基于典型句型的词语搭配定量分析及提取算法

王 璐 张仰森
(北京信息科技大学 北京 100192)

摘 要 在分析现有的词语搭配自动提取算法的不足后,提出了一种新的词语搭配提取算法,尝试从非结构化语言知识到结构化语言知识的转化。基于词语搭配的语言学知识,构建了基于典型句型的词语搭配模型,其以动词、名词及形容词为中心词分类搭配,以实词为主干提取搭配,利用共现频率及互信息等统计学模型在大规模语料库中进行筛选,固化这些搭配知识,构建搭配知识库。

关键词 词语搭配,典型句型,互信息,搭配数据库

Quantitative Analysis and Extracting Arithmetic of Collocations Basic on Typical Patterns

WANG Lu ZHANG Yang-sen
(Beijing Information Science and Technology University, Beijing 100192, China)

Abstract The shortcoming of the existing automatic extraction algorithm was analyzed, and a new model was proposed, trying to transform unstructured language knowledge into structural language knowledge. The language knowledge was introduced to a extraction model based on typical patterns, and collocations were classed by noun, verb and adjective as center, and by substantive as backbone. Then, concurrence frequency and MI etc were used to screen in large-scale corpus. Finally, this knowledge was solidified to build collocation database.

Keywords Collocation, Typical patterns, MI, Collocation database

1 概述

词语搭配是传统语言学研究的一个基本领域,但在智能语言信息处理方面还是一个新兴的热门领域。一切自然语言处理系统归根结底都是基于知识系统,而搭配知识库是其中重要而不可缺少的部分。由此,如何简单、有效、全面地提取词语搭配成为了关键问题。另外针对中文而言,词与词的搭配只要满足意义逻辑的要求就可以搭配,不像英文还有时态、单复数等的变化。词语搭配问题在中文信息处理中显得尤为重要。

长期以来,词语搭配的研究都属于传统语言学领域,也取得了很多成果,例如《现代汉语实词搭配词典》(商务印书馆,1992)等。但是 Smadja[1] 在 1993 年就指出,传统的手工编辑的 Oxford English Dictionary(OED)的准确率大约只有 4%,效率与自动获取方法差距偏大;此外传统方法受人为因素干扰过大,使相互间的协调发展相当困难;而且即使是最资深的专家,在信息量急剧膨胀的现代也难以全面掌握语言发展全貌。语料库语言学的兴起为词语搭配自动获取和知识库的研究开辟了有力的基础和广阔的前景。

2 对词语搭配的相关研究

在自然语言处理领域,对于什么是搭配,目前尚未有一个权威性的定义。其中最具影响的是美国宾州大学 Benson 教授[2] 在编撰 BBI Combinatory Dictionary of English(1985,1986,1989,1990)中给出的关于搭配的定义:搭配是一种具有任意性的、重复出现的词的组合(A collocation is an arbitrary and recurrent word combination)。清华大学的孙茂松教授对其进行了补充:搭配通常是具有一定结构的,同时搭配是与领域有关的[3]。

国外最早开始搭配的计算机定量分析的是 choueka 等[4](1983)。他们从《纽约时代周刊》(New York Times)约 1100 万词的文本中提取了数以千计的英语常用的搭配,如 fried chicken, home run, Magic Johnson 等。Smadja[1](1993)的 Xtract 系统是迄今为止关于搭配定量分析的最新、最完整的工作。在一个规模为一千万词的股票市场新闻报告语料库上运行 Xtract 所得到的结果显示,搭配提取的准确率达到了 80%(如果不诉诸词性自动标注技术,准确率约为 40%)。至于国内方面,孙茂松等[3](1997)提出的包过强度、离散度及尖峰 3 项指标在内的搭配定量评估体系,以一个约 710 万词的新华社语料库为工作平台,达到了约为 33.94% 的准确率,是最早、最全面的搭配提取工作。曲维光等[5](2004)提出了一种基于框架的词语搭配抽取方法,其应用大规模分词和词性标注语料,引入相对词序比的方法进行筛选,抽取的平均准确率达到了 84.73%,在准确率上有了很大的提高。

研究现有中文词语搭配抽取方法发现,以下方面有待改进:

本文受国家自然科学基金(60873013,61070119),北京大学计算语言学教育部重点实验室开放课题基金(KLCL-1005),北京市属市管高等学校人才强教计划项目(PHR201007131)资助。

王 璐(1987-),女,硕士生,主要研究方向为智能信息处理;张仰森(1962-),男,教授,硕士生导师,主要研究方向为人工智能、中文信息处理。

(1) 实验所用的语料，大多只经过分词处理，没有经过词性标注，缺失了词语搭配要利用的重要语言信息，没有充分利用大规模语料库资源；

(2) 抽取搭配词汇的同时，没有充分考虑搭配的结构信息；

(3) 搭配抽取窗口的单纯定义，缺乏理论依据，尚待实践，且候选集过大增加了算法复杂度；

(4) 搭配抽取方案中没有充分利用语言学知识；

(5) 国内对词语搭配的研究多属于一些方法性研究，很多有很好试验结果的方法只适用某些特定词类之间，缺乏进一步扩展全面适用的工作，不利于下一步构建搭配知识库。

本文针对目前中文词语搭配自动抽取方法中存在的问题，做出了相应的改进，提出了一个新的词语搭配提取模型。通过在大规模语料上的实验，验证了该模型的有效性。

3 词语搭配的语言学知识引入

对于搭配来讲，可以通过引入相关语言学知识来进一步筛选各搭配中符合语言学规律的词语。由于本文用到的训练语料经过分词和词性标注，因此可以将词语搭配中词性的组合性限定规则用于词语搭配的自动抽取，以进一步利用语言知识。

搭配主要在实词间进行，本文选取名词(用 N 表示)、动词(用 V 表示)、形容词(用 A 表示) 3 类作为中心词构建模型。通过语言知识，知道一个实词能够与哪些实词搭配具有突出的选择性特征。在汉语中很少有搭配能力完全相同的实词，因此研究汉语实词搭配的最佳方法是逐词具体描写。利用现有的语言学资源《现代汉语搭配词典》和《现代汉语实词搭配词典》的相关知识和要求进行整理、整合，经人工筛选用于后续的词语搭配中。

例如，《现代汉语实词搭配词典》中采用下列的搭配框架描写现代汉语实词搭配的状貌：

名词搭配框架

名词 míngcí

〈名〉释义

[主]①～＋动；②～＋形；

[宾]动＋～；

[中]①名＋～；②动＋～；③形＋～；④数量＋～；

[定]～＋名；

动词搭配框架

动词 dòngcí

〈动〉释义

[主]①～＋动；②～＋形；

[谓]①名＋～；②动＋～；③形＋～；④～＋名；⑤～＋形(宾)；⑥～＋形(补)；⑦～＋数量；⑧能愿＋～；

[宾]动＋～；

[中]①名＋～；②动＋～；③形＋～；

[定]～＋名；

[状]～＋动；

形容词搭配框架

形容词 xíngróngcí

〈形〉释义

[谓]①名＋～；②动＋～；③～＋形；④～＋数；⑤能愿＋～；

[宾]动＋～；

[补]动＋～；

[中]形＋～；

[定]～＋名；

[状]～＋动；

本模型不需要使用语法关系，对以上的搭配框架进行整合补充得到了本文的基于典型句型的词语搭配模型。

4 基于典型句型的词语搭配模型

4.1 词语搭配的粗选模型

词语搭配具有结构性，表现在组合关系上，就是词与词之间存在相对固定的结构和位置。在孙茂松等[3](1997)提出的搭配定量评估体系中，利用离散度及尖峰两项指标衡量词语搭配的结构性取得了一定的效果，但离散度和尖峰要在强度信息不足以据之做出裁决的时候才能突出作用。也就是说，结构信息是从属于强度信息之下的，而这与词语搭配的定义要求不符；同时，对于一些高强度但结构信息不明显的词语搭配无效。

整合利用《现代汉语实词搭配词典》中的搭配框架，这里不需要使用语法关系，只需简化词性之间的关系。而搭配主要是实词间进行，实词主要由名词、动词、形容词组成，以它们为中心词主导搭配，能更好地组织、存储搭配以及进行后期的处理。

本文抽取以名词、动词和形容词为中心词的语义搭配属性对：

1)名词

名词；动词；形容词；量词。

2)动词

副词；动词。

3)形容词

副词；形容词。

以名词为中心词的词语搭配知识表示：

S－〉N＋N|V＋A＋N|A＋N|Q＋A＋N|N＋V|N＋Q

A－〉NULL|A|A＋D|A＋A

以动词为中心词的词语搭配知识表示：

S－〉V＋V|F＋V

以形容词为中心词的词语搭配知识表示：

S－〉A＋A|F＋A

S：词语搭配；N：名词；V：动词；A：形容词；Q：量词；F：副词；D：的。

说明：对于名词和形容词以及名词和动词的搭配在以名词为中心词的搭配中已提取，后面就不再重复提取。本模型对一些复杂搭配以及现代汉语中变化的语法现象不做讨论，例如涉及成语的搭配或者副词修饰名词等搭配。通过该模型抽取搭配候选集，再通过以下的统计学模型进行筛选，得到最终的搭配。

4.2 词语搭配的统计筛选模型

本文在前期的词语搭配的粗选模型基础上利用同现概率及互信息在设定的基于典型句型的模型下抽取搭配。第一次将统计学信息从属于结构信息，在典型句型筛选的基础上进一步筛选。

互信息定义：

$$I(x,y)=\log_2 \frac{P(x,y)}{P(x)\times P(y)}$$

式中，$P(x,y)$ 表示词语 x 和词语 y 在语料中的共现频率；

$P(x)$，$P(y)$分别表示 x，y 在语料中各自出现的频率。针对Benson教授对词语搭配的定义中的两个特性做如下说明：

重复性：x，y 共现次数越多，$P(x,y)$越大，$I(x,y)$亦随之越大，表明重复性越强。反之，则重复性越弱。

任意性：x，y 受约束程度越深，意味着 x，y 与其它词的共现机会越少，即 $P(x)$，$P(y)$的值减少，在 $P(x,y)$值不变的情况下，会使 $I(x,y)$的值变大，表明 x，y 的任意性加强。反之，则表明任意性减弱。

本文选取 $P(x,y)=4$，$I(x,y)=4$ 为阈值，大于该阈值的候选集为搭配。例如，中心词为“能力/n”的互信息最大的前10个搭配，如表1所列。

表1　中心词为“能力/n”的互信息最大的前10个搭配

搭配	搭配词性	中心词	中心词词性	互信息
养家	vn	能力	N	11.6749112434186
放空	Vi	能力	N	11.5463114030111
抗寒	V	能力	N	11.5463114030111
绸缎	N	能力	N	11.5463114030111
信贷资金	N	能力	N	11.5143327240386
抗病	Vn	能力	N	11.5143327240386
应变	Vn	能力	N	11.1924046291512
偿债	Vn	能力	N	11.0992952247598
作战	Vn	能力	N	10.8135952818462
模拟	Vn	能力	N	10.7133853912332

5　搭配知识库构建

高准确率的词语搭配算法提取出了大量的高质量词语搭配，如何有效存储固化，为后期其他自然语言处理做知识库来源或者下一步的语义分析，就涉及到数据库的建设和设计。本数据库中现在只涉及到字词级的搭配数据，也很好地考虑了后期语义搭配的扩展。表的设计：现在只局限于保存提出的字词级搭配，故只有一个表：字词级搭配。其含有5个属性：中心词 、中心词词性、搭配、搭配词性、互信息。

	中心词	中心词词性	搭配	搭配词性	互信息
1	世纪	n	新	/a	7.10398886725415
2	宗教	n	引导	/v	9.79314986949061
3	精神	n	创造	/vn	4.57986417863636
4	水平	n	高	/a	6.15280918948913
5	甬道	n	宽	/a	10.9215295841765
6	灯	n	腾	/v	13.9399752514973
7	孩子	n	失明	/vi	10.7190990626083
8	文化	n	钱币	/n	6.09661247297844
9	城	n	模范	/n	11.4744219625393
10	农田	n	占	/v	9.39323207300667
11	干涉主义	n	是	/vl	5.25977088150756
12	全人类	n	走向	/v	7.98190593196326
13	设备	n	钻井	/vn	10.4537909052544
14	制度	n	税收	/n	4.23595862038016
15	达到	v	能够	/vu	4.21755275819846
16	深化	v	继续	/v	6.99904443344949
17	业绩	n	光辉	/a	11.8058919768386
18	能力	n	有	/vx	4.27076201430414
19	巢	n	区分	/v	14.8050456714112
20	情况	n	布	/n	6.44812215516766

图1　数据库示例

设置说明：数据库以中心词为主干存储，避免重复也为后期语义知识合并提供主干。存储中心词及搭配的词性是为了更好地提取语义信息以及后期的按类统计等工作。互信息是其重要的统计学信息，它一定程度上覆盖了共现概率，所以本数据库不再存储共现概率。中心词和搭配的顺序是按照典型句型直接收录的，一般情况下的顺序为搭配在中心词前，除了以名词为中心词、动词或量词为中心词两种情况下，中心词在前而搭配词在后，但该搭配改变顺序也是正确的，故本数据库没有考虑顺序的问题，一律存为中心词和搭配。共收集了144812个搭配，示例如图1所示。

6　实验及结果

6.1　语料的准备

本文使用的训练语料为北京大学计算语言学研究所研制的2600多万字的《人民日报》基本标注语料库。它对全部语料已完成词语切分和词性标注等基本加工。该项成果通过了合作单位Fujitsu的验收，基本满足词语搭配抽取的需要。

6.2　实验结果

本实验共获得144812个搭配，对每种搭配类型都取一个中心词为例进行试验，结果如下：名词与名词搭配(N+N)，以中心词为“企业”为例，共抽取73个候选搭配，经人工校对，71个正确，准确率为97.2%。抽取的搭配如表2所列。

表2　名词与名词搭配(N+N)，中心词为“企业”的搭配

公有制	全资	精英	商业	集体	骨干	龙头	外资	乡镇	股份制
名牌	蚕丝绸	下属	纺织	工业	特大型	煤炭	高新技术	军工	所
有制	重点	陶瓷	粮棉	汽车	经营性	高能耗	基层	低效	私人
科技型	集团型	经营型	活力	竞争性	渔业	渔机	水产	家族	林
业	台资	工厂	同类	服装	钢铁	保障性	扫描仪	股	混合型	机
械	加工型	水泥	股权	公司制	餐饮	鞋	内资	类型	制造业	省
内	劣势	华北	松散型	建筑业	老字号	软件	出口型	主力	酒
家	国有制	合作制	侨资	零部件	中央				

名词与动词搭配(N+V)，以中心词为“企业”为例，共抽取235个候选搭配，经人工校对，220个正确，准确率为93%。抽取的搭配如表3所列。

表3　名词与动词搭配(N+V)，中心词为“企业”的搭配

集资	亏损	新建	召开	重组	垄断	投资	国有	脱钩	改革	专营
办	出口	生产	属	加工	瘫痪	面临	合资	受淹	撤销	私有
分担	海运	兼并	控股	出钱	采购	夯	争光	经营	参股	扣除
摸清	陷入	投产	供电	离退休	免	补发	变相	免税	剥离	赚
出资	干预	认证	融资	介绍	插减轻	脱困	改制	背{bei1}	半停
产	需要	添	送货	破产	施工	试行	运输	挖掘	附营	减员	卖
预期	改组	逃税	铲除	赋予	应当	着力	不惜	私有化	发电
经销	靠拢	富余	唱戏	缓解	改造	淘汰	增添	注入	村办	兴办
制药	增资	超群	注册	伴随	促进	扶植	插手	亦可	简化	审
批	拆除	挽回	并存	赞助	注意	协办	孵化	信息化	明晰	迁出
捕捞	介入	增效	免征	成长	大于	为由	推销	享受	征收	印
染	借款	深加工	联营	摘垮	抵押	依赖	剩余	冻结	倾听	购销
摊派	扭亏	炼铁	炼钢	冶炼	焕发	限于便利	创办	挣钱	输送
紧跟	巡礼	推介	担	赢利	互通	下岗	自营	转换	摆脱	促使
运营	面对	开行	拖欠	认定	再造	遏制	交费	转制	签约	减
负	核销	讲明	谋取	无望	虚开	濒临	争	印刷	吸纳	陷于	看
好	卸掉	带来	听听	解困	亟须	创建	妨碍	获利	榨油	转产
确保	迈进	集团化	排污	明示	涌入	放开	关闭	亏	停产	迁建
代发	抢滩	植	应对	送给	分割	培育	达标	承揽	衡量	缴清
苦练	领取	检查	负债	自愿	破解	雇用	获	查阅	验证	听取
核查	连任	熟悉	串通	划归	谋划	见长{jian4zhang3}	寻呼	定
价	投标	西进	参展	制假	累计	贴近	识别	招标	喜获	责令

名词与形容词搭配(N+A)，以中心词为“企业”为例，共抽取11个候选搭配，经人工校对，10个正确，准确率为90%。

(下转第270页)

1) 针对数据的不确定性特点，提出一种基于元组存在性的概率数据模型。

2) 结合基于元组存在性的概率数据模型，提出相关的概率关系代数和概率数据库查询处理算法。

参 考 文 献

[1] Imielinski T, Lipski W. Incomplete information in relational databases[C]//Journal of the ACM. October 1984, 31: 761-791

[2] Dalvi N, Suciu D. Management of Probabilistic Data: Foundations and Challenges[C]//PODS. 2007: 1-12

[3] Benjelloun O, Sarma A D, Halevy A, et al. ULDBs: Databases with uncertainty and lineage[C]//VLDB. 2006: 953-964

[4] Sen P, Deshpande A. Representing and querying correlated tuples in probabilistic databases[C]//ICDE. 2007

[5] Cavallo R, Pittarelli M. The theory of probabilistic databases[C]// Proceedings of 13th Int. Conf. on Very Large Data Bases (VLDB'87). Brighton, England, September 1987: 71-81

[6] Cheng R, Kalashnikov D, Prabhakar S. Evaluating probabilistic queries over imprecise data[C]// International Conference on Management of Data. 2003

[7] Garofalakis M, Suciu D, et al. Probabilistic Data Management [J]. IEEE Data Engineering Bulletin, 2006, 29(1)

[8] Choenni S, Blok H E, Leertouwer E. Handling uncertainty and ignorance in databases: A rule to combine dependent data[C]// Database Systems for Advanced Applications. 2006

[9] Xia Yu-ni, Prabhakar S, Lei Shan, et al. Indexing continuously changing data with mean-variance tree[C]// ACM Symposium on Applied Computing. 2005

(上接第 234 页)

抽取的搭配如表 4 所列。

表 4 名词与形容词搭配(N+A)，中心词为“企业”的搭配

困难	知名	像样	赚钱	单一	雄厚	著名	微小	吃力	正规	不错

名词与量词搭配(N+Q)，以中心词为“企业”为例，共抽取 4 个候选搭配，经人工校对，3 个正确，准确率为 75%。抽取的搭配如表 5 所列。

表 5 名词与形容词搭配(N+A)，中心词为“企业”的搭配

家	户	批	一下子

动词与动词搭配(V+V)，以中心词为“突破”为例，共抽取 16 个候选搭配，经人工校对，15 个正确，准确率为 93%。抽取的搭配如表 6 所列。

表 6 动词与动词搭配(V+V)，心词为“突破”的搭配

可	创新	取得	获	快攻	予以	无	有所	分区
寻求	亩产	加以	求	鉴定	争取	可望		

动词与副词搭配(V+F)，以中心词为“突破”为例，共抽取 8 个候选搭配，经人工校对，8 个正确，准确率为 100%。抽取的搭配如表 7 所列。

表 7 动词与副词搭配(V+F)，中心词为“突破”的搭配

首先	已	轻易	重点	一举全力	难以	由此

形容词与形容词搭配(A+A)，以中心词为“繁荣”为例，共抽取 4 个候选搭配，经人工校对，3 个正确，准确率为 75%。抽取的搭配如表 8 所列。

表 8 形容词与形容词搭配(A+A)，中心词为“繁荣”的搭配

全面	稳定	空前	繁荣

形容词与副词搭配(A+F)，以中心词为“繁荣”为例，共抽取 4 个候选搭配，经人工校对，4 个正确，准确率为 100%。抽取的搭配如表 9 所列。

表 9 形容词与副词搭配(A+F)，中心词为“繁荣”的搭配

更加	长期	共同	真正

将各个类型的准确率统一计算，则：

平均准确率=(71+220+10+3+15+8+3+4)/(73+235+11+4+16+8+4+4)=94.1%

相对于基于框架的提取方法准确率为 83.73%，本文的方法不仅在准确率上有了很大的提高，而且在全面性上也有了很大的突破。通过对以名词、动词、形容词为中心词的主要实词的统计，基本提取了各种搭配类型，这些信息对于后期建立搭配知识库有很大的作用。

结束语 本文提出了一种全新的基于典型句型的词语搭配自动提取模型。引入语言学知识，利用标注词性的大规模语料库，构建基于典型句型的模型，利用统计量共现频率及互信息筛选候选搭配，获得了 94.1%的准确率及较高的召回率。

分析错误提取的词语搭配，主要有以下几个方面：

1)构建典型句型考虑的主要是两个词，对于单个词考虑不够。例如“企业”+“添”及“企业”+“赚”。

2)只考虑了简单句型，对复杂句型引起的局部句型相同从而产生的错误考虑不够。

3)对语言学知识的运用还很粗浅，只利用了词与词之间的搭配限制，还没有利用相关的语义制约关系。

下一步工作：(1) 提高分词与词性标注的性能，为词语搭配研究提供质量更好的语料。(2) 对词性相同的搭配，典型句型的限制要进一步加强，限制其结构。(3) 对两个词以上的搭配以及短语组块间的搭配进行深入研究，寻找更好的解决方案。(4)引入语义信息，加强词语搭配的融合，扩张数据库的实用性。

参 考 文 献

[1] Smadja F. Retrieving Collocations from Text: Xtract[J]. Computational Linguistics, 1993, 19(1): 143-177

[2] Benson M, Benson E, Ilson R. The BBI Combinatory Dictionary of English: A Guide to Word Combinations[M]. John Benjamins, Amsterdam and Philadelphia, 1986

[3] 孙茂松，黄昌宁，方捷. 汉语搭配定量分析初探[J]. 中国语文，1997(1)：29-38

[4] Choueka Y, Klein T, Neuwitz E. Automatic Retrieval of Frequent Idiomatic and Collocational Expressions in a Large Corpus [J]. Literacy and Linguistic Computing, 1983, 4(1): 34-38

[5] 曲维光，陈小荷，吉根林. 基于框架的词语搭配自动抽取方法[J]. 计算机工程，2004，30(23)：22-24