Web数据反馈的搭配抽取方法

林建方[1]，牛 成[2]，李 生[1]，郑德权[1]

（1. 哈尔滨工业大学 语言语音教育部－微软重点实验室，哈尔滨 150001，jflin@mtlab.hit.edu.cn；
2. 微软亚洲研究院，北京 100080）

摘 要： 为了提高搭配(Collocation)抽取的精度，提出一种新的互联网数据的搭配抽取方法. 传统的搭配抽取统计方法都是基于语料库的，常受到语料库规模的影响和制约，而在互联网数据中蕴含着丰富的知识和信息，基于Web的词汇相关性度量方法，充分利用搭配在谷歌中的页面数模拟其对应语料库的词频数，并分别选取共现频率、互信息、卡方检验3种经典统计关联度量方法. 实验结果表明召回率、精确率均好于对应的基于语料库的方法，这说明互联网中大量数据应用于自然语言处理各种任务的可行性.

关键词： 搭配；共现频率；互信息；卡方检验；语料库；Web
中图分类号： TP391 **文献标志码：** A **文章编号：** 0367－6234(2010)02－0281－05

Automatic collocation extraction using web feedback data

LIN Jian-fang[1], NIU Cheng[2], LI Sheng[1], ZHENG De-quan[1]

(1. MOE-MS Key Laboratory of Natural Language Processing and Speech, Harbin Institute of Technology, Harbin 150001, China, jflin@mtlab.hit.edu.cn; 2. Microsoft Research Asia, Beijing 100080, China)

Abstract: To improve the precison of collocation extraction, this paper proposes a new method based on Internet data. For the constraint by the corpus scale for traditional collocation extraction approach based on linguistic corpus, we acquire collocations from Web, which contains plenty of information and knowledge. Three classical association measures of co-occurrence frequency, mutual information and χ^2-test are used to automatically extract the collocation. Based on the experimental results, the benchmarks show that the performance of this new Web-based approach is superior to that of traditional approach in both precision and recall. Thus the data from Internet may be applied in many NLP applications.

Key words: collocation; co-occurrence frequency; mutual information; χ^2-test; corpora; web

搭配是两个或多个词组合在一起表达某种特殊含义的词汇现象，它不能随意地更改和替换搭配中的组成部分[1－2]. 随着互联网数据和大规模语料库成为计算语言学研究的重要知识来源，搭配的自动抽取在机器翻译、信息检索、词义消歧，句法分析，自然语言生成等应用任务中得到广泛应用[3－6]. Choueka等认为搭配是重复的相邻词串，他只用N－grams(2到6)方法进行搭配抽取，并没有考虑到搭配的两个词间距的情况，Church和Hanks使用互信息来测量相互关联的两个单词，然而没有考虑两个词的语法制约关系，导致如“teacher-student”常共现的词组被抽取出来，Smadja的Xtract系统是关于搭配定量分析经典工作，提出了度量词语对之间搭配强度的计算公式，引入了位置信息以及相关统计数据分布的离散度计算公式，并集成了语料库语言学中词性自动标注技术. Lin(1998)使用浅层句法分析得到依赖三元组，通过互信息进行搭配抽取. 近年来国内的学者借鉴国外的方法，也开始进行汉语词语搭配的自动获取研究[7－8]. 本文定义搭配为具有一定句法关系的惯用词对，例如“<失去(lose)～OBJ～控制(control)>”是一个具有动宾关系的汉语搭配. 通常搭配抽取分两步进行[9]：1)根据词法

收稿日期： 2008－12－17.
基金项目： 国家自然科学基金重点资助项目(60736044)；国家科技发展计划探索类资助项目(2006AA01Z150).
作者简介： 林建方(1978—)，男，博士研究生；
李 生(1943—)，男，教授，博士生导师.

(如词性标注)、句法(如浅层句法分析)或语义等语言学知识从语料库中抽取搭配词语候选集;2)利用某种统计测度评价方法,计算候选搭配集合的关联权重并排序. 搭配抽取的统计方法主要分为:1)基于共现关系的方法(例如基于相对和绝对共现频率);2)基于信息论的方法(例如互信息,熵);3)基于统计的方法(例如卡方检验, t 检验).

本文据此提出一种基于 Web 数据反馈的搭配抽取方法,即根据候选搭配及构成词在谷歌出现的页面数模拟在语料库出现的次数,并分别选用 3 种经典搭配抽取方法,为共现频率、互信息、卡方检验,试验结果表明 Web 的方法效果好于基于语料库的方法.

1 关联度量方法描述

1.1 共现频率法

对于共现频率法很多研究都是基于"滑动窗口"的方法[10-11],即以当前词为中心,在一定大小的窗口范围内(5 ~6 个词)选择搭配候选词对,尽管这些单词之间关联性较大,并可抽取出远距离搭配,如:"file a lawsuit"、"file a class file action lawsuit". 然而由于没有考虑到搭配的结构性导致如"doctor-nurse"、"teacher-student"等非搭配被抽取出来. 这里是根据浅层句法分析后的结果进行共现频率统计,例如"go ~ verb:Mod:adv ~ only",计算"only go"或"go only"在整个语料库中的出现词数,对于每种句法关系都得到按词频大小降序排列的候选搭配词汇表,试验表明这样可以大大消除噪音.

1.2 互信息法

两个单词的互信息越大,其关联性越大,当单词频率较低时互信息更大. 然而互信息倾向于对小概率事件赋予较大的值,同时对于稀疏数据存在过评价问题,本文采用加权互信息方法计算句法分析后的依存三元组单词之间的值为

$$\mathrm{WMI}(w_1,r,w_2) = p(w_1,r,w_2)\log\frac{p(w_1,r,w_2)}{p(w_1\mid r)p(w_2\mid r)p(r)}. \quad (1)$$

式(1)通过加上一个加权因子 $p(w_1,r,w_2)$ 来调节上述情况,给定一个阈值,大于该阈值的三元组才能作为候选搭配,其中各项计算为

$$p(w_1,r,w_2) = \frac{\mathrm{count}(w_1,r,w_2)}{N}p(r) = \frac{\mathrm{count}(*,r,*)}{N},$$

$$p(w_1\mid r) = \frac{\mathrm{count}(w_1,r,*)}{\mathrm{count}(*,r,*)}p(w_2\mid r) = \frac{\mathrm{count}(*,r,w_2)}{\mathrm{count}(*,r,*)}.$$

式中: $\mathrm{count}(w_1, r, w_2)$ 为 head 为单词 w_1, Relation Type 为 r, Modifier 为单词 w_2 所有搭配个数, $\mathrm{count}(*, r, *)$ 为所有满足 r 关系的所有搭配的个数, $\mathrm{count}(w_1, r, *)$ 为 head 为单词 w_1, 在 r 关系情况下,所有与其有共现关系搭配的个数, $\mathrm{count}(*, r, w_2)$ 为 Modifier 为单词 w_2, 在 r 关系情况下,所有与其有共现关系的搭配的个数, N 为整个语料库中抽取出的英语三元组个数, $*$ 为任意的单词或关系类型.

1.3 卡方检验法

假设检验主要通过检验某一样本的平均数与正太分布总体的平均数之间的差异判断共现的词语是否能构成搭配,卡方检验法的意图在于,考察二元组构成词频分布关系,计算 n 个独立变量的观测值与期望值的差异之和,以判断二元组相互独立的原假设是否成立,对于二元组关系的卡方检验为

$$\chi^2 = \sum_{i,j}\frac{(c_{i,j}-E_{i,j})^2}{E_{i,j}} = \frac{N(c_{11}c_{22}-c_{12}c_{21})^2}{(c_{11}+c_{12})(c_{11}+c_{21})(c_{12}+c_{22})(c_{21}+c_{22})}. \quad (2)$$

式中: $c_{12} = \mathrm{count}(w_1)$, $c_{21} = \mathrm{count}(w_2)$, $c_{11} = \mathrm{count}(w_1,w_2)$, $c_{22} = N - c_{11} - c_{12} - c_{21}$.

卡方值反映的是共现的词语间多大程度上存在典型搭配关系,它给研究者提供的是一种把握性,对卡方值高的共现序列,研究者便有足够的把握确定其为显著搭配. 例如在置信水平 $\alpha = 0.05$, $\chi^2 < 3.841$ 时,这两个单词是完全独立的,否则二元组相关,可以构成搭配.

2 搭配抽取流程及评价

2.1 搭配抽取的流程

抽取搭配的流程图和算法如图 1 和算法 1 所示.

算法 1 英语搭配知识的获取.

步骤 1 语料预处理. 对语料库中的文本进行预处理,去掉杂质,把文章中的句子做断句处理,仅保留长度 >25 个单词的句子.

步骤 2 浅层句法分析. 使用 NLPWIN 对语料库中的句子做浅层句法分析,因此本文中的搭配都表示成依存三元组(triple),它们表示句子中词

和词之间依存句法关系.

步骤 3 候选搭配排序. 通过预先设立阈值过滤噪音,对于依存句法分析后产生的三元组,这里只保留大于阈值的三元组,对于每种统计关联度量方法都得到候选搭配结果的有序集合.

步骤 4 Google 模拟语料库. 对于每种关系的三元组,本文都选取一定数目的候选搭配发送到 Google 上面,再利用此搭配组合在 Google 出现的页面数模拟在语料中出现的次数,仍通过 3 种统计测度评价方法:共现频率、互信息及卡方检验得到各种关系的排序列表.

步骤 5 结果评价. 对比机器抽取的搭配结果和专家评测对所得到的结果进行评价,计算出精确率、召回率,并同时得到其对应曲线变化图.

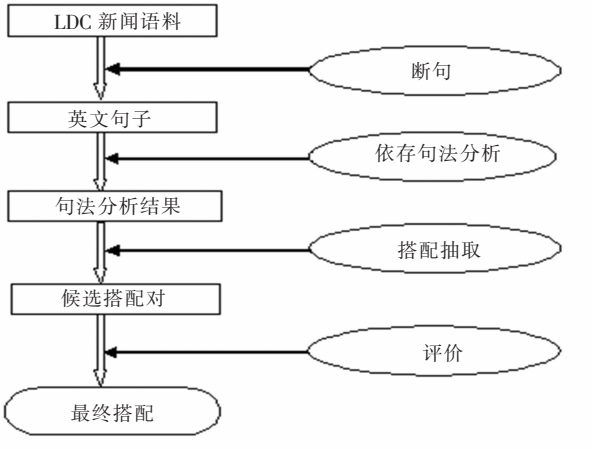

图 1　搭配抽取流程图

2.2　评价方法

搭配抽取的评价方法分为两种:1)根据人工答案或者已有的搭配知识库作为标准搭配集合,计算抽取算法的精确率和召回率;2)评价搭配算法应用到机器翻译、信息检索等系统对性能的提高. 本文随机从每种关系中取 5 000 个做实验测试样本. 首先进行人工标注,通过 3 个语言学家进行判断,并只接受那些至少被两个专家评测为正确的搭配,所得到的测试数据集如表 2 所示,对实验结果采用精确率、召回率两种评价指标[12],其定义为

$$\text{Precision} = \frac{\text{测试语料被正确识别的搭配数目}}{\text{被机器实际判定}}, \tag{4}$$

$$\text{Recall} = \frac{\text{测试语料被正确识别的搭配数目}}{\text{测试语料中实际存在的搭配数目}}. \tag{5}$$

3　实验及结果

3.1　实验数据

本文的语料来自 LDC2003 英文新闻语料,其中包括法新社(1994 ~ 1997 年、2001 ~ 2002 年)、美联社(1994 ~ 2002 年)、纽约时报(1994 ~ 2002 年)和新华社英文版(1995 ~ 2001 年),涉及时事新闻、计算机、体 育、教育、环境、娱乐、艺术等多个领域,共 4 111 240 篇文档,共抽取出 1 亿个英语句子和 370 万英语搭配,本文只考虑 3 种最重要的依存关系类型:1)动宾关系 < verb, OBJ, noun > ;2)形容词修饰名词 < noun, ATTR, Adj > ; 3)副词修饰动词 < verb, Mod, adv > ;

表 1 是测试语料的详细情况,其中,#Token 为在语料库中出现的总的次数,#Type 为去除重复出现的搭配个数,为了剔除一些噪音,本实验把阈值设为 70,即在语料中出现数目 > 70 的那些搭配才抽取出来,例如,类似 < company, ATTR, dummy > 搭配可以剔除掉,实验表明这样可以大大消除噪音.

表 1　实验数据集

类别	#Type	#Token
verb,OBJ,noun	95,813	44,629,987
noun, ATTR,Adj	88,132	39,807,202
verb,Mod,adv	17,939	10,136,019

表 2　测试数据集

类别	#共计	#正例
verb,OBJ,noun	5,000	940
noun, ATTR,Adj	5,000	1,250
verb,Mod,adv	5,000	425

3.2　实验结果

为了更好的比较各种统计方法,这里计算出精确率、召回率随着前 N 个候选搭配(N -best)变化图,如图 2 ~ 图 4 所示. 其中, WebFreq 为基于谷歌页面数目的共现频率方法,CorpusFreq 为基于语料库的共现频率方法,其他类似. X 轴为要处理的候选搭配组合的比例, Y 轴为计算出来的精确率和召回率,在精确率变化图中与 X 轴平行的线为正确搭配所占比例,即表 2 中每种关系中正例所占比例,它表示对于随机选择一定数目的候选搭配集合所计算出的精确率值.

从精确率和召回率变化中(图2~图4)可以看出:

1)在精确率图中,在前一半候选搭配结果中,各种抽取方法差异还是比较明显,后一半比较接近,同时它们精确率值超过基准线(Baseline). 同样对于召回率变化图,有的抽取方法在前一半(约 40% ~ 50%)把相当比例的搭配(70% ~ 80%)抽取出来,在后 20% 结果比较接近.

2)由图 2 中可以看出,在动宾关系(< verb,

OBJ, noun >) 中,基于语料库的共现词频方法(CorpurFreq)好于对应的基于 Web 的方法(WebFreq),其它每一种基于 Web 的方法都超过基于语料库的方法,即基于 Web 的卡方检验超过基于语料库的卡方检验方法,基于 Web 的 MI 好过基于语料库的 MI 方法.

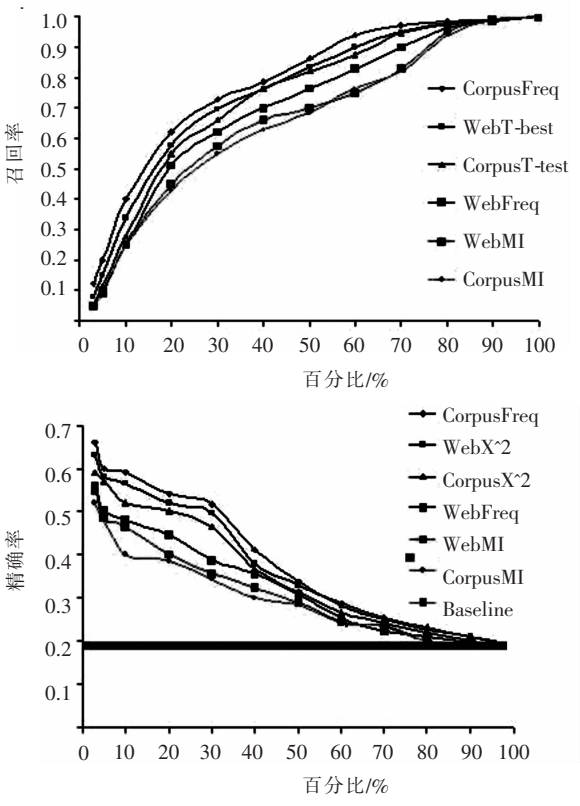

图 2 动宾(verb,OBJ,noun)关系召回率、精确率对比图

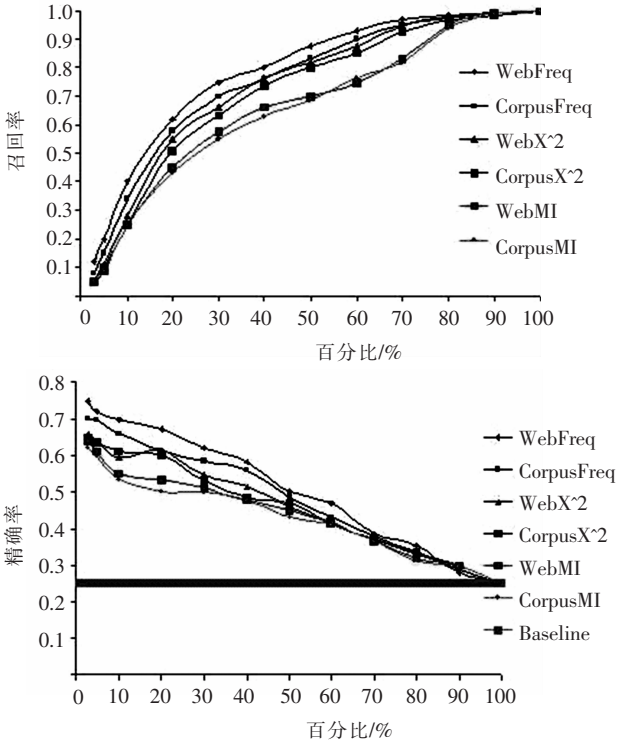

图 3 形容词修饰名词关系(noun,ATTR, adj) 召回率、精确率对比图

3)针对每种关系各种关联度量方法的效果不一样,例如在形容词修饰名词关系 < noun, ATTR, adj > 和动宾关系(< verb, OBJ, noun >)中,基于共现频率的方法好于基于卡方检验方法,基于互信息的没有前两种方法好;而在副词修饰动词 < verb, Mod, adv > 关系中,基于卡方最好,基于互信息的其次,共现频率的没有前两种好.

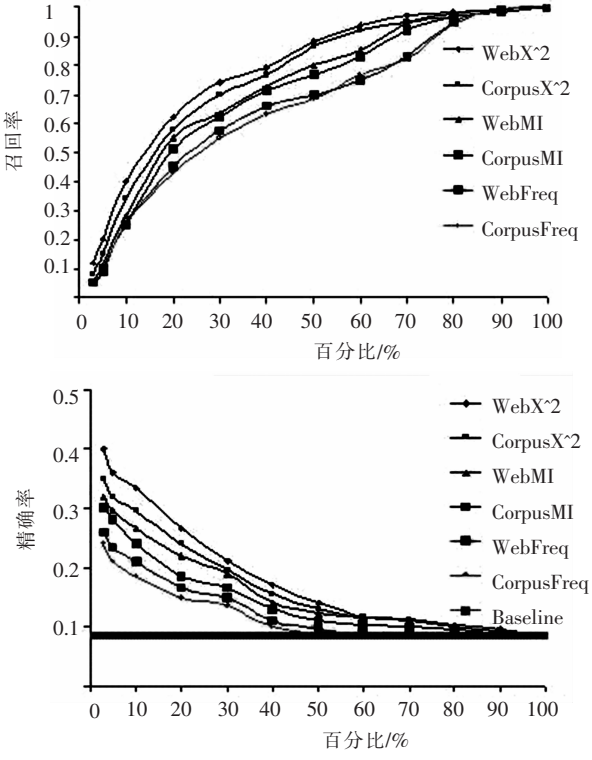

图 4 副词修饰动词(verb, Mod, adv)召回率、精确率对比图

3.3 讨 论

尽管 Web 的方法结果超过基于语料库的方法,但有些工作要在下列工作中进行拓展[13-14]:

1)尝试利用搭配字典进行评价.

2)提高句法分析的准确率.

3)不同的统计关联方法适合不同的搭配类型,这就要扩大搭配抽取的范围及测试语料库的规模.

4 结 论

1)提出更精确地针对每一类关系搭配抽取算法,例如在主谓(verb, OBJ, noun)关系和副词修饰动词(verb, Mod, adv)中卡方检验效果最好,而形容词修饰名词(noun, ATTR, adj)中基于共现词频的方法最好.

2)提出更精确的评价方法,例如利用通用搭配词典或互联网数据进行更精确地评价实验结果.

3)中英文双语搭配对自然语言处理的很多任务是非常有用的资源,例如跨语言信息检索,机器翻译系统等,提出将单语抽取的方法引入到双语搭配中.

参考文献:

[1] 全昌勤, 刘辉, 何婷婷. 基于统计模型的词语搭配自动获取方法的分析与比较[J]. 计算机应用研究, 2005, 22(9): 55 - 57.

[2] WERMTER J, HAHN U. Collocation extraction based on modifiability statistics[C]//Proceedings of the 20th International Conference on Computational Linguistics. Morristown: Association for Computational Linguistics, 2004: 980 - 986.

[3] Manning C D, SCHUTZE H. Foundations of Statistical Natural Language Processing[M]. England: MIT Press Cambridge, 1999:1 - 189.

[4] JIN Peng, SUN Xu, WU Yunfang, *et al*. Word clustering for collocation-based word sense disambiguation [C]//Proceedings of 8th International Conference on Intelligent Text Processing and Computational Linguistics. Heidelberg: Springer-Verlag, 2007: 267 - 274.

[5] XU Ruifeng, LU Qin, WONG Kam-Fai, *et al*. Annotating Chinese collocations with multi information [C]//Proceedings of Linguistic Annotation Workshop (Association for Computational Linguistics). [s. n.]: [s. l.], 2007: 61 - 68.

[6] SERETAN V, WEHRLI E. Collocation translation based on sentence alignment and parsing[C]//Proceedings of the International Conference on Automatic Natural Language Processing. [s. n.]:[s. l.], 2007: 5 - 8.

[7] 姚建民, 屈蕴茜, 朱巧明, 等. 大规模语料库中自动搭配获取的统计方法研究[J]. 计算机工程与设计, 2007, 28 (9): 2154 - 2155.

[8] 王大亮, 涂序彦,佟子健,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.

[9] SERETAN V, WEHRLI E. Accurate collocation extraction using a multilingual parser[C]//Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Assiciation for Computational Linguist. Morristown: Assiciation for Computational Linguist, 2006: 953 - 960.

[10] EVERT S, KRENN B. Methods for the qualitative evaluation of lexical association measures[C]//Proceedings of the 39th Annual Meeting on Assiciation for Computational Linguist. Morristown: Assiciation for Computational Linguist, 2001: 188 - 195.

[11] 郭锋, 李绍滋, 周昌乐. 基于词汇吸引与排斥模型的共现词提取[J]. 中文信息学报, 2004, 18(6): 16 - 23.

[12] 王素格, 杨军玲, 张武. 自动获取汉语词语搭配[J]. 中文信息学报, 2006,20(6): 31 - 37.

[13] WERMTER J, HAHN U. You can't beat frequency (unless you use linguistic knowledge)—A qualitative evaluation of association measures for collocation and term extraction[C]//Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual meeting of the Assiciation for Computational Linguist. Morristown: Assiciation for Computational Linguist, 2006: 785 - 792.

[14] WU Hua, ZHOU Ming. Synonymous collocation extraction using translation information[C]//Proceedings of the 41th Annual Meeting on Assiciation for Computational Linguist. Morristown: Assiciation for Computational Linguist, 2003:120 - 127.

(编辑　张　红)