基于语料统计的量词对名词语义选择倾向的研究[1]

王萌 贾玉祥 俞士汶

北京大学计算语言学教育部重点实验室，北京，100871

E-mail: wm@pku.edu.cn yxjia@pku.edu.cn yusw@pku.edu.cn

摘要：量词系统是现代汉语语法的重要特点。本文从计算的角度，采用基于信息论和知识的方法，从大规模语料库中自动获取量词对名词中心语的语义选择倾向，对现代汉语中量名搭配进行定量分析，并考察了 24 个常用量词所搭配名词的语义分布情况。实验表明，自动获取的语义选择倾向符合语言学家对量词的定性分析和认知，并可为量词的分类提供定量参考和依据。

关键词：量词 量名搭配 语义选择倾向 数量名短语 自动获取

Research on the Selectional Preference of Classifier and Noun Based on Corpus

Wang Meng, Jia Yuxiang, Yu Shiwen

Key Laboratory of Computational Linguistics (Peking University), Ministry of Education, 100871

E-mail: wm@pku.edu.cn yxjia@pku.edu.cn yusw@pku.edu.cn

Abstract: Classifier system is an important characteristic in Chinese grammar. This paper adopts an information and knowledge-based method to automatically acquire the classifier-noun selectional preference from a large-scale corpus, and analyzes the classifier-noun collocations quantitatively and the distributions of noun classes of 24 frequent classifiers. Experiment results show that the automatically acquired selectional preference agree with the qualitative analyses and cognitions of linguists, and provide quantitative reference for the classification of classifier.

Keywords: Classifier, Classifier-noun collocation, Selectional preference, Numeral-Classifier-Noun phrase, Automatic acquisition

1 引言

作为现代汉语的一个重要语言特点，量词系统受到了语言学界的普遍关注。上世纪三四十年代，汉语语言学界对量词进行了定名和划类。此后，学者对量词展开了深入而丰富的研究，主要经历了从静态的分类研究到把量词与相关词类、与句法结构联系起来的动态研究过程[1]，取得了一系列成果[2][3]。但是，纵观以上研究，多限于对量词做定性的描述或区分，而基于语料库的定量分析并不多见。本文试图从计算的角度，从语料库中自动获取量词对名词的选择倾向，对现代汉语中量名搭配进行定量分析，并考察了 24 个常用量词所搭配名词的语义分布情况。该结果不仅为研究者从全局上对量词进行定量考察提供数据，同时也为定性分析的研究成果提供了依据。

2 量词对名词的选择倾向

2.1 量名搭配

[1] 基金支持：国家 973 课题（文本内容理解的数据基础，编号：2004CB318102）

汉语中的量词十分丰富，《现代汉语语法信息词典》[4]量词库中共收录量词 525 个。量词主要是表示计量单位，一般而言，量词在句法组成上是必不可少的部分，它与前面的数词及其所修饰的名词构成数量名短语，例如，“两部电影”、“五条鱼”、“一张纸”等。考察量名搭配的特点，一方面，量词所修饰的名词通常是有限的（“个”是使用最为广泛的量词，几乎所有的个体名词都能论“个”），只能与某些类名词搭配，而这一类中的名词通常会有一些相近的特点。例如，量词“条”，可以说“一条鱼”、“一条蛇”及“一条河”，“条”隐含了其修饰的名词是具备“长”、“柔软”以及“像绳子一样”等特征的物体；量词“张”则隐含了物体有延展的平面，如“一张桌子”、“一张床”等。此外，还有一些约定俗成的搭配，例如“一把扇子”、“一把锁”等。另一方面，名词也可有不同的量词和它搭配，不同量词反映了名词不同方面的特征，例如，名词“布”，既可以用“块”修饰，表现了布的局部特征，又可用“匹”修饰，表现的是较大整体的卷状形态。从量词与名词的搭配关系中可以看出，量词通过其隐含的语义信息，对名词存在选择限制。

2.2 选择倾向

选择倾向（selectional preference）一般指谓词（predicate）对其论元（argument）的在语义上的限制。比如“吃”的宾语倾向于<食物>（用尖括号表示一个语义类），“鸣叫”的主语倾向是<鸟>。选择倾向是人类知识的重要组成部分，在统计自然语言处理领域也受到广泛关注，被广泛用于词义消歧、句法消歧、未知词语意思推断等诸多任务。

选择倾向方面的研究，主要集中在英文的动词对宾语的选择限制。Resink（1993）[5]提出了获取选择倾向的基本框架，并实现了一种选择倾向模型，该模型基于 WordNet 语义分类体系，通过信息论的计算模型，从语料中获取动词对宾语的选择倾向。实际上，这个模型可以应用到任何词类的语义约束中，例如，动词和主语、动词和介词短语以及形容词和名词。本文将该模型应用到量词和名词的语义约束中，采用《中文概念词典》[2]（Chinese Concept Dictionary，CCD）[6]作为语义分类体系，通过信息论相对熵的计算模型，从语料中自动获取量词对名词的选择倾向，并对结果进行了定量分析。

3 获取方法

选择倾向形式化描述为一个映射 *selects*: $(m, r, c) \rightarrow a$。r 是语法关系（这里指量词对名词），m 是量词，c 是语义类，a 是一个实数，表征 m 选择 c 的可能性。选择倾向的获取就是从训练语料中学习这一映射。

3.1 CCD 名词语义分类体系

CCD 的总体结构沿用 WordNet 框架，以同义词集合（Synset）表示概念（Concept），在概念之间定义关系。概念按词类分为动词、名词、形容词、副词四种类型，分别存贮在四个文件。名词概念（语义类）有 66025 个，含名词 104167 个，概念关系包括上位（Hypernym）、下位（Hyponym）、整体（Holonym）、部分（Meronym）、反义（Antonym）等语义关系。如果概念 A 是概念 B 的上位，则 B 是 A 的下位，下位关系是传递和反对称的。一个概念既可以有多个上位概念，也可以有多个下位概念。概念通过上下位关系形成层次结构。

[2] CCD 是基于 WordNet 构建的，它根据汉语的特点，继承并优化了 WordNet 的语义分类体系，为中文选择倾向的研究提供了基础，本文采用的是 2006 年的版本。

图 1 是 CCD 名词概念上下位关系的示意图。其中，箭头由下位指向上位，实线表示直接上下位关系，虚线表示间接上下位关系，中间省略了一些层次；虚线三角形表示下层语义结构从略。可见一个上位概念包含多个下位概念，而一个下位概念也可以有多个直接上位，如<水>既属于<物体>，又属于<液体>。这样，CCD 名词概念构成一个有向无环图，而不是一个严格的树结构。

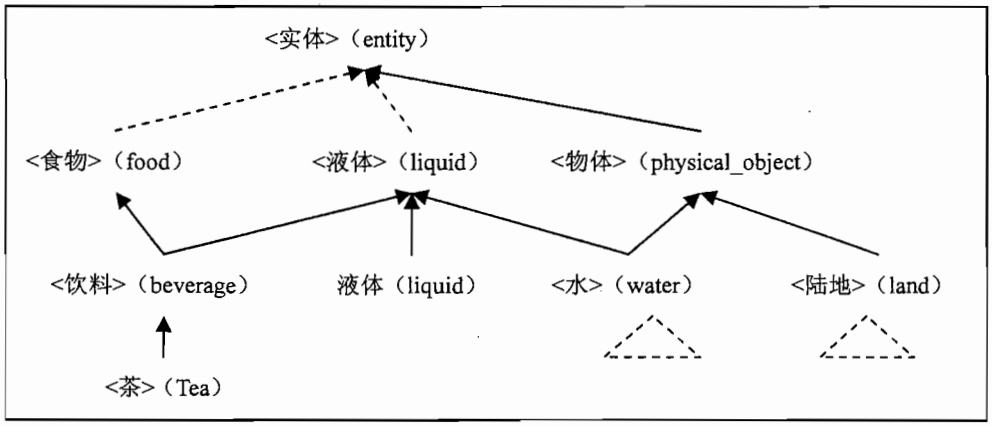

图 1 CCD 名词语义分类体系示意图

3.2 信息论模型

大部分量词都倾向于修饰特定的名词，例如，量词“辆”倾向于修饰交通工具，“幅”倾向于修饰图画，“名”倾向于修饰人。设 Pr（c）表示名词语义类的总体概率分布，Pr（c|m）表示名词类作为量词 m 的中心语时的概率分布。信息论中的相对熵（又称 KL 距离，Kullback-Leibler divergence）可以度量两个概率分布之间差异。因此量词对其所修饰中心名词语义类的选择倾向强度（selectional preference strength）被定义为名词类的先验分布与待考察量词所修饰名词类的分布之间的相对熵，如公式 1 所示：

$$s_r(\mathrm{m}) = D(\Pr(c \mid \mathrm{m}) \| \Pr(c)) = \sum_c \Pr(c \mid \mathrm{m}) \log \frac{\Pr(c \mid \mathrm{m})}{\Pr(c)} \quad \text{（公式 1）}$$

选择倾向强度度量了量词约束它所修饰的名词的强度，选择优先强度越大，量词就越倾向于选择某些语义类。基于选择倾向强度，可以定义量词与名词语义类的选择关联度（selectional association），如公式 2 所示：

$$A_r(\mathrm{m}, c) = \frac{1}{s_r(\mathrm{m})} \Pr(c \mid \mathrm{m}) \log \frac{\Pr(c \mid \mathrm{m})}{\Pr(c)} \quad \text{（公式 2）}$$

公式（2）是一个比值，体现了某语义类对量词选择优先强度的相对贡献，选择关联度越大，量词对该语义类的选择倾向性越强。

3.3 参数估计

本文使用最大似然估计（Maximum Likelihood Estimation，MLE）对公式 1 和公式 2 中的概率 $\Pr(c)$ 和 $\Pr(c \mid \mathrm{m})$ 作出估计，分别如公式 3 和公式 4 所示。

$$\hat{\Pr}(c)=\frac{freq(c)}{\sum_{c'} freq(c')} \quad \text{（公式 3）} \qquad \hat{\Pr}(c\mid \mathrm{m})=\frac{freq(\mathrm{m},c)}{freq(\mathrm{m})} \quad \text{（公式 4）}$$

文本中出现的是词 w，不是语义类 c，需要借助语义分类体系，将词语映射到语义类，通过计算词频 $freq(w)$或共现词频 $freq(m,w)$来估计语义类出现的频次 $freq(c)$或共现频次 $freq(\mathrm{m},c)$，如公式 5 及公式 6 所示。其中 $classes(w)$是包含词 w 的语义类集合，由 w 所在的各个概念及其所有上位概念组成；$words(c)$是语义类 c 所包含的词语集合，如果 $c\in classes(w)$，则 $w\in words(c)$。通过 CCD 获得每个词 w 所在的语义类集合 $classes(w)$后，即可得到 $words(c)$。

一个词可能有多个义项，每个义项对应于 CCD 中的一个概念。这里对词的义项不做区分，假设词 w 的出现对每个义项及其所有上位概念均起作用，而且对这些语义类的贡献均等。因此，词频 $freq(w)$要除以语义类的个数$|classes(w)|$以得到语义类的频次 $freq(c)$，共现频次的计算同理。

$$freq(c)=\sum_{w\in words(c)}\frac{1}{|classes(w)|}freq(w) \quad \text{（公式 5）}$$

$$freq(\mathrm{m},c)=\sum_{w\in words(c)}\frac{1}{|classes(w)|}freq(\mathrm{m},w) \quad \text{（公式 6）}$$

借助 CCD 语义分类体系获得 $classes(w)$的方法如图 2 所示。例如，“树”一词有两个义项，两个义项及其所有上位概念构成的语义类集合 $classes(w)$，可以得到$|classes(w)|$=12。其中“=>”左边是下位概念，右边是上位概念。

Sense1
树--（一个高的多年生长的木本植物，有主干和枝杈……） =>木本植物--（一种有坚硬的木质化的茎或木质部分的植物） =>维管植物--（……） =>植物，植物生命体--（……） =>生命形式，生物体，生物--（任何活的实体） =>实体--（任何存在的东西（有生命或无生命））
Sense 2
树，树图--（从一个根出发的所有枝杈构成的图形；“家谱树”） =>平面图，二维图形--（2-维图形） =>图形--（点、线、面所组合而成的可视的形） =>形状，形式--（事物的空间排列；“几何是关于形状的数学科学”） =>属性--（属于一个个体的抽象或特征） =>抽象--（从具体的事例中抽取公共特征从而形成的一般概念）

图 2 从 CCD 中获取 $classes(w)$

4 实验结果

4.1 量名搭配的获取

我们从 1998 年上半年的《人民日报》基本标注语料（分词和词性标注）中自动获取量名搭

配对，为此实现了一个“复杂数量名短语”自动识别算法，有着较高的准确率和召回率，该方法详见文献[7]。该算法的输入是经过分词和词性标注的语料，输出结果则标注了数量名短语。例如，“这/r 是/v 美/j 中/j 两/m 军/n 之间/f 签署/v 的/u [第一/m 个/q 有关/vn 军事/n 安全/an 磋商/vn 机制/n 的/u 协定/n]”，方括号中的内容即为数量名短语的识别结果，从中可以得到一个量名搭配对<个，协定>。按照上述方法，我们从98年上半年《人民日报》语料中共获取量名搭配对152,430对，所有参数估计均在该二元对上进行。

4.2 选择倾向

选取24个常用量词，考察量词所搭配名词的语义类情况。从语料库中自动获取量词对名词的语义选择倾向，得到量词对各层次所有语义类的选择程度 $A_r(m, c)$，进而可以分析名词语义类的分布情况，得到量词最倾向于修饰哪个语义类的名词。

表1给出了24个量词的选择优先强度 $S_r(m)$，并按照选择优先强度的降序排列。可以看出，量词“个”作为汉语中最广泛使用的量词，对其所修饰名词的约束强度最弱，几乎没有选择倾向性；集合量词“批”和“种”的选择倾向性次之。而量词“匹”作为专职量词[3]，对名词的选择倾向性最明显。这样的结果与语言学家对量词的分类和认知是非常一致的。

序号	量词	$S_r(m)$	序号	量词	$S_r(m)$
1	匹	3.6642	13	张	1.3181
2	辆	2.9405	14	家	1.2915
3	句	2.7387	15	座	1.2839
4	所	2.7352	16	件	1.2495
5	只	2.1176	17	位	1.2018
6	幅	2.0049	18	名	1.1013
7	元	1.8683	19	项	1.0847
8	起	1.6587	20	次	0.9800
9	段	1.5633	21	种	0.5747
10	支	1.4652	22	本	0.5444
11	条	1.4251	23	批	0.4528
12	部	1.3685	24	个	0.1352

表1 量词的选择倾向强度 $S_r(m)$

表2分别给出量词“匹”、“辆”和“支”的名词语义类，并按选择关联度从大到小排序，每个量词只给出前5个。

量词	语义类编号	语义类词语	选择关联度
	01875130	马 白马 马科动物	0.0681
	01874317	奇蹄动物 奇蹄类动物	0.0681
匹	01875414	马 白马 马属	0.0644
	01871837	有蹄类 有蹄动物	0.0568
	01878442	坐骑 鞍马 驯马 骑马 乘用马	0.0363

	03569523	车 交通工具 运输工具	0.1050
	02495376	传送器 运输机 交通工具 运输工具 运输系统	0.1040
辆	03018224	汽车 机动车辆 发动机车辆	0.0916
	02383458	汽车 机动车 机动车辆	0.0672
	02859872	仪器 工具	0.0671
	06209853	人马 兵丁 兵卒 兵员 兵家 兵源 军人 军队 军马	0.0625
	06232518	队伍 军事编队 军队编制 军队队形	0.0547
支	06233236	行列 队伍	0.0547
	06235973	列 排 行 队	0.0544
	06212281	会友 队伍 会员身份 全体会员	0.0536

表2 部分量词所搭配名词的语义类及其关联强度

从表2中可见，“匹”和“辆”作为选择倾向性最高的两个量词，它们所修饰名词的语义类非常集中，例如，“辆”排名最前的都是<车，交通工具>类名词。表中获得的名词语义类处于不同的抽象层次，排名越靠前的名词语义类是量词最常搭配的语义类。例如，表2中显示量词“支”与<军队>类名词搭配的选择关联度最高，“支”与<歌曲>类名词搭配的选择关联度低于<军队>类（为0.00153，表2中未列出），这说明虽然“支”可以与不同语义类的名词搭配，但是在实际使用过程中和<军队>类名词搭配的情况更多见。

5 结语

本文首次采用基于信息论和知识的计算模型，从大规模语料库中自动获取汉语量词对名词的语义选择倾向，定量地考察了量词所搭配名词的语义分布情况。实验表明，自动获取的语义选择倾向符合语言学家对量词的定性分析和认知，不但从全局的、定量的角度为量词的研究提供了佐证，二者互为补充，而且，获取结果也可作为语言学者对量词分类的定量参考和依据。

参 考 文 献

[1] 王冬梅. 现代汉语量词研究综述.. 扬州大学学报 (人文社会科学版). 1997年第6期.

[2] 刘子平. 汉语量词词典. 内蒙古教育出版社. 1996

[3] 何杰. 现代汉语量词研究. 民族出版社. 2001

[4] 俞士汶、朱学锋、王惠等. 现代汉语语法信息词典详解（第二版）. 北京：清华大学出版社，2003年2月

[5] P. Resnik. Selection and Information: A Classed-Based Approach to Lexical Relationships. Ph.D. thesis. University of Pennsylvania, Philadelphia, PA. 1993.

[6] 于江生, 俞士汶. 中文概念词典的结构. 中文信息学报, 16(4). 2002. 12-20

[7] Meng Wang，Shiwen Yu， Huiming Duan， Exploiting Salient Word Dependency for Chinese NP Identification: A Study on Classifier Noun Phrase，IEEE NLP-KE，Beijing，2008.

[8] 贾玉祥、俞士汶. 语义选择限制的自动获取及其在隐喻处理中的应用. 第四届学生计算语言学研讨会论文集. 山西 太原. 2008年7月

[9] 朱德熙. 语法讲义. 北京：商务印书馆，1982年9月