面向依存文法分析的搭配抽取方法研究

车万翔　刘挺　秦兵　李生
哈尔滨工业大学信息检索组，黑龙江 150001
E-mail:{car, tliu, qinb, sli}@ir.hit.edu.cn

摘要：本文通过对经分词和词性标注的大规模语料库（1.8GB）的统计，计算出语料库中出现的词对个数、距离及方差，并应用 t 检验的改进方法，得到了词对之间的“搭配强度系数”值 R，以此来衡量它们之间这种搭配关系的强弱。这一系数直接面向依存文法分析，以此得到一个句子中各个词的搭配关系强弱序列表，以后将要从此表中得到依存文法树。目前我们可以在智能搜索引擎等多种场合找到此种方法的应用。

关键词：搭配　搭配强度系数　t 检验　依存文法　智能搜索引擎

A Method to Fetch Collocations Orienting Dependency Grammar

Che Wanxiang　Liu Ting　Qin Bing　Li Sheng
Information Retrieval Group, Harbin Institute of Technology 15001
E-mail:{car, tliu, qinb, sli}@ir.hit.edu.cn

ABSTARCT: In this paper, we statistic a very large corpus (1.8GB) and work out the word-pairs' number, distance's mean and variance. And then we use a modified t test method to fetch a "Collocations Coefficient" R in order to weigh how strong of their relationship. This coefficient orients to the analysis of dependency grammar straight. In this way, we have a word-pairs list in a sentence sorted by R's order. Later, we can get a dependency grammar tree from this list. Now we can find several applications using this method such as intelligent searching engine, etc.

Keywords: collocations　Collocations Coefficient　t test　dependency grammar　intelligent searching engine

1 引言

语法分析，一直是自然语言处理的一个重点和难点，目前应用的方法主要有短语结构文法和依存文法分析等。其中的依存文法分析主要是通过分析句子内词语之间的依存关系，以

此来揭示其句法结构，由于其更能准确刻画出句子中词与词之间的联系而得到了人们的广泛的重视。传统构造依存文法树的方法主要是依靠人工劳动，根据经验来获取词语间的关系(搭配)。但是由于存在着人的知识有限，效率不高，更改繁琐等缺点，人工获取搭配的方法不能大规模应用。所以，本文试图利用统计的方法，在大规模语料库中自动获取搭配，这样便弥补了人工获取的不足，得到了较为理想的结果。

2 搭配的定义

本文中搭配的定义是指任意两个词在大规模语料库中的出现。但是，并非所有的词之间都能构成搭配，在此我们认为如果词与词之间的距离大于 7 或者在它们之间出现标点符号，就不认为是搭配。

同样，不是所有词对都是好的搭配，因为由于语料库的膨胀，各种词对都有可能是随机出现，而且随着词与词之间距离波动程度的增加，它们之间搭配的程度也必然减弱。用什么方法消除这种随机性的影响呢？最简单的就是规定一个词对出现个数 n 和它们之间距离方差 σ^2 的阈值，凡是个数高于 n 或方差小于 σ^2 的词对就认为它们是好的搭配。但显然存在一些词，虽然它们出现的频率也非常高而且方差也可能非常小，但大多数情况下，并不能认为它们与其它词能构成好的搭配。于是，我们在下面应用经典的 t 检验及其变形公式定义了 R 的值，得出了较好的评价标准。

3 抽取搭配的方法

3.1 语料库的选择

由于很小的语料库不能真实反映现实中的语言现象，所以，在此处用到的语料库非常巨大，约为 1.8GB 的经过分词和词性标注的真实文本。其中，分词程序应用的是哈尔滨工业大学计算机学院机器翻译研究室的分词系统，该系统分词正确率在 93%左右，完全能够达到我们应用需要，而且由于语料库的庞大，也可以湮灭其中的分词错误。在每个词后面都有词性标注，符号表示如：/ng 为名词，/vg 为动词等等。

例句：你们/r 应当/vz 穿/vg 得/usdf 漂漂亮亮/z 的/usde ！/wj

3.2 数据结构

为能快速查填句子中词与词之间的搭配，我们使用了二元查找树的数据结构。其中，各节点存放语料库中每个词、词性以及出现的次数。特别的，我们在每个节点后又增加了一个

next 节点，用来存放出现在与该节点词搭配词的二元查找树。在新树的节点中，我们增加了两个域，分别存放词对的平均距离和距离的方差。上面给出的例句存储的树如图 1(简化)

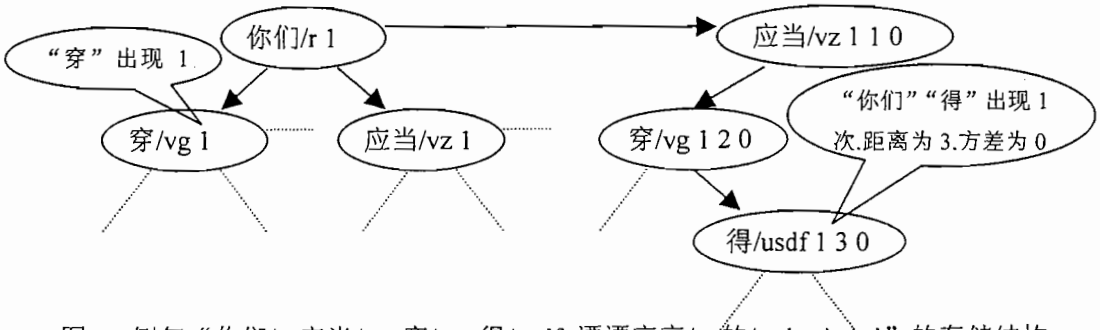

图 1. 例句“你们/r 应当/vz 穿/vg 得/usdf 漂漂亮亮/z 的/usde ！ /wj”的存储结构

当插入词时候，从根开始查找，如果插入的词小于该节点，则插入左子树；大于插入右子树；相等，则次数加 1。同理，插入词对时，先找到第一个词，再将第二个词插入它的 next 域中，遇到相同的词时，既要将次数加 1，平均距离和距离的方差也应重新计算。

平均距离的计算公式为：

$$l_n = \frac{l_{n-1} \times C_{n-1} + l_{n-1}}{C_{n-1} + 1} \tag{1}$$

方差的计算公式为：

$$\sigma_n^2 = \frac{C_{n-1} \times \sigma_{n-1}^2}{C_{n-1} + 1} + \frac{C_{n-1} \times (l_a^2 + l_{n-1}(l_{n-1} - 2 \times l_a))}{(C_{n-1} + 1)^2} \tag{2}$$

其中，l_n 为新的平均距离，l_{n-1} 为原来的平均距离，C_{n-1} 为原来词对出现的次数，l_a 为新加入的词对的距离，σ_n^2 为新的距离的方差，σ_{n-1}^2 为原方差。

3.3 抽取结果

由于语料库的巨大，虽然我们采取了较为高效的数据结构和算法，但是仍然非常耗时。在使用了一台安装有 PIII800 CPU，512M 内存，运行 Linux 操作系统的普通 PC 机上，经过两步的迭代运算，总共用时约为 15 个小时。

生成文件的格式为：

+ A 词 词性 出现次数# A 后词 1 词性 出现次数 平均距离 方差# A 后词 2 词性 出现次数 平均距离 方差# ……

+ B 词 词性 出现次数# B 后词 1 词性 出现次数 平均距离 方差# B 后词 2 词性 出现次数 平均距离 方差# ……

最后得到的文件大小为 1.5GB。

4 搭配强度系数

我们定义系数 R 来衡量两个词之间搭配的程度。本文使用的是类似于 t 检验的方法。下面，以例句中出现的“穿”和“漂漂亮亮”为例，具体说明。

其中，“穿”和“漂漂亮亮”出现的次数和方差信息如表 1：

表 1“穿”和“漂漂亮亮”统计

	“穿 /vg”	非“穿 /vg”
“漂漂亮亮 /z”	25 次 平均距离 2.08 方差 0.15 （穿...漂漂亮亮）	225 次 如：（打扮...漂漂亮亮）
非“漂漂亮亮 /z”	41038 次 如：（穿...衣服）	约为 1×10^9 (其它非“穿...漂漂亮亮”)

t 检验公式为：$t=\dfrac{\bar{x}-\mu}{\sqrt{\dfrac{s^2}{N}}}$ （3）

其中，$\bar{x}$ 是样本的期望，s^2 是样本的方差，N 是样本的大小，μ 是概率分布的期望。

首先，应用公式（3）计算“穿”和“漂漂亮亮”之间的距离 t 检验。

由例句和统计数据有：$\bar{x}$=2，μ=2.08，s^2=0.0384，N=25，有：

$$t_l=\frac{\bar{x}-\mu}{\sqrt{\dfrac{s^2}{N}}}=\frac{2-2.08}{\sqrt{\dfrac{0.15}{25}}}=-1.03 \qquad (4)$$

公式（4）中的 s^2 有可能为 0。此时，如果 $\bar{x}$ 和 μ 相等，则形成 $\dfrac{0}{0}$，根据经验，我们可设此时的 t_l 为 1。相反，如果 $\bar{x}$ 和 μ 不相等，t_l 应该很大，不妨设其为 10。

下面计算“穿”和“漂漂亮亮”出现次数的 t 检验：

P（“穿”）$=\dfrac{41038}{1\times10^9}=4.1038\times10^{-5}$

P（“漂漂亮亮”）$=\dfrac{225}{1\times10^9}=2.25\times10^{-7}$

当零假设发生时，即“穿”和“漂漂亮亮”相互独立时，有：

H_0：P（“穿”“漂漂亮亮”）=P（“穿”）P（“漂漂亮亮”）

$$=4.1038\times10^{-5}\times2.25\times10^{-7}=9.233\times10^{-12}$$

如果零假设为真，则设随机事件“穿”和“漂漂亮亮”出现为 1，其它事件为 0，是贝努里分布，其中 $p=9.233\times10^{-12}$，则期望 $\mu=p=9.233\times10^{-12}$，方差 $\sigma^2=p(1-p)\approx p$

在测试集中“穿”和“漂漂亮亮”出现了 25 次，则 $\bar{x}=\frac{25}{1\times10^9}=2.5\times10^{-8}$

$$t_c=\frac{\bar{x}-\mu}{\sqrt{\frac{s^2}{N}}}=\frac{2.5\times10^{-8}-9.233\times10^{-12}}{\sqrt{\frac{2.5\times10^{-8}}{1\times10^9}}}\approx5 \qquad (5)$$

于是，我们可以定义搭配强度系数 R，来衡量两个词之间搭配的强弱关系。公式为：

$$R=\left|\frac{t_c}{t_l}\right| \qquad (6)$$

所以，$R_{\text{“穿”“漂漂亮亮”}}=\left|\frac{5}{-1.03}\right|=4.94$

有了以上的数据和计算方法，我们就可以利用以上公式计算任何句子中各个词之间的 R 值。对有些在训练语料库中未出现的词规定其 R 值为 0。

表 2 列出了句子“她/r 喜欢/vg 穿/vg 漂亮/a 的/usde 衣服/ng 。/wj”中各个词的 R 值。

表 2.“她/r 喜欢/vg 穿/vg 漂亮/a 的/usde 衣服/ng 。/wj” R 值统计表

词 1	词 2	$R_{\text{“词1”“词2”}}$
她/r	喜欢/vg	10.21
	穿 /vg	1.62
	漂亮/a	2.63
	的/	1.77
	衣服/ng	2.04
喜欢/vg	穿 /vg	1.79
	漂亮/a	5.39
	的/	1.89
	衣服/ng	2.84
穿 /vg	漂亮/a	0.69
	的/usde z	1.63
	衣服/ng	3.48
漂亮/a	的/usde	2.43
	衣服/ng	4.35
的/used	衣服/ng	1.89

按照 R 从大到小的顺序将词对输出到表 3 如下

表 3. R 值顺序表

词对	$R_{\text{“词1”“词2”}}$	词对	$R_{\text{“词1”“词2”}}$
她/r 喜欢/vg	10.21	喜欢/vg 的/used	1.89

喜欢/vg 漂亮/a	5.39	的/usde 衣服/ng	1.89
漂亮/a 衣服/ng	4.35	喜欢/vg 穿/vg	1.78
穿/vg 衣服/ng	3.48	她/r 的/used	1.76
喜欢/vg 衣服/ng	2.84	穿/vg 的/used	1.63
她/r 漂亮/a	2.63	她/r 穿/vg	1.62
漂亮/a 的/used	2.43	穿/vg 漂亮/a	0.69
她/r 衣服/ng	2.03		

5 将搭配应用于依存文法分析

七十年代，Robinson 提出了依存文法中关于依存关系的四条公理[2]而汉语由于其特殊性，又存在第五条公理[3]。依存文法分析的关键问题就是要得到句子中任意两个词之间的依存关系的强弱，而本文定义的搭配强度系数 R 就是这种依存关系的一种反映，所以可以直接将 R 值用于依存文法分析。

具体方法是仿照表 3 由大到小列出句子中各词对的搭配强度系数 R，根据 R 值的大小，在不违背以上任意一个公理的情况下，利用回朔技术等方法，得到句子中词与词之间的依存关系。

图 2 是由例句所得到的词对关系：

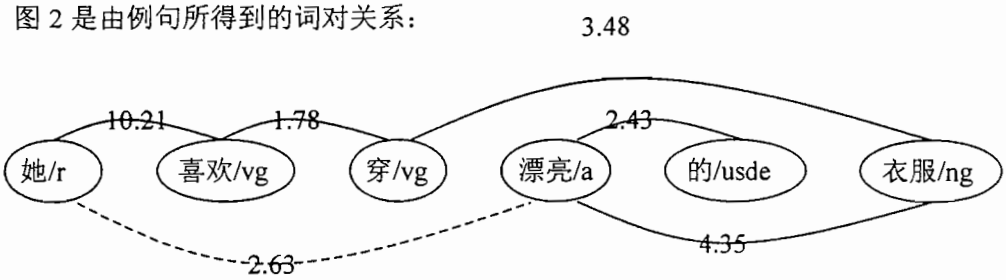

图 2. R 值关系图

图 2 中，各 R 值较高的词对被用线连上，其值在连线上标出。根据经验，我们可以发现，并非所有 R 值大的词对之间都存在依存关系，如："她"和"漂亮"，其 R 值高达 2.63，可是我们为什么不在它们认为它们之间有依存关系呢？这主要是因为在这个句子之中，"喜欢"应该是中心词，根据公理的第五条规定，中心词不能被弧跨越，所以我们认为它们之间不存在依存关系。

表 4 是任意给出的两个句子的 R 值列表（取 R 值较高的 5 个词对）：

表 4. 任意句子的 R 值列表

他们/r 经常/d 开展/vg 灭/vg 鼠/ng 运动/ng。/wj		他/r 没有/d 遇到/vg 危险/ng。/wj	
词对	$R_{"词1""词2"}$	词对	$R_{"词1""词2"}$
开展/vg 运动/ng	4.20	没有/d 遇到/vg	11.68

他们/r 开展/vg	1.70	遇到/vg 危险/ng	3.39
经常/d 开展/vg	1.69	他/r 没有/d	1.74
他们/r 经常/d	1.37	他/r 遇到/vg	1.17
灭/vg 鼠/ng	1.22	他/r 危险/ng	0.66

从表 4 中我们也可以看出，利用搭配强度系数，我们基本上可以将句子分清层次和部分。我们又从《南方周末》电子版中随机抽取了 100 个句子，结果是大部分句子基本符合汉语的语法规则，但是同样有一些句子面临着如图 2 虚线所示的问题。可见，本文提出的方法在依存文法分析的初级阶段有着很好的作用，也是深入的分析过程的基础。关于深入分析所涉及到的具体的细节工作，如中心词的选择等，有待于以后进一步应用词性等信息来完善。

6 搭配的应用

目前的中文搜索引擎大都是基于词的匹配，其中有一些已经加入了与、或、非等连接符号，还有一些所谓的整句提问的搜索，也只是去掉一些非关键词，然后再用剩余的词进行检索，这还远达不到智能的地步，因为在所提问的句子存在大量的歧义不能有效的消除。但是，如果我们对所提问的句子中的各词对进行搭配强度系数计算，然后在词对之后加上 R 值，用 R 较高的词对到数据库中进行检索，必然会消除单个词检索带来的歧义，因为一个词此时不仅有自身的含义，而且含义还受到与其搭配的词的限制，再加上我们对每个词还给予了词性的信息，句子中词的歧义被大大的削弱了。

7 总结

以上是我们为应用依存文法分析构造依存法树搭配抽取方法方面所作的初步研究和探索，基本上达到了预想的目标。实验证明我们提出的搭配抽取方法和评价方法是可行的，合理的，但是距离能够真正完美的进行依存文法分析，还有很长的路要走，这也是我们下一步的目标。

参考文献

[1] Christopher D.Manning, Hinrich Schutze:
"Foundations of Statistical Natural Language Processing", MIT Press, 1999

[2] 刘伟权等："建立现代汉语依存关系的层次体系"，《中文信息学报》, 1996 年 10(2)

[3] 黄昌宁等："语料库、知识获取和句法分析"《中文信息学报》,1992 年 6(3)

[4] Dekang Lin: "Extracting Collocations from Text Corpora", Department of Computer Science University of Manitoba, Canada, 1998