第37卷 第23期 计 算 机 工 程 2011年12月
Vol.37 No.23 Computer Engineering December 2011

·人工智能及识别技术· 文章编号：1000—3428(2011)23—0177—04 文献标识码：A 中图分类号：TP391

基于词共现有向图的中文合成词提取算法

刘兴林[1,2]，郑启伦[1]，马千里[1]

(1. 华南理工大学计算机科学与工程学院，广州 510640；2. 五邑大学计算机学院，广东 江门 529020)

摘 要：分词系统由于未将合成词收录进词典，因此不能识别合成词。针对该问题，提出一种基于词共现有向图的中文合成词提取算法。采用词性探测方法从文本中获取词串，由所获词串生成词共现有向图，并借鉴Bellman-Ford算法思想，从词共现有向图中搜索多源点长度最长且权重值满足给定条件的路径，该路径所对应的词串即为合成词。实验结果显示，该算法的合成词提取正确率达到91.16%。

关键词：合成词提取；词性探测；词共现有向图；自然语言处理；Bellman-Ford算法

Chinese Compound Word Extraction Algorithm Based on Word Co-occurrence Directed Graph

LIU Xing-lin[1,2], ZHENG Qi-lun[1], MA Qian-li[1]

(1. School of Computer Science and Engineering, South China University of Technology, Guangzhou 510640, China;
2. School of Computer Science, Wuyi University, Jiangmen 529020, China)

【Abstract】Word segmentation systems do not include compound words into their dictionaries, so they can not recognize compound words. To address this problem, this paper proposes a Chinese compound word extraction algorithm based on word co-occurrence graph. It gets word strings from a document through by part-of-speech detecting, generates word co-occurrence directed graph,, and borrows the idea of the Bellman-Ford algorithm to search the longest paths with weight values satisfy the given conditions for multiple starting points in the word co-occurrence directed graph. The word strings corresponding to the paths are considered as compound words. Experimental results show that the algorithm achieves 91.16% upon the precision.

【Key words】compound word extraction; part-of-speech detection; word co-occurrence directed graph; Natural Language Processing(NLP); Bellman-Ford algorithm

DOI: 10.3969/j.issn.1000-3428.2011.23.060

1 概述

中文分词是自然语言处理(Natural Language Processing, NLP)领域的一项非常重要的基础性工作，分词准确性直接影响后续的应用处理。基于字符串匹配的分词算法是目前应用最为广泛、分词效果最好的算法，但是需要词典的支持，因此，当分词系统碰到一些词典未收录的词时，会错误地将未收录词切分成多个语素或多个词。如“人均国民生产总值”被切分成“人均/j 国民/n 生产/vn 总值/n”；“上海世博会”被切分成“上海/ns 世/ng 博会/nr”。

中文文本中的词语可以分为2类：原子词和合成词。原子词是语言中用于组合形成其他新词的短词，一般不遵循意义组合原理，原子词较稳定，不易产生新词。合成词由多个原子词构成，遵循意义组合原理，而且表达了一个完整的概念。由于文本中大部分词语是合成词，并且合成词识别广泛应用于机器翻译、文本信息检索、信息抽取等领域，因此合成词的识别研究具有重要的意义。

本文提出一种基于词共现有向图的合成词提取算法，通过词性探测获取文本中的词串，生成词共现有向图，进而提取出合成词。其中，词性探测是指在文本分词标注的基础上，采用统计的方法获得合成词各原子词的词性情况，进而形成构词规则；词共现是指一篇文本中几个连续(词间无其他任何符号)的原子词出现在同一个句子中，这些连续的原子词构成一个词串，词共现有向图根据文本中的词共现词串生成，图的顶点是词串中的原子词，边是两端顶点在文本中共现的位置。

2 相关研究

文献[1]提出并实现了一种结合词性分析与串频统计的词语提取方法，该方法先对原子词进行词性分析，并建立保留词性表，删除词性表以及停用词表，进而对识别到的词串进行串频统计，当串频达到了给定的阈值时，提取为词语，总体准确率达到了87.5%。

文献[2]借鉴人类的认知心理模式，提出一种基于词序列频率有向网的组合词抽取算法，以识别自由文本中的组合词。算法首先建立描述文本中的词序列出现频率的有向网，然后通过独特的矩阵运算，逐步把组合词提取出来。算法的优点是无须借助专业的语言知识，提取准确率达到了90.2%。

文献[3]提出了一种基于统计和规则的未登录词识别方

基金项目：广东省自然科学基金资助项目(9451064101003233, S2011010003681)；广东省科技计划基金资助项目(2010B010600039)；华南理工大学中央高校基本科研业务费基金资助项目(2009ZM0125, 2009ZM0189, 2009ZM0255)

作者简介：刘兴林(1976－)，男，实验师、博士研究生，主研方向：文本知识获取，智能计算，数据挖掘；郑启伦，教授、博士、博士生导师；马千里，讲师、博士

收稿日期：2011-06-01 **E-mail：**jmxlliu@163.com

法。该方法先对文本进行分词，同时生成临时词典，并利用规则和频度信息给临时词典中的每个字串赋权值，利用贪心算法获得每个碎片的最长路径，由此提取未登录词，实验结果表明其准确率达到了 81.25%。

文献[4]提出一种基于遗传算法的隐马尔可夫模型识别方法，该方法是在高准确率词性标注的基础上实现的。在训练阶段，用遗传算法获取隐马尔可夫模型参数；在识别阶段，先用一种改进的 Viterbi 算法进行动态规划，识别同层名词短语，然后将逐层扫描算法和改进的 Viterbi 算法相结合来识别嵌套名词短语。实验结果表明，此联合算法达到了 94.78%的准确率和 94.29%的召回率。

文献[5]采用基于最大熵模型的方法来获取概念，通过对领域文本进行挖掘得到名词性短语，使用改进的 TF-IDF 公式从中抽取具有领域性的短语，并经人工修正后得到本体概念，算法准确率达到了 81.31%。

文献[6]对英语 BaseNP 和自由名词短语进行识别，使用数据的不同表示方法生成不同的分类器，通过投票技术将各种结果进行合并，设法提高在标准数据集上的 BaseNP 和自由名词短语的识别效果。将该方法在 3 个标准数据集上进行 BaseNP 和自由名词短语的识别，实验结果显示，准确率达到了 93.63%。

文献[1-3]的算法与本文算法类似，但文献[1,3]的算法正确率较低，文献[2]算法的时空复杂度高，算法开销较大。文献[4-6]的算法仅是对名词短语的识别，局限性较大。为解决上述算法中存在的问题，提高合成词提取的正确率，本文提出一种改进的合成词提取算法。

3 基于词共现有向图的合成词提取算法

本文认为，一个词串是合成词必须满足以下 3 个条件：

(1)该词串由句子中 $L(L \geqslant 2)$ 个无间隔的原子词构成。

(2)该词串在文本中多次出现。

(3)在该词串的前面或后面加上其他原子词所形成的新词串出现的次数明显减少。

为便于下文的叙述，将上述关于合成词的判定规则称为 Regulation-1。下面先给出一个已分词并标注词性的文本片段 Sample-1 作为本文的示例文本。

①新/a 的/u 产业/n 革命/vn ，/w 知识/n 经济/n 革命/vn 悄然/d 兴起/v 。/w ②知识/n 经济/n 革命/vn 将/d 造就/v 知识/n 经济/n ，/w 人类/n 将/d 进入/v 知识/n 经济/n 时代/n 。/w ③知识/n 经济/n 革命/vn 是/v 深刻/a 的/u 革命/vn ，/w 必将/d 对/p 世界/n 经济/n 格局/n 产生/v 深远/a 影响/vn 。/w

Sample-1 采用 ICTCLAS3.0 进行分词标注，其中，①、②、③是句子编号，是为便于叙述而加上的，实际分词标注不含句子编号。

很明显，Sample-1 中“知识经济革命”和“知识经济”是 2 个满足合成词判定条件的词串，而且都表达了一个完整的概念，但分词系统把它们切分为“知识/n 经济/n 革命/vn”和“知识/n 经济/n”。本文的研究工作正是基于 Regulation-1 展开的。

3.1 词性探测

词性探测显然是建立在对文本分词标注的基础之上，文献[1]的研究发现，所有构成合成词的原子词中，词性主要集中于名词、动词和形容词，而代词、副词、介词、连词、助词、叹词、语气词和拟声词等不会构成合成词。笔者对复旦大学上海(国际)数据库研究中心 NLP 小组提供的文本集中的 1 600 篇政治经济类论文进行分析统计，结果与上述研究结论基本吻合。因此，建立一个过滤词性表，在提取词串时，若原子词的词性是过滤词性，则将该原子词过滤掉。

另外，经过统计发现，部分符合词性要求的原子词也不会构成合成词，这些原子词称为停用词，如“所在/n”、“是/v”、“要/v”、“看来/v”、“认为/v”、“不少/a”，因此，建立一个停用词表，用于过滤那些符合词性要求的原子词，以减小词串的长度和数量。

依上述方法从 Sample-1 提取到的词串如表 1 所示，其中，位置是一个三元组，记录词串出现的句子编号、句内起始位置和结束位置。

表 1 Sample-1 中的词串

词串	位置	长度
产业/n 革命/vn	1,3,4	2
知识/n 经济/n 革命/vn	1,6,8	3
知识/n 经济/n 革命/vn	2,1,3	3
造就/v 知识/n 经济/n	2,5,7	3
知识/n 经济/n 时代/n	2,12,14	3
知识/n 经济/n 革命/vn	3,1,3	3
世界/n 经济/n 格局/n 产生/v 深远/a 影响/vn	3,12,17	6

3.2 词共现有向图

词共现有向图记为：$G:<V,E>$，其中，V 指文本中的原子词集；E 是由词对构成的集合，边的起点对应词对的首词，边的终点对应词对的末词。有向边的权是一个集合，是词对在文本中共现的位置集，每个元素是一个三元组 $<sno,start,end>$，标识词对所在的句子编号以及在句子中的起始和结束位置。

本文关于词共现有向图的一些基本术语与文献[2]类似。V 的元素用 $v_1,v_2,\cdots,v_{|V|}$ 表示，e_{ij} 表示以 v_i 为起点、以 v_j 为终点的有向边，s_{ij} 表示边 e_{ij} 上的集合，$w_{ij}=|s_{ij}|$ 表示边 e_{ij} 的权重值，$p<v_i,\cdots,v_j>$ 表示顶点词串 $v_i,\cdots,v_j$ 所对应的路径，$ps<v_i,\cdots,v_j>$ 表示对应路径上所有边的集合的交集，也称为 $p<v_i,\cdots,v_j>$ 的集合，$len<v_i,\cdots,v_j>$ 为路径 $p<v_i,\cdots,v_j>$ 的长度，$weight<v_i,\cdots,v_j>=|ps<v_i,\cdots,v_j>|$ 表示 $p<v_i,\cdots,v_j>$ 的权重值。定义一个类交集运算 $\cap^s$，用于词共现有向图上边的集合交集运算：

$$X \cap^s Y=\{<sno,start,end>|<sno,start,mid>\in X,\ <sno,mid,end>\in Y\}$$

显然，$X \cap^s Y \neq Y \cap^s X$。因此，在进行类交集运算时，必须保证左操作数的边(或路径)尾顶点是右操作数的边(或路径)头顶点。

综上所述，在词共现有向图中，当路径 $p<v_i,\cdots,v_j>$ 所对应的词串满足 Regulation-1 的 3 个条件，则这个词串是合成词。

3.3 生成词共现有向图

由表 1 中各词串可得，$V=$ {产业，革命，知识，经济，造就，时代，世界，格局，产生，深远，影响}，$|V|=11$，则 Sample-1 对应的词共现有向图如图 1 所示。而利用文献[7]的算法，Sample-1 生成的词序列有向图如图 2 所示。通过比较可以发现，本文算法生成的词共现有向图规模精简得多，时空复杂度明显低于文献[7]的算法。

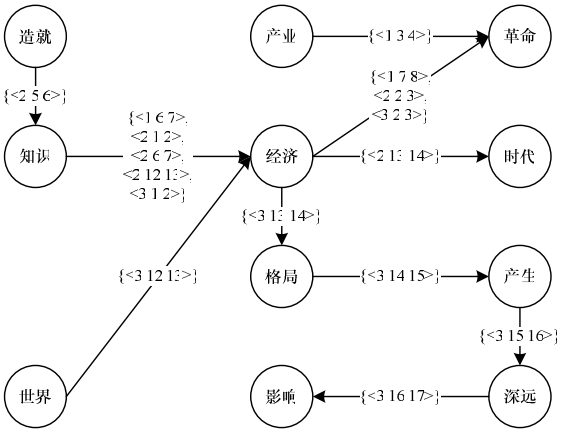

图 1　Sample-1 的词共现有向图

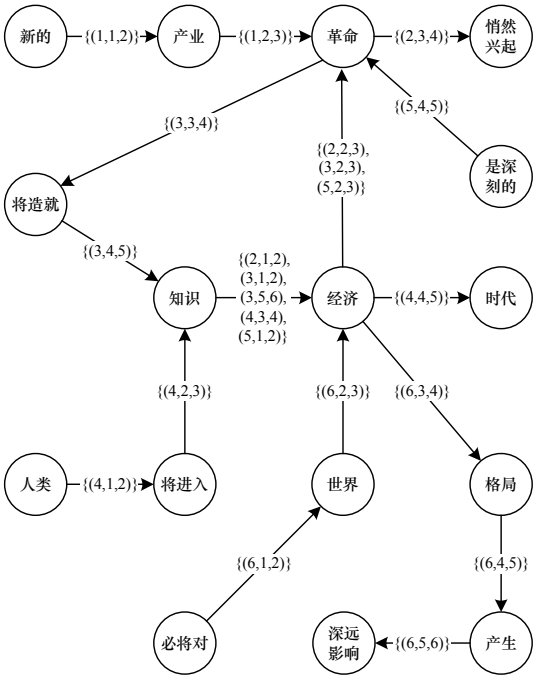

图 2　Sample-1 的词序列有向图

3.4　合成词提取算法

首先设定一个整数阈值 T，当：

$$weight<v_i,\cdots,v_j>=\left|ps<v_i,\cdots,v_j>\right|<T$$

时，认为 $p<v_i,\cdots,v_j>$ 所对应的词串不是合成词；当：

$$weight<v_i,\cdots,v_j>=\left|ps<v_i,\cdots,v_j>\right|\geqslant T$$

$$weight<v_{i-1}v_i,\cdots,v_j>=\left|ps<v_{i-1}v_i,\cdots,v_j>\right|<T$$

$$weight<v_i,\cdots,v_jv_{j+1}>=\left|ps<v_i,\cdots,v_jv_{j+1}>\right|<T$$

时，认为 $p<v_i,\cdots,v_j>$ 所对应的词串是合成词。

一般情况下设 $T=2$，根据这个原则，可以对图 1 进行化简，即删除 $w_{ij}=\left|s_{ij}\right|<T$ 的边，同时去掉孤立顶点，这个过程，本文称之为降维。图 1 降维后得到图 3。

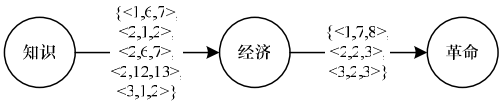

图 3　Sample-1 降维后的词共现有向图

借鉴 Bellman-Ford 算法思想，本文设计了求解词共现有向图中多源点路径长度最长($\geqslant L$)并且权重值满足给定条件($\geqslant T$)的路径算法，用于合成词提取。

算法步骤如下：

(1)置 $s=aMatrix(1,0)$，$d=\varnothing$，$path=s$，$len(path)=1$，$ps(path)=\varnothing$，$weight(path)=0$。

(2)设 s 为源点，在词共现有向图中搜索 s 的下一顶点 d，如果失败转第(3)步；否则，执行以下 4 步：

1)计算 $ps(path)=ps(path)\cap^{s}ps<s,d>$；

2)更新 $weight(path)$；

3)如果 $weight(path)\geqslant T$，置 $path=path\,\&\,d$，$len(path)=len(path)+1$，令 $s=d$，$d=\varnothing$。

4)转第 1)步。

(3)如果 $len(path)<L$，转第(7)步。

(4)存储提取到的合成词。

(5)在词共现有向图中删除已提取到的合成词所对应的路径信息。

(6)对词共现有向图执行降维操作。

(7)如果图不为空，转第(1)步。

(8)输出提取到的合成词，算法结束。

通过多次迭代搜索，算法总能找到图中满足 Regulation-1 的最长路径，最终可提取出合成词。

由于是长度优先，因此总是先将更长的合成词识别出，这样的好处在于，当一个合成词包含了另外一个合成词时，算法能先后识别出这 2 个合成词。若是权重值优先，则只能识别出长度较短的合成词。

3.5　合成词过滤

经统计分析发现，某些词不能出现在合成词的前面或后面，如“主义”、“时期”等不能是构成合成词的第 1 个词，“奠定”、“打下”等不能是构成合成词的最后一个词，因此采用首词删除和末词删除规则对提取到的合成词进行过滤。首先建立首词删除表和末词删除表，检测提取到的合成词的首词和末词是否包含在首词过滤表和末词过滤表中，若是，则将其删除，再对该合成词剩余部分采用 Regulation-1 的第(1)个条件进行判定，若满足条件，则保留，否则，放弃。

4　实验分析与比较

实验分 2 组进行：第 1 组使用复旦大学上海(国际)数据库研究中心 NLP 小组提供的约 20 MB 的文本集(含 1 600 篇政治经济类论文)，随机选取 10 篇文本，平均字数为 6 130；第 2 组使用 2003 年“863”评测的 10 篇文本，平均字数为 3 983。所有文本均采用中科院分词系统 ICTCLAS 进行分词标注。

4.1　实验结果分析

对提取到的合成词进行人工评价，对比原文，若合成词表达了一个完整的概念，则标识为正确的，否则为不正确。

2 组实验结果如表 2 所示。可以看出，合成词提取正确率均达到了 90%以上，总体平均正确率为 91.16%，可见，本文算法在不同类型的数据集上均能取得较高的正确率。

表 2　2 组数据集的合成词提取结果

数据集	文本总字数	提取总数	正确数	正确率
NLP 小组文本集	61 309	726	670	0.922 9
“863”评测文本集	39 831	321	289	0.900 3

对提取不正确的合成词进行统计分析，发现合成词不正确(共 88 个)的情况主要有 3 类：(1)合成词不能表达一个完整的概念，这种情况最为常见，如“积极财政”、“参与国际”、“收入家庭”。这种情况共发生 46 次，占所有不正确合成词

的 52.27%。(2)提取到的合成词不符合构词规则，即提取到的合成词中“动词+…+动词”结构的合成词基本上都不正确，这种情况共发生 19 次，占 20.60%，如“处在/v 工业化/vn”、“建设/vn 提供/v”、“增长/vn 培育/v”。(3)分词系统将某些成语、缩略词等分词标注后，其中的某个原子词词性标注为过滤词性，如提取到的“大物博”，原文为“地大物博”，分词系统分词标注为“地/u 大/a 物/ng 博/ag”，算法将“地/u”过滤掉了，因此，无法正确提取，这种情况共发生 12 次，占 13.64%。第(1)种情况较难解决，这是本文算法一个较大的缺陷；第(2)种情况可以通过增加合成词过滤规则来解决；第(3)种情况解决起来并不困难，但是如果将过滤词性纳入提取范围，会带来一些其他的问题，如词共现有向图的规模急剧增大以及正确率下降。

4.2 与其他算法的比较

由于文献[1-3]算法的实质与本文算法一致，因此将四者进行合成词提取性能的比较实验，结果如图 4 所示。从中可以看出，本文算法的性能最优越。

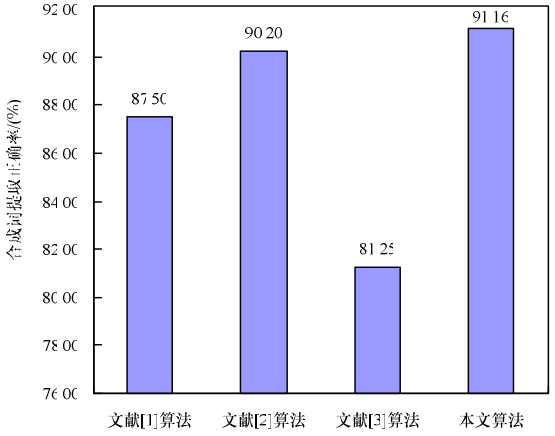

图 4 4 种算法提取性能比较

5 结束语

本文方法的合成词提取平均正确率为 91.16%。由于合成词的提取是建立在分词系统之上，而且是通过词性探测获得词串的，因此分词标注的准确性对合成词的正确提取有较大的影响。另外，本文算法不能提取出权重值小于 T 的合成词，T 值的设定对合成词的提取也有较大的影响，增大 T 可提高正确率，但会降低召回率。

下一步的工作主要有：(1)采用语义识别对合成词进行过滤，解决提取到的合成词符合判定规则但不能表达一个完整概念的问题，进一步提高合成词提取正确率。(2)优化参数 T 的设置，在正确率和召回率之间取得最佳平衡。(3)继续完善算法，提高其运行效率。

参考文献

[1] 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.

[2] 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.

[3] 周 蕾, 朱巧明. 基于统计和规则的未登录词识别方法研究[J]. 计算机工程, 2007, 33(8): 196-198.

[4] 李 荣, 郑家恒, 郭梅英. 基于遗传算法的隐马尔可夫模型在名词短语识别中的应用研究[J]. 计算机科学, 2009, 36(10): 244-246.

[5] 韦小丽, 孙 涌, 张书奎, 等. 基于最大熵模型的本体概念获取方法[J]. 计算机工程, 2009, 35(24): 114-116, 120.

[6] Erik F, Sang Tjong-Kim. Noun Phrase Recognition by System Combination[C]//Proceedings of ANLP2NAACL'00. Seattle, USA: [s. n.], 2000: 50-55.

[7] 陈建超. 基于海量互联网网页文本的中文概念知识库构建算法研究及应用[D]. 广州: 华南理工大学计算机科学与工程学院, 2009.

编辑 张 帆

(上接第 176 页)

由实验数据可知，在进行多匝道调节联合控制后，交通流运行顺畅、平稳，各区间均未发生交通阻塞现象，车流量也保持在理想水平，且入口匝道也不会发生回溢现象，达到了预期控制效果。

6 结束语

本文分析了多匝道调节联合控制交通流的特性，确立了多匝道调节联合控制模型，在保证交通流较大的情况下考虑了入口匝道排队长度问题，并基于蚁群算法求解此协调控制问题。结果表明，主线交通流量在联合控制作用下得以顺畅运行，且各入口匝道也未产生回溢现象，达到了预期控制目标。对于目前城市快速路交通流量飞速增长的现象，今后应当继续加强多种动态交通控制措施的协调研究，建立更加适合国内快速路的交通流模型。

参考文献

[1] 侯忠生, 晏静文. 带有迭代学习前馈的快速路无模型自适应入口匝道控制[J]. 自动化学报, 2009, 35(5): 588-589.

[2] 赵忠杰, 周林英, 王英伟. 基于免疫算法的多匝道 MF-AILC 控制方法[J]. 系统仿真技术, 2009, 5(1): 40-44.

[3] Hou Zhongsheng, Xu Jianxin, Zhong Hongwei. Freeway Traffic Control Using Iterative Learning Control Based Ramp Metering and Speed Signaling[J]. IEEE Transactions on Vehicular Technology, 2007, 56(2): 466-477.

[4] Andreas H, Schutter B, Hellendoorn H. Model Predictive Control for Optimal Coordination of Ramp Metering and Variable Speed Limits[J]. Transportation Research Part C: Emerging Technologies, 2005, 13(3): 185-209.

[5] 刘智勇. 智能交通控制理论及其应用[M]. 北京: 科学出版社, 2003.

[6] 史忠科, 黄辉先, 曲仕茹, 等. 交通控制系统导论[M]. 北京: 科学出版社, 2003.

[7] 陈 俊, 沈 洁, 秦 玲. 蚁群算法求解连续空间优化问题的一种方法[J]. 软件学报, 2002, 13(12): 2317-2323.

[8] 池元成, 蔡国飙. 基于蚁群算法的多目标优化[J]. 计算机工程, 2009, 35(15): 167-169.

编辑 顾逸斐