，虽然有很多解决方法，但基本上都停留在简单的特征组合层面，没有形成系统的解决方案。

3 新词的词性猜测

与普通词性标注不同，新词词性猜测的主要困难是，没有词典和统计数据的支持。一般意义上的词性标注是基于词典，依据规则或统计数据，确定被标注词词性（POS）的过程。词性标注的前提是：(1)被标注词在词典中出现；(2)有针对被标注词的标注规则或统计数据。而对于新词来讲，以上先验信息或数据都不存在。新词词性猜测准确率低已经成为影响中文词性标注性能提高的最大障碍。

目前针对新词词性猜测开展的研究相对较少。已有工作中使用的研究方法主要是基于统计或统计和规则相结合的方法。

统计所用特征，包括外部特征和内部特征，其中外部特征包括上下文信息（相邻词、相邻字以及相邻标记）、整篇文档的信息等；内部特征包括字串长度、字串前缀（后缀）、字串中每个字符的具体特征（位置，词性）等。同英语相比，可用统计特征不足，一直是困扰中文信息处理的难点。

3.1 基于单类特征的词性猜测方法

这类方法中研究者只使用内部特征或外部特征训练模型，猜测新词词性。

Wu[8]等将未登录词的词性猜测看作分类问题，使用内部特征作为分类依据，根据统计结果判断词性。他们认为字符在同类词的相同位置上的作用是相同的，提出使用字符位置似然概率模型 $P(cat, pos, len)$ 来猜测词性。其中，cat 表示包含该字符词的词性，pos 表示字符在词中的位置编号，len 表示词的长度。如，$P(v,2,3)$表示某字符长度为 3 时动词的第二个字符位置的概率。通过使用词中每个字符的联合概率来计算词性概率，然后根据阈值确定词性。该方法形式简单，计算方便，对于 2,3 字新词的词性猜测效果较好，但对于“长词”词性猜测效果较差；如能将新词的上下文信息加以考虑，可能会提高准确率。

Nakagawa[36]等构造了一个基于全文特征的概率模型，该模型将文档中出现的所有未登录词都纳入考虑范围，使用吉布斯采样来进行参数估计和解码。中文新词词性标注的准确率为 67.85%。虽然研究中使用了全局特征，但如能以某种方式加入新词内部信息和特征，可能会进一步提高词性猜测效果。

3.2 基于组合特征的词性猜测方法

这类方法的基本思路是，首先使用多个特征训练标注模型，然后根据被标注新词的内外部特征直接对其进行词性猜测[11,28,36-38]。

Nakagawa[37]等基于 SVM 框架，使用外部特征和内部特征（前缀，后缀等）来训练模型，实现了未登录词的标注系统。通过两遍扫描来提高词性猜测的准确率，第一遍扫描使用未登录词前面标记，第二遍关注其前面和后面两个标记。该方法词性猜测效果较好，但在 SVM 模型中使用确定性的搜索方法，只考虑了局部信息，没有考虑到句子全局信息。

Lu[39]提出一个基于组合模型的 POS 猜测方法。该组合模型由一个基于规则的模型和两个基于统计的模型组合构成。其中基于规则的模型包含使用未登录词长度、类型等内部特征的词性猜测规则，统计模型包括使用上下文信息的三元文法模型和基于特定长度、特定词性的字符位置似然概率模型（Wu[8]等的模型）。该组合模型的词性猜测效果较好，准确率达到 89%，比以往的方法有较大提高。将多种模型组合，可以有效发挥每个模型的优势，以获得较高的准确率。但对于二字词猜测的准确率不高，系统较复杂，且手工编写的词性猜测规则的完备性难以保证。

Goh[28]等使用未登录词的外部特征和内部特征（外部特征启用一元和二元上下文特征，内部特征涉及词缀和词长）来训练 ME 模型，进行未登录词的 POS 猜测。该模型准确率达 91.07%，但内部特征挖掘不充分，准确率依然有很大的提升空间。

Qiu[11]等基于 CRF 模型，使用内部特征和全局特征来进行 POS 猜测。首先使用词典词的内部特征训练 CRF 模型，利用训练好的 CRF 进行词性猜测，然后对猜测结果进行置信度评测，对于评分低的标注结果，使用全局特征进行校正。利用搜索引擎查询未登录词，获得包含该词的多个句子，通过分析句子中未登录词的上下文标记来获得其全局特征。系统最后使用规则模板，根据全局特征对词性进行投票判决。该模型准确率达到 94.2%，是当前的最好水平。但该模型没有很好地应用新词的上下文特征；将全文判决交由不相关的搜索引擎投票机制决定，可靠性难以保证；投票所用的规则模板难以完全覆盖各种复杂情况。

3.3 新词词性猜测技术分析

从以上分析可见，只用单类特征（内部特征或外部特征）的词性猜测方法，效果相对较差。将内外特征组合用来训练统计模型，会取得较好的词性猜测效果。至于在词性猜测中内部特征和外部特征哪个更重要，目前研究者更重视前者。

对于目前的新词词性猜测方法，内部特征、外部特征以及机器学习模型三者之间的配合还不完全合理，使得两类特征没有得到充分的利用；虽然在一些研究方法中使用了高效的机器学习模型，但由于训练模板的限制，对内部特征的探索和使用不充分，词性猜测的准确率尚待提高。

在前述所有的词性猜测方法中，对 2 字词的猜测准确率都相对较低（现阶段最好水平低于 88%）。由于 2 字词在新词中比重最大，这类词的词性猜测问题已经成为目前新词词性猜测的瓶颈。如能发掘新特征来提高这类词的猜测效果，必然会有效提高词性猜测的整体准确率。

结束语 新词识别中，对于候选新词的获取，在有大规模训练语料的前提下，使用通用训练模板，用字符位置标记作为补充特征，训练 CRF 或 ME 模型，进行候选新词的提取，会取得较高的准确率和召回率；如没有大规模训练语料支持时，使用高效的重复串统计算法，比如 Pat-Array 等，来统计 N 元重复字串，以获得候选新词。这样虽然所用的资源不多，但可取得很高的新词召回率。这种情况下，会导致垃圾字串增多，需要加强过滤机制。

对于垃圾字串的过滤，笔者认为应综合使用统计特征和启发式过滤规则，最大限度地将垃圾串滤掉，以提高过滤准确性。将各种过滤特征进行系统整合，将是一个很好的垃圾串过滤方案。发掘新颖有效的过滤特征和规则，也不失为一种有前途的解决思路。

对于新词词性猜测，基于现有统计特征，挖掘新特征来提高词性猜测性能，是当前研究要实现的目标。应用统计学习模型作为主体框架，使用外部特征配合词的内部特征，会获得较好的标注效果[40-42]。因为，汉字本身携带了大量信息，这可以作为特征来提高词性猜测的性能。

参 考 文 献

[1] 黄昌宁，赵海. 中文分词十年回顾[J]. 中文信息学报，2007，21(3)：8-19

[2] Gao J，Li M，Wu A，et al. Chinese Word Segmentation and Named Entity Recognition：A Pragmatic Approach[J]. Computational Linguistics，2005，31(4)：531-572

[3] Liu T，Liu B-Q，Wang X-L，et al. The Effectiveness Study of Local Maximum Feature for Chinese Unknown Word Identification [J]. Journal of Chinese Language and Computing，2007，17(1)：15-26

[4] Chen K-J，Ma W. Unknown Word Extraction for Chinese Documents[C]// Proceedings of COLING 2002. Taipei，2002：169-175

[5] Ling GC，Asahara M，Matsumoto Y. Chinese Unknown Word Identification Using Character-based Tagging and Chunking[C]// Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. Sapporo，Japan，2003：197-200

[6] Zhang H，Liu Q，Cheng X，et al. Chinese lexical analysis using hierarchical hidden Markov model[C]// The second SIGHAN workshop on Chinese language processing. Sapporo，Japan，2003：63-70

[7] 周正宇，李宗葛. 一种新的基于统计的词典扩展方法[J]. 中文信息学报，2001，15(5)：46-51

[8] Wu A，Jiang Z. Statistically-Enhanced New Word Identification in a Rule-Based Chinese System[C]//Proceedings of the Second Chinese Language Processing Workshop. Hong Kong，China，2000：46-51

[9] Gao J，Goodman J，Li M，et al. Toward a Unified Approach to Statistical Language Modeling for Chinese[J]. ACM Transactions on Asian Language Information Processing，2002，1(1)：3-33

[10] 邹纲，刘洋，刘群，等. 面向 Internet 的中文新词语检测[J]. 中文信息学报，2004，18(6)：1-9

[11] Qiu L，Hu C，Zhao K. A Method for Automatic POS Guessing of Chinese Unknown Words[C]//Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008). Manchester，2008：705-712

[12] 梁婷，叶大荣. 应用构词法则与类神经网路于中文新词萃取[C]// Proceedings of Research on Computational Linguistics Conference XIII. 2000：21-40

[13] Nie J-Y，Hannan M-L，Jin W. Unknown Word Detection and Segmentation of Chinese using Statistical and Heuristic Knowledge[J]. Communications of COLIPS，1995：47-57

[14] Isozaki H. Japanese named entity recognition based on a simple rule generator and decision tree learning[C]//Proceedings of the 39th Annual Meeting on Association for Computational Linguistics Toulouse. France，2001：306-313

[15] 刘华. 一种快速获取领域新词语的新方法[J]. 中文信息学报，2006，20(5)：17-23

[16] Chen K-J，Bai M-H. Unknown Word Detection for Chinese by a Corpus based Learning Method[J]. Computational Linguistics and Chinese Language Processing，1998，3(1)：27-44

[17] Chen A. Chinese Word Segmentation Using Minimal Linguistic Knowledge[C]//Proceedings of the second SIGHAN workshop on Chinese language. Sapporo，Japan，2003：148-151

[18] Wu A. Chinese Word Segmentation in MSR-NLP[C]//Proceedings of the Second SIGHAN Workshop on Chinese Language. Sapporo，Japan，2003：172-175

[19] Wang Z，Liu T. Chinese Unknown Word Identification Based on Local Bigram Model[J]. International Journal of Computer Processing of Oriental Languages，2005，18(3)：185-196

[20] Ma W-Y，Chen K-J. A bottom-up merging algorithm for Chinese unknown word extraction [C] // Proceedings of the Second SIGHAN Workshop on Chinese Language Processing. Sapporo，Japan，2003：31-38

[21] Zhang H，Liu Q. Automatic Recognition of Chinese Unknown Words Based on Roles Tagging [C] // Proceedings of the 1st SIGHAN Workshop on Chinese Language Processing. Taipei，2002：71-78

[22] Zhang H-P，Liu Q，Yu H-K，et al. Chinese Name Entity Recognition Using Role Model[J]. Computational Linguistics and Chinese Language Processing，2003，8(2)：29-60

[23] Fu G-h，Luke K-k. Chinese unknown word identification as known word tagging[C]//Proceedings of the Third International Conference on Machine Learning and Cybernetics. Shanghai，2004：2612-2617

[24] Fu G，Luke K K. Chinese Unknown Word Identification Using Class based LM[C]// Proceedings of The First International Joint Conference on Natural Language Processing. Hainan Island，China，2004：262-269

[25] Sornlertlamvanich V，Potipiti T，Charoenporn T. Automatic corpus-based Thai word extraction with the C4. 5 learning algorithm[C]//Proceedings of COLING 2000. Nancy-Saarbrucken-Luxembourg，2000：802-807

[26] Li H，Huang C-N，Gao J，et al. The Use of SVM for Chinese New Word Identification[C]//Proceedings of First International Joint Conference on Natural Language Processing. Sanya Hai-Nan island，China，2004：723-732

[27] Goh C-L，Masayuki，Asahara，et al. Training Multi-Classifiers for Chinese Unknown Word Detection[J]. Journal of Chinese Language and Computing，2005，15(1)：1-12

[28] Goh C-L，Asahara M，Matsumoto Y. Machine Learning-based Methods to Chinese Unknown Word Detection and POS Tag Guessing[J]. Journal of Chinese Language and Computing，2006，16(4)：185-206

[29] Xue N. Chinese Word Segmentation as Character Tagging[J]. Computational Linguistics and Chinese Language Processing，2003，8(1)：29-48

[30] Peng F，Feng F，McCallum A. Chinese Segmentation and New Word Detection using Conditional Random Fields[C]//Proceedings of The 20th International Conference on Computational Linguistics. University of Geneva，Switzerland，2004：562-568

[31] 郑家恒，李文花. 基于构词法的网络新词自动识别初探[J]. 山西大学学报：自然科学版，2002，25(2)：115-119

[32] 崔世起，刘群，孟遥，等. 基于大规模语料库的新词检测[J]. 计算机研究与发展，2006，43(5)：927-932

（下转第 16 页）

[11] Kottler S, Kaufmann M, Sinz C. Computation of Renameable Horn Backdoors[C]//IN SAT 2008. LNCS 4996. Berlin Heidebery:Springer-Verlag,2008:154-160
[12] Dilkina B,Gomes C P,Sabharwal A. Tradeoffs in the complexity of backdoor detection[C]//Proc. CP'07. LNCS 4741. 2007:256-270
[13] Nishimura N,Ragde P,Szeider S. Solving #SAT Using Vertex Covers[C]//Proceedings of SAT'06. LNCS 4121. 2006
[14] Downey R G, Fellows M R. Parameterized Complexity [M]. Springer Verlag,1999
[15] Egly U,Tompits H,Woltran S. On quantifier shifting for quantified Boolean formulas[C]//Proc. SAT'02 Workshop on Theory and Applications of Quantified Boolean Formulas. Informal Proceedings,2002:48-61
[16] Paris L,Ostrowski R,Siegel P,et al. From Horn Strong Backdoor Sets to Ordered Strong Backdoor Sets[C]// MICAI 2007. LNAI 4827. 2007:105-117
[17] Szeider S. Matched Formulas and Backdoor Sets[C]//Proceedings of SAT2007. Journal on Satisfiability, Boolean Modeling and Computation,2008
[18] Benoist E, Hèbrard J J. Ordered formulas [C] // Les cahiers du GREYC, CNRSUPRES-A 6072. Number 14, Universitè de Caen-Basse-Normandie(1999)
[19] Szeider S. Generalizations of matched CNF formulas[J]. Annals of Mathematics and Artificial Intelligence,2005,43(1-4):223-238
[20] Russell S,Norvig P. Artificial Intelligence:A Modern Approach (2nd ed)[M]. Prentice Hall,2002
[21] Dowling W F,Gallier J H. Linear-time, algorithms for testing the satisfiability of propositional Horn formulas[J]. J. Logic Programming,1984,1(3):267-284
[22] Aspvall B,Plass M F,Tarjan R E. A linear-time algorithm for testing the truth of certain quantified Boolean formulas[J]. Informatiion Processing Letters,1979,8(3):121-123
[23] Kleine H, Karpinski M, Flogel A. Resolution for quantified Boolean formulas[J]. Information and Computation,1995,117 (1):12-18
[24] Davis M,Logemann G,Loveland D. A machine program for theoremproving[J]. Comm. ACM,1962,5:394-397
[25] Garey M R, Johnson D R. Computers and Intractability [M]. New York:W. H. Freeman and Company,1979
[26] Mittal S,Falkenhainer B. Dynamic Constraint Satisfaction Problems[C]//Proc. AAAI-90. 1990
[27] Jeroslow R,Wang J. Solving propositional satisfiability problems [C]// Annals of mathematics and Artificial intelligence. Springer,1990
[28] Cook S. The complexity of theorem proving procedures [C] // Proc. 3rd Annual ACM Symposium on the Theory of Computation:151-158
[29] Dechter R, Pearl J. Network - based heuristics for constraint - satisfaction problems[J]. Artif. Intell,1987,34(1):1-38
[30] Gomes C, Selman B, Kautz H. Boosting Combinatorial Search through Randomization [C] // AAAI98. New Providence, RI, 1998:431-438
[31] Li C M. Anbulagan. Heuristics based on unit propagation for satisfiability problems[C]//IJCAI'97. 1997:366-371
[32] Mnasson N, Zecchina R, Kirkpatrick S. Determining computational complexity for characteristic 'phase transitions'[J]. Nature,400:133-137
[33] Moskewicz M W, Madigan C F, Malik S, et al. Chaff : Engineering an Efficient SAT Solver[C]//38th DAC. 2001
[34] Tsang E. Foundations of Constraint Satisfaction[M]. Academic Press,1993
[35] Gomes C, selman B, Crato N, et al. Heavy-tailed Phenomena in Satisfiability and Constraint satifaction Problems[J]. J. of Autom. Reasoning,2000,24:67-100
[36] Bacchus F,Dalmao S,Pitassi T. Algorithms and complexity results for #SAT and Bayesian inference[C]//44th Annual IEEE Symposium on Foundations of Computer Science (FOCS'03). 2003:304-351
[37] Buning H, Karpinski M, Flogel A. Resolution for Quantified Boolean formulas[M]. Information and Computation, Elsevier, 1995
[38] 堵丁柱,葛可一,王洁. 计算复杂性导论[M]. 北京:高等教育出版社,2003
[39] http://users. info. unicaen. fr/zanutti/rechch/
[40] Mannila H,Mehlhorn K. A fast algorithm for renaming a set of clauses as a Horn set[J]. Information Processing Letters,1985
[41] Nie Xu-min,Guo Qing. Renaming a Set of Non-Horn Clause[J]. Journal of Computer Science and Technology,2000
[42] Bubeck U,Kleine B H. Bounded universal expansion for preprocessing QBF[C]//Proc. 10th Int. Conf. on Theory and Applications of Satisfiability Testing(SAT'07). 2007
[43] Purtilo J J,Callahan J R. Parse tree annotations[J]. Communications of the ACM,1989

(上接第 10 页)

[33] 刘挺,吴岩,王开铸. 串频统计和词形匹配相结合的汉语自动分词系统[J]. 中文信息学报,1998,12(1):17-25
[34] 曹勇刚,曹羽中,金茂忠,等. 面向信息检索的自适应中文分词系统[J]. 软件学报,2006,17(3):356-363
[35] 罗智勇,宋柔. 基于多特征的自适应新词识别[J]. 北京工业大学学报,2007,33(7):718-725
[36] Nakagawa T, Matsumoto Y. Guessing Parts-of-Speech of Unknown Words Using Global Information[C]// Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. Sydney, Australia,2006:705-712
[37] Nakagawa T, Kudoh T, Matsumoto Y. Unknown Word Guessing and Part-of-Speech Tagging Using Support Vector Machines[C]//Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium. 2001:325-331
[38] Chen A, Zhang Y, Sun G. A Two-Stage Approach to Chinese Part-of-Speech Tagging [C]// Proceeding of the Sixth Sighan Workshop on Chinese Language Hyderabad. India,2008:82-85
[39] Lu X. Hybrid Methods for POS Guessing of Chinese Unknown Words[C]// Proceedings of the ACL Student Research Workshop. Michigan,USA,2005:1-6
[40] Nakagawa T. Chinese and Japanese Word Segmentation Using Word-Level and character-level information[C]//Proceedings of the 20th International Conference on Computational Linguistics. 2004:466-472
[41] Ng H T, Low J K. Chinese Part-of-Speech Tagging: One-at-a-Time or All-at-Once? Word-Based or Character-Based? [C]// The Conference on Empirical Methods on Natural Language Processing. 2004:277-284
[42] Wang Z, Huang C, Zhu J. Which Performs Better on In-Vocabulary Word Segmentation-Based on Word or Character[C]//Proceeding of the Sixth Sighan Workshop on Chinese Language. Hyderabad, India,2008:61-68