基于位置标签与词性结合的组合词抽取方法*

欧阳柳波，周伟光
（湖南大学 信息科学与工程学院，长沙 410082）

摘　要：现有分词系统不能及时收录新词语，因而不能有效识别领域组合词。针对此问题，提出一种位置标签与词性相结合的组合词抽取方法。首先对语料进行文本预处理、添加位置标签、加权词频过滤等建立词条的位置标签集；然后依据位置标签集计算词条在句子中的相邻度判定组合词；最后制定反规则对抽取结果进行过滤，并对垃圾串进行两端逐步消减再判定进一步识别组合词。通过在不同语料库上进行实验，表明本方法具有更高的准确率。

关键词：组合词抽取；位置标签集；相邻度；反规则过滤；新词发现
中图分类号：TP391.1　　**文献标志码**：A　　**文章编号**：1001-3695(2016)04-1062-04
doi:10.3969/j.issn.1001-3695.2016.04.022

Compound word extraction based on location tag and POS

Ouyang Liubo, Zhou Weiguang
(*College of Computer Science & Electronic Engineering*, *Hunan University*, *Changsha* 410082, *China*)

Abstract: Now existing segmentation systems cannot recruit new words timely, so they cannot identify compound words effectively. To solve that, this paper proposed a method of compound word extraction based on location tag and POS (part of speech). First, this method established location tag set for each item by processing corpus texts, adding location tag for each item and filtering items with weighted term frequency. Then it counted adjacent degree to judge compound words on the basis of location tag set. Finally, formulated reverse rules and filtered garbage strings with them, detected combined words further from garbage strings by removing item from the head and the tail. Experiments were carried out on different corpora, and the results show that this method has higher precision.

Key words: compound word extraction; location tag set; adjacent degree; reverse rule filtering; new words detecting

汉语自动分词是中文信息处理的一项基础性工作，但其重要性不言而喻。随着社会的发展，科技的进步，人类的知识越来越丰富，越来越多的组合词不断出现用来表达新概念，如“移动互联网”“数据挖掘”“软件开发过程”等。传统分词系统不能及时把它们收录进词库，无法有效识别它们，而是被切分成“移动/vn 互联网/n”“数据/n 挖掘/v”“软件/n 开发/vn 过程/n”。杨梅[1]在现代汉语合成词构词研究中提出，文本中的词分为组合词和原子词。原子词(atomic word)是语言中用于组合形成其他新词的短词；组合词(compound word)由多个原子词构成，遵循意义组合原理且表达一个完整的概念。由于表达文本关键意义的分词语大部分都是组合词，且当前的分词系统难以识别这些组合词，所以研究组合词的识别方法显得非常迫切且有意义。另外，组合词抽取在机器翻译、文本信息检索、信息抽取、文本分类等领域具有重要的应用价值。

1　相关研究

目前新词识别的主要方法有三种：基于统计、规则和基于统计与规则结合的方法。基于统计的方法，一般通过对语料中的词条进行统计，运用数学的方法来发现新词，不依赖句法语义信息，也不受限于特定的领域。该方法的移植性和通用性很强，但是语料库的规模和候选词的词频对抽取效果影响较大。基于规则的方法，认为句子中的词总是按照一系列规则出现，提取词语搭配规则并结合词性或语义，通过匹配找出新词。这种方法准确率较高，但是规则的定义和提取非常复杂、繁琐和多变，难度较大，召回率较低[2]。而基于统计与规则结合的方法在一定程度上能够兼顾其优缺点，这样既能保证抽取的效率，也能在一定程度上保证抽取的质量，越来越受到重视。

陈建超等人[3]提出的利用词序列频率有向网的合成词提取方法，根据文本信息建立词序列有向矩阵，根据有向矩阵的统计结果获取组合词，抽取出的垃圾串较多。霍帅等人[4]提出的新词发现方法，利用词关联性信息的迭代上下文熵对微博内容的新词进行识别，结合了统计特征与词法特征，准确率达到了 88.1%。Sun 等人[5]提出一种利用新词构成规则与词串统计的方法识别新概念，但对连续出现的长串新词无法有效识别。Peng 等人[6]提出一种利用条件随机域的中文词法切分方法，通过计算置信度来识别新概念。于娟等人[7]提出一种词性分析与串频统计结合的词语提取方法，首先进行无用原子词删除，然后进行以原子词为步长提词，再进行不成词删除与人工提词获取词语。张新等人[8]提出一种基于规则与统计的本体概念获取方法，通过长度递减与串频统计的方法获取词

收稿日期：2014-11-19；**修回日期**：2014-12-24　　**基金项目**：国家自然科学基金资助项目(61472132)；湖南省产学研结合重大科技成果转换资助项目(2010XK6024)；国家核高基重大专项资助项目(2012ZX01045-004-005-002)

作者简介：欧阳柳波(1972-)，男，湖南长沙人，副教授，博士，主要研究方向为智能信息处理、软件与知识工程等(oylb@hnu.edu.cn)；周伟光(1989-)，男(通信作者)，硕士研究生，主要研究方向为中文信息处理.

串,再通过规则过滤与领域归属分析获取领域概念。本文提出一种基于词条位置标签与词性相结合的组合词提取算法。

2　算法分析与设计

2.1　总体介绍

组合词通常由两个或者两个以上的原子词构成。现有的分词系统对未包含的新概念不能有效地识别,而是将其切分成多个能单个识别的原子词。如果两个原子词条按照固定的相邻次序在同一句子中出现多次,就有可能成为组合词。但是这只是形成组合词的必要条件,而非充分条件。只基于统计的词串提取会包含较多的垃圾词串,而本文依据词条的位置标签集进行组合词判定,并在此基础上制定进行垃圾词串过滤,对垃圾串进行两端逐步消减再判定,进一步识别垃圾词串包含的组合词,从而拥有较高的正确率。组合词抽取过程如图 1 所示。

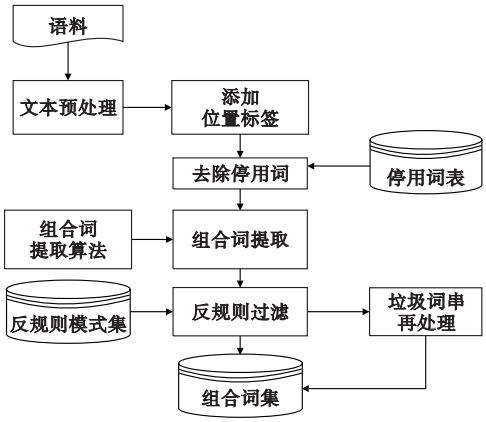

图1　组合词抽取过程图

2.2　相关概念

概念 1　位置标签　指词条在句子中的位置标志,由句子编号、起始位置和结束位置组成。一个词条 c 的位置标签可用一个三元组来标志,具体表示如下:$\mathrm{Tag\_}L(c)=(s,b,e)$。其中,$s$ 表示该词条所在的句子标志,b 表示构成该词条的第一个原子词条在句子中的位置,e 表示构成该词条的最后一个原子词条在句子中的位置。对于原子词条来说,$b=e$。

概念 2　组合度　假如组合词 C 是由原子词条进行 k 次组合而成,就称 k 为组合词 C 的组合度。原子词的组合度取为 0。例如,“移动互联网”由原子词“移动”和“互联网”进行一次组合,其组合度为 1;“自然语言处理”由原子词“自然”“语言”和“处理”组成,需要组合两次,其组合度为 2。

概念 3　同现度　指两个原子词条在同一个句子中出现的次数,即两个词条位置标签集中句子标志相同的个数。同现度反映了两个词条的可组合程度,组合度越大,表明可组合程度越高。假设 T_1 是包含词条 c_i 的句子集合,T_2 是包含词条 c_j 的句子集合。令 $T=T_1\cap T_2$,即代表词条 c_i、c_j 同时出现的句子集合,则句子集合 T 中的元素个数就是词条 c_i、c_j 的同现度。

概念 4　相邻度　指同现度大于设定阈值的两个词条按照固定位置相邻的次数。当相邻度大于设定阈值时,则认定该两个词条可组成一个组合词。如果词条 c_i、c_j 的同现度大于设定阈值,且 T 是同时包含词条 c_i、c_j 的句子 t 的集合,那么词条 c_i、c_j 的相邻度 $N(c_i,c_j)$ 计算方法如下:

a)初始化 $n=0$;

b)选取集合 T 中的一个句子,判断词条 c_i、c_j 是否按照固定的位置相邻(形如 c_ic_j);

c)如果词条 c_i、c_j 相邻(即满足条件:$\mathrm{Tag\_}L(c_i).e=\mathrm{Tag\_}L(c_j).b-1$)则 n 自增 1;否则返回 b);

d)直至集合 T 中的句子被全部选取完。

最终,n 即为词条 c_i、c_j 的相邻度。

2.3　组合词抽取过程

2.3.1　文本预处理

由于分词工具无法对文本中大量的特殊符号、公式及图片进行自动处理,为保证自动分词正确处理,首先要对文本中的特殊符号公式以及无意义的空格、空行等进行去除。再利用中科院的 ICTCLAS 进行自动分词及词性标注,下面给出一个文本片段进行分词及词性标注的结果,如例 1 所示。

例 1:软件/n 工程/n 是/v 一/m 门/q 研究/v 用/p 工程化/v 方法/n 构建/v 和/p 维护/v 有效/a 的/u、/w 实用/a 的/u 和/c 高质量/n 的/u　软件/n 的/u 学科/n。/w 它/r 涉及/v 到/v 程序设计/n 语言/n、/w 数据库/n、/w 软件/n 开发/v 工具/n、/w 系统/n 平台/n、/w 标准/n、/w 设计/v 模式/n 等/v 方面/n。/w

2.3.2　添加位置标签

在自动分词的基础上,根据概念 1,为文本中的每个词条添加位置标签,表示每个词条在文中的位置。方法如下:

a)扫描文本,如果发现逗号、句号、顿号、问号、感叹号、分号、冒号标点符号,则统一替换为句号;

b)以句号为分割标志,对文本进行句子切分,为每个句子添加句子标志,得到句子 t 的集合 $T\{t_1,t_2,t_3,\cdots,t_i\cdots,t_n\}$;

c)依次对 T 中的每个句子 t_i 进行扫描,以分词界限为标志,进行词条切分,对每个原子词条 c 添加位置标签。

d)直至文本中的每个位置的原子词条都拥有一个唯一的位置标签。

为文本中的每个原子词添加位置标签后,每个词条在文本中都有唯一的标志,形成原子词条集。即使相同的词条在语料中出现多次,就会有多个唯一的位置标签。例 1 的文本片段添加位置标签后结果如例 2 所示。

例 2　软件(1,1,1)工程(1,2,2)是(1,3,3)一(1,4,4)门(1,5,5)研究(1,6,6)用(1,7,7)工程化(1,8,8)方法(1,9,9)构建(1,10,10)和(1,11,11)维护(1,12,12)有效(1,13,13)的(1,14,14)、实用(2,1,1)的(2,2,2)和(2,3,3)高质量(2,4,4)的(2,5,5)软件(2,6,6)的(2,7,7)学科(2,8,8)。它(3,1,1)涉及(3,2,2)到(3,3,3)程序设计(3,4,4)语言(3,5,5)、数据库(4,1,1)、软件(5,1,1)开发(5,2,2)工具(5,3,3)、系统(6,1,1)平台(6,2,2)、标准(7,1,1)、设计(8,1,1)模式(8,2,2)等(8,3,3)方面(8,4,4)。

2.3.3　去除停用词

为词条添加位置标签后,发现有一部分词条是中文领域的助词、连接词、感叹词、语气词、介词,如“而且”“通过”“啊”“等”“的”。它们本身并无实际的意义,只有放在句子中起辅助的作用,并且它们一般不会和其他词语组成表达具有实际内涵意义的合成词。去除停用词,不但能提高组合词提取的准确率,又能降低抽取算法的时间复杂度及空间复杂度。

去除停用词常用的方法就是首先建立停用此表,然后进行过滤。停用词表是在借鉴现有停用词表的基础上,进行修改,形成符合本文的停用词表。然后是对处理得到的词条进行停

用词过滤。停用词表部分举例如表 1 所示。

表 1 部分停用词

停用词表				
是	把	比如	按照	拿
且	别	不仅	除非	那里
得	尽管	但是	根据	凭借
和	即使	接着	进而	如果
并非	就是	连同	并且	另外

2.3.4 加权词频过滤

由于一个词条 c 可能在文中出现多次,因此一个词条会有多个位置标签,从而构成该词条的位置标签集。词条的位置标签集获取过程如下:

a)从原子词条集取一个词条,判断该词条是否已有位置标签集;

b)如果已有,则把该词条的位置标签添加到该词条的位置标签集中;

c)如果没有,则为该词条建立一个位置标签集,并把该词条的位置标签添加到位置标签集;

d)循环步骤 a) ~c),直至原子词条集中的词条被取完。

一个词条 c 的位置标签集用 $L(c)=\{\text{Tag\_}L(c)_1,\text{Tag\_}L(c)_2,\text{Tag\_}L(c)_3,\text{Tag\_}L(c)_4,\cdots\}$ 来表示,其中集合的长度 l 即集合中元素的个数,代表该词条在该文本中出现的频次。部分词条位置标签如表 2 所示。

表 2 部分词条位置标签

词条	位置标签集
软件	{(1,1,1),(1,5,5),(2,8,8),(5,11,11),…}
工程	{(1,2,2),(1,6,6),(2,9,9),(3,7,7),(5,12,12),…}
网络	{(2,6,6),(15,12,12),(41,6,6)}
系统	{(13,6,6),(25,12,12),(34,6,6),(39,6,6)}
智能	{(53,6,6),(61,12,12),(91,6,6)}
产品	{(65,2,2),(79,2,2)}

在词频统计的基础上使用 TF-IDF(term frequency-inverse document frequency)算法对原子词条进行加权词频过滤。某一特定文件内的高词语频率,以及该词语在整个文件集合中的低文件频率,可以产生出高权重的 TF-IDF。因此,TF-IDF 倾向于过滤掉常见的词语,保留重要的词语。具体计算方法如式(1)所示,其中 $tf_{ik}(d_i)$ 表示特征词 t_k 在文本 d_i 中出现的频率;$\text{idf}(t_k)$ 表示特征词 t_k 文本强度;N 表示文档集中的文本总数;n_k 表示特征词 t_k 的文本频数;分母为归一化因子[9]。

$$w_{ik}=tf_{ik}(d_i)\times\text{idf}(t_k)=\frac{tf_{ik}(d_i)\times\log\left(\frac{N}{n_k}+0.01\right)}{\sqrt{\sum_{k=1}^{n}\left((tf_{ik}(d_i))^2\times\log^2\left(\frac{N}{n_k}+0.01\right)\right)}} \tag{1}$$

计算词条 c_i 在语料的权重 w_i,设定阈值 W。如果 w_i 小于 W,则从原子词条集中删除,否则保留。

2.3.5 组合词抽取

由概念 4 可知,相邻度反映两个词条在文本中按照固定次序相邻的频次。依据词条的位置标签集,首先计算词条的同现度,如果同现度小于设定阈值,则不再计算其相邻度,这样可以很大程度上减少算法的运算量。如果词条的同现度满足设定阈值,则根据词条的位置标签计算词条的相邻度作进一步的判定。具体抽取过程如下:

a)从原子词条集中选取两个原子词条 c_i、c_j,其位置标签集分别为 $\{\text{Tag\_}L(c_i)_1,\text{Tag\_}L(c_i)_2,\text{Tag\_}L(c_i)_3,\cdots,\text{Tag\_}L(c_i)_m\}$、$\{\text{Tag\_}L(c_j)_1,\text{Tag\_}L(c_j)_2,\text{Tag\_}L(c_j)_3,\cdots,\text{Tag\_}L(c_j)_n\}$;

b)根据词条 c_i、c_j 的位置标签集,依据概念 3,计算词条 c_i、c_j 的同现度 N;如果 N 小于设定阈值 h_1,返回步骤 a);

c)根据概念 4,计算词条 c_i、c_j 的相邻度 M,如果 M 小于设定阈值 h_2,返回步骤 a);

d)对词条词 c_i、c_j 进行合并,形成组合概念 c_ic_j,使 $c_ic_j.\text{Tag\_}L(c_ic_j).b=c_i.\text{Tag\_}L(c_i).b$ 且 $c_ic_j.\text{Tag\_}L(c_ic_j).e=c_j.\text{Tag\_}L(c_j).e$,并将合并后的词条 c_ic_j 加入原子词条集;

e)修改 c_i、c_j 的位置标签集,即从 c_i 的位置标签集中删除满足 c_ic_j 组合条件的 c_i 的位置标签,从 c_j 的位置标签集中删除满足 c_ic_j 组合条件的 c_j 的位置标签;

f)循环步骤 a) ~e),直至原子词条集中不再出现新的组合词;输出组合度大于零的原子词即为组合词。

在文献[10]中的抽取方法中,当两个原子词条进行复合之后,直接删除了该词条,从而排除了该词条与其他词条进行组合的可能性。而本方法步骤 e)是从词条的位置标签集中只删除了满足当前组合条件的标签,从而避免了此种情况的发生,进一步保证了抽取结果的准确性。

2.4 反规则过滤

在上面的基础上进行组合词抽取,正确率与召回率已经较高,但还有很大的提高空间。以上的方法本质上依据的仍然是词串共现的统计思想,没有结合词语本身的意义。一些符合抽取算法的词串被抽取出来,但并不能表达一个完整的概念,而是常见的汉语搭配,如“互联网成为”。组合词是指由两个或两个以上的词语组成并且能表达一个完整的新概念,因此组合词通常都是名词性短语。在自然语言理解领域,词性组合搭配规则分析是最常用的分析方法之一[7]。通过对上面的提取算法中得到候选组合词进行分析,可以发现名词性短语的词性构成方式是多种多样的,因此难以穷举出所有名词性短语的组合规则来对候选短语进行筛选[11]。

表 3 反规则模式部分举例

序号	反规则	举例
1	动词 + 动词	成功/v 实施/v
2	介词 + 名称	在/p 网络/n
3	形容词 + 名称	高/a 水平/n
4	介词 + 动词	被/p 清除/v
5	副词 + 动词	非常/d 有效/v

本文利用上面的提取算法对语料进行训练分析,观察不能构成组合词的相关短语的构词特征,人工总结出不能构成组合词的短语搭配规则(本文称为反规则),对抽取结果进行过滤,以此提高准确率。部分反规则如表 3 所示。

对于被反规则过滤掉的词串(本文称为垃圾串)并没有直接删除,而是进行垃圾串再处理,即进行两端逐步消减再判定。由于一些词串因为包含了其他句子成分,不符合组合词的条件。如果删除首词或尾词之后,就有可能成为组合词。对垃圾串进行首词删除,然后根据已有的词条位置标签集进行再判定,若符合组合词的条件,则加入到组合词集。若不符合,则同理进行尾词删除再判定。若都不符合,则对该垃圾串进行首尾

词同时删除再判定。

3　实验分析

为验证本文方法对组合词提取的准确性与有效性,本文设计了两组实验:一组在复旦大学上海数据库研究中心 NLP 小组提供的文本集上进行实验,随机选取计算机领域的 20 篇论文,平均字数 5 214,称为语料 1;另一组从万方数据库随机选取计算机领域的 20 篇论文,平均字数 6 376,称为语料 2。

3.1　实验结果

对抽取到的组合词进行人工评判,结合相关上下文知识,如果该组合词能够表达一个完整的概念,则判定为正确,否则判定为错误。两组实验的结果如表 4 所示。可以看出,组合词的抽取正确率平均达到了 93.63%,本文方法在不同的语料库都取得较高的正确率。

表 4　组合词抽取结果准确率

数据集	文字总数	提取总数	正确总数	准确率/%
语料 1	104 823	1 092	1 021	93.45
语料 2	127 524	1 356	1 272	93.81

表 5　组合词部分抽取结果展示

语料库 1		语料库 2	
神经网络	计算机软件	移动互联网	大数据
信息处理	分布式计算	遗传算法	模式识别
数据挖掘	信息系统	数据挖掘技术	机器学习
软件开发过程	多线程	文本处理	文本聚类
搜索引擎	故障树	软件开发过程	信息处理
决策树	自动检测	云计算	个性化推荐
循环网络	属性约简	移动推荐	搜索引擎

对抽取结果中正确的组合词进行统计分析,发现由两个和三个原子词组成的组合词占主要部分,达到 83.67%,部分正确结果如表 5 所示。对抽取结果中不正确的短语进行统计分析,发现主要有两类情况:a)抽取出的短语符合抽取算法及构词规则,但仍然不能表达一个完整的概念,如“数据迁移”“语义标注”等。这种情况占所有不正确抽取结果的 62.3%,是因为汉语表达方式多样性所致;b)由于相同的词在不同的语境中有不一样的词性,分词系统不能完全准确地标注出词语的正确词性,这就导致有些不符合构词规则的候选词不能过滤掉,影响结果的准确率。

3.2　实验比较分析

另外选取了近期比较有代表性的组合词抽取方法进行对比实验,文献[3]的基于词序列频率有向网的中文组合词提取算法,记为 T1;文献[10] 基于混合判定模型的复合概念抽取方法,记为 T2;文献[12]的基于词共现有向图的中文合成词提取算法,记为 T3;本文提出的方法记为 T0。四种组合词提取方法的准确率如图 2 所示。

文献[12,13]在文献[3]的词序列有向网的方法基础上进行了改进,建立首尾词表检查并过滤不符合要求的合成词,但范围受限导致可移植性较差,其本质仍是词频统计的思想,虽然进行了,没有结合词性语义的内容,与本文相比垃圾串较多,准确率偏低。文献[10]提出的混合判定模型,虽然有噪声词处理但完全没有结合词性、语义等语法特征,相比本文准确率偏低。而本文利用词条的位置信息抽取组合词之后进行了反规则进行过滤,并对垃圾串进行了特定的识别处理,准确率较高。

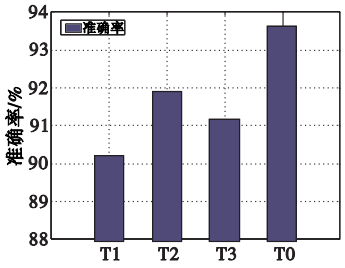

图2　本文方法与其他文献方法准确率结果图

4　结束语

本文提出了一种基于词条位置信息与反规则过滤的中文组合词提取方法,利用词条的位置信息提取组合词,并在此基础上进行反规则过滤,并对垃圾串进行了两端逐步消减再判定,进一步提高了组合词的准确率。通过实验证明,组合词提取的准确率达到了 93.63%。本方法对基于语料库的本体构建、文本分类、关键词抽取等具有较高的应用价值[14]。

下一步的主要工作有:a)对组合词提取算法进行改进优化,优化参数设置;b)结合语义计算方法与领域属性判定方法,对抽取到的组合词进一步识别,解决符合抽取算法但不能表达一个完整概念的问题。

参考文献:

[1] 杨梅. 现代汉语合成词构词研究[D]. 南京:南京师范大学,2006.

[2] 张海军,史树敏,朱朝勇,等. 中文新词识别技术综述[J]. 计算机科学,2010,37(3):6-16.

[3] 陈建超,郑启伦,李庆阳,等. 基于词序列频率有向网的中文组合词提取算法[J]. 计算机应用研究,2009,26(10):3746-3749.

[4] 霍帅,张敏,刘奕群,等. 基于微博内容的新词发现方法[J]. 模式识别与人工智能,2014,27(2):141-145.

[5] Sun Xu, Zhang Yaozhong, Matsuzaki T, *et al*. A discriminative latent variable chinese segmenter with hybrid word/character information [C]//The Annual Conference of the North American Chapter of the Association for Computational Linguistics. Morristown, NJ: Association for Computational Linguistics,2009:56-64.

[6] Peng Fuchun, Feng Fangfang, McCallum A. Chinese segmentation and new word detection using conditional random fields [C]//Proc of the 20th International Conference on Computational Linguistics. Morristown, NJ: Association for Computational Linguistics, 2004: 562-568.

[7] 于娟,党延忠. 结合词性分析与串频统计的词语提取方法[J]. 系统工程理论与实践,2010,30(1):105-111.

[8] 张新,党延忠. 基于规则与统计的本体概念自动获取方法研究[J]. 情报学报,2007,26(6):813-820.

[9] 李学明,李海瑞,薛亮,等. 基于信息增益与信息熵的 TFIDF 算法[J]. 计算机工程,2012,38(8):37-40.

[10] 欧阳柳波,邹北骥,刘丽杰. 一种基于混合判定模型的复合概念抽取方法[J]. 电子学报,2013,41(3):488-495.

[11] 王大亮,涂序彦,郑雪峰,等. 多策略融合的搭配抽取方法[J]. 清华大学学报:自然科学版,2008,48(4):608-612.

[12] 刘兴林,郑启伦,马千里. 基于词共现有向图的中文合成词提取算法[J]. 计算机工程,2011,37(23):177-180.

[13] 刘兴林,郑启伦,马千里. 中文合成词识别及分词修正[J]. 计算机应用研究,2011,28(8):2905-2908.

[14] Zhang Ruiqiang, Yasuda K, Sumita E. Chinese word segmentation and statistical machine translation [J]. ACM Trans on Speech and Language Processing,2008,5(2):1-19.