语义选择限制的自动获取及其在隐喻处理中的应用[1]

贾玉祥 俞士汶

北京大学计算语言学研究所，北京，100871

yxjia@pku.edu.cn yusw@pku.edu.cn

摘要：语义选择限制是人类知识的重要组成部分，被广泛用于多种自然语言处理任务。本文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动词对宾语的语义选择限制，并将其用于动词隐喻的处理。实验表明，自动获取的语义选择限制与人工判断有较高的一致性。本文方法可以推广到动词对主语、形容词对名词中心语等语义选择限制的获取。

关键词：语义选择限制 自动获取 隐喻处理

Automatic Acquisition of Selectional Preference and Its Application to Metaphor Processing

Jia Yuxiang, Yu Shiwen

The Institute of Computational Linguistics, Peking University, Beijing, 100871

yxjia@pku.edu.cn yusw@pku.edu.cn

Abstract: Selectional preference is an important part of human knowledge, and has been widely used in many natural language processing tasks. This paper adopts an information and knowledge-based method to automatically acquire the verb-object selectional preference from a large-scale corpus, and applies the selectional preference to verb metaphor processing. Experiment results show that the automatically acquired selectional preference agree well with human judgment. This method can be generalized to other grammatical relations, such as verb-subject and adjective-noun, etc.

Keywords: Selectional preference, Automatic acquisition, Metaphor processing

1 引言

语义选择限制（selectional preference）是指谓词（predicate）对其论元（argument）的在语义上的限制。比如“吃”的宾语一般为<食物>（尖括号括起来表示一个语义类，而不是一个词），“思考”的主语要求是<人>。语义选择限制是人类知识的重要组成部分，很早就被提出并应用于语义理论[1]，在自然语言处理领域也受到广泛关注，被用于词义消歧[2]、句法消歧[3]、语义角色标注[4]、语义词典构建[5]、未知词语意思推断等诸多任务。

从语料中自动获取语义选择限制的研究始于文献[6]，它提出了获取语义选择限制的基本框架，即基于某一语义分类体系，通过某种计算模型，从语料中获取谓词对论元的选择限制。[6]使用信息论的方法，以 WordNet 作为其语义分类体系。以 WordNet 为基础，[7]提出基于最小描述长度（Minimum Description Length，MDL）的方法，[8]提出基于假设检验的方法，[9]使用隐马尔科夫模型，而[10]使用贝叶斯网络。

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

语义选择限制方面的研究，主要以英文、动词对宾语的选择限制为主。多语 WordNet 的发展，也为其他语言的研究提供了契机，如德语基于 GermaNet 的研究。中文方面，中文概念词典[11]（Chinese Concept Dictionary，CCD）是一本类 WordNet 词典，它根据汉语的特点，继承并优化了 WordNet 的语义分类体系，为中文语义选择限制的研究提供了基础。[12]基于语料库，探讨了个别动词对宾语的语义选择限制情况，而语义选择限制自动获取方面的研究还未见报道。

隐喻本质上是一种认知现象，在人们的自然语言中普遍存在，是自然语言理解不可回避的问题[13]。可以认为隐喻产生的原因是谓词与其逻辑主语或宾语之间存在冲突，即主语或宾语违反了谓词的语义选择限制。如“汽车喝汽油”，“喝”的主语要求是有生命的语义类，而“汽车”是无生命的，因此这句话是隐喻用法。语义选择限制是隐喻处理的重要知识源[14]，但汉语隐喻处理的研究才刚刚起步，还未对这一知识源加以利用。

本文以 CCD 2006 版的语义分类体系为基础，使用经典的基于信息论的方法[6]，从大规模中文语料中获取动词对宾语的语义选择限制，并将其用于隐喻的自动处理。该方法可以推广到动词与主语、形容词与名词中心语等语法关系。

2 获取方法

语义选择限制形式化描述为一个映射 *selects*: $(p, r, c) \rightarrow a$。r 是语法关系（如动词与宾语），p 是谓词，c 是语义类，a 是一个实数，表征 p 选择 c 的可能性。语义选择限制的获取就是从训练语料中学习这一映射。

2.1 CCD 名词语义分类体系

CCD 用同义词集合表示概念，概念按词类分为动词、名词、形容词、副词四种类型。名词概念（语义类）有 66025 个，含名词 104167 个，概念之间存在上位（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，可以认为 A 是 B 的超集，B 是 A 的子集。概念通过上下位关系形成层次结构。

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；尖括号表示概念，不带尖括号表示词本身。可见一个上位概念由多个下位概念组成，同时一个下位概念也可以有多个直接上位，如<饮料>既属于<食物>，又属于<液体>。统计显示，CCD 中有 872 个名词概念有两个以上的直接上位概念。这样，以每个最上层概念为根节点，构成一个有向无环图，而不是一个严格的树结构。

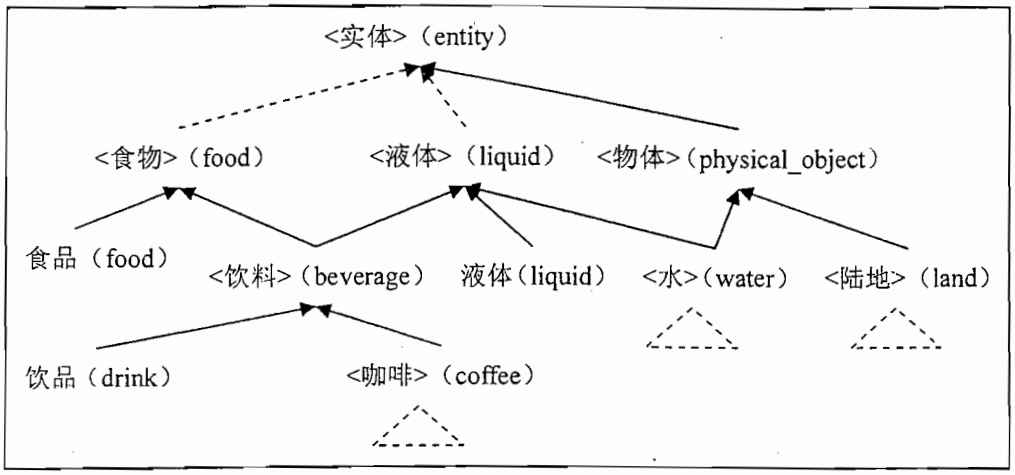

图 1 CCD 名词语义分类体系示意图

2.2 信息论模型

在一般语料中，语义类<人>做主语的先验概率 Pr(<人>)大于<鸟>做主语的先验概率 Pr(<鸟>)。但当谓词是“鸣叫”时，<人>的后验概率 Pr(<人>|鸣叫)小于先验概率，而<鸟>的后验概率 Pr(<鸟>|鸣叫)大于先验概率，且 Pr(<鸟>|鸣叫)>Pr(<人>|鸣叫)。论元语义类的后验概率分布和先验概率分布之间的差异，体现了谓词对论元语义类的选择性。谓词“鸣叫”倾向于选择<鸟>作为主语。信息论中的相对熵（又称 KL 距离，Kullback-Leibler divergence）是度量两个概率分布之间差异的指标。因此谓词 p 在语法关系 r 下对论元语义类的选择优先强度（selectional preference strength）定义如下：

$$s_r(p) = D(\Pr(c \mid p) \| \Pr(c)) = \sum_c \Pr(c \mid p) \log \frac{\Pr(c \mid p)}{\Pr(c)} \qquad \text{（公式 1）}$$

选择优先强度越大，谓词就越倾向于选择某些语义类。到底倾向于选择哪些语义类，定义如下谓词对论元语义类的选择关联度（selectional association）：

$$A_r(p,c) = \frac{1}{s_r(p)} \Pr(c \mid p) \log \frac{\Pr(c \mid p)}{\Pr(c)} \qquad \text{（公式 2）}$$

即该语义类对谓词选择优先强度的相对贡献，体现了该语义类做谓词论元的适合程度。选择关联度越大，谓词对该语义类的选择倾向性越强。选择关联度计算函数即前文提到的映射 *selects*: $(p, r, c) \rightarrow a$。

2.3 参数估计

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

$$\hat{\Pr}(c) = \frac{freq(c)}{\sum_{c'} freq(c')} \qquad \text{（公式 3）}$$

$$\hat{\Pr}(c \mid p) = \frac{freq(p,c)}{freq(p)} \qquad \text{（公式 4）}$$

文本中出现的是词 w，不是语义类 c。用词频 $freq(w)$或共现词频 $freq(p,w)$来估计语义类出现的频率 $freq(c)$或共现频率 $freq(p,c)$（见公式 5 及公式 6），需要借助语义分类体系。一个词可能有多个义项，每个义项对应于 CCD 中的一个概念（语义类）。这里对词的义项不做区分，假设词的出现对每个义项均起作用，并且对义项的所有上位概念均起作用。包含词 w 的语义类集合 $classes(w)$是由 w 所在的各个概念及其所有上位概念组成，而且 w 对这些语义类的贡献均等，即词频要除以语义类的个数$|classes(w)|$。假设每个语义类 c 所包含的词的集合为 $words(c)$，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 $words(c)$。

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

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

借助 CCD 语义分类体系获得 $classes(w)$的方法如图 2 所示。“晚餐”一词有三个义项，三个义项及其所有上位概念构成的语义类集合 $classes(w)$大小$|classes(w)|$=13。用二元组表示一个概念节点（如<05629070, 晚餐>），其中，数字（05629070）是概念在 CCD 中的唯一编码，词语（“晚餐”）是该节点同义词集合中的一个代表词语。“=>”左边是下位概念，右边是上位概念。CCD 中最大的情况，$|classes($“点”$)|$=81，$|words($<00001740, 万物>$)|$=59196。

w = 晚餐
Sense 1 <05629070, 晚餐> => <05627549, 饭> => <05625967, 食品> => <00011575, 营养物> => <00010572, 物质> => <00009457, 实体> => <00001740, 万物>
Sense 2 <05629292, 晚餐> => <05627549, 饭> => <05625967, 食品> => <00011575, 营养物> => <00010572, 物质> => <00009457, 实体> => <00001740, 万物>
Sense 3 <06130676, 晚餐> => <06126145, 社交集会> => <05978845, 集合> => <05962976, 社会团体> => <00017954, 群体>
classes(w)={<05629070, 晚餐>, <05627549, 饭>, <05625967, 食品>, <00011575, 营养物>, <00010572, 物质>, <00009457, 实体>, <00001740, 万物>, <05629292, 晚餐>, <06130676, 晚餐>, <06126145, 社交集会>, <05978845, 集合>, <05962976, 社会团体>, <00017954, 群体>}

图 2 从 CCD 中获取 $classes(w)$

3 实验与分析

3.1 语料库

从 2000 年人民日报全年语料中自动抽取<动词, 宾语中心词>二元对，共 792770 对，所有参数估计均在该二元对上进行。动词宾语中心词的抽取，是在分词、标注的基础上，采用简单启发式规则实现。例如，

边界确定：目标动词之后，下一个动词或标点之前。

歧义消解：如果有多个候选名词，则选择最后一个。

对句子“ 果断/ad 采取/v 一/m 系列/q 宏观/n 经济/n 政策/n 措施/n ，/wd”来说，选择“采取”之后，“，”之前的所有名词的最后一个，即“措施”，作为宾语中心词。

规则大大简化了句子的分析，也会带来一些误差。如句子“ 采取/v 市民/n 代表/n 座谈/v 、/wu 张贴/v 公开栏/n 等/u 形式/n ，/wd”中，“采取”的宾语中心词误选为“代表”。然而，从实验结果看，在大规模数据的情况下，这些误差对语义选择限制的影响是可以接受的。

3.2 语义选择限制

文献[12]选取 46 个高频动词，考察动词宾语语义类的情况，只给出做宾语的顶层语义类，如“发挥”的宾语语义类是<属性>，“举行”的宾语语义类是<事情>。本文对[12]中的所有动词，从语料库中自动获取对宾语的语义选择限制，得到动词对各层次所有语义类的选择程度 $A_r(p,c)$，进而得到更具体更适合的语义类，如“发挥”的宾语语义类<作用>，“举行”的宾语语义类<社交集会>等。

表 1 给出了选择优先强度 $S_r(p)$最大的 10 个动词，这些动词对宾语语义类的选择倾向性非常明显；并给出了各动词最可能选做宾语的语义类，即选择关联度 $A_r(p,c)$最大的语义类。可以看出，所选出的语义类与人的认知非常一致。

表 1 $S_r(p)$最大的 10 个动词

动词	$S_r(p)$	宾语语义类	$A_r(p,c)$
附	3.36459	<05252378, 图片>	0.107091
改正	2.79074	<03680230, 错>	0.022959
采取	2.51549	<00113217, 措施>	0.168648
会见	2.01743	<00004123, 人>	0.049084
伤害	1.99238	<05561386, 感情>	0.0756803
震撼	1.70509	<07594286, 心灵>	0.0806628
发挥	1.48463	<00606005, 作用>	0.0392447
举行	1.34798	<06126145, 社交集会>	0.0678182
建筑	1.21997	<02839091, 住房>	0.0358957
表示	1.14342	<05578169, 感谢>	0.0935749

表 2 给出动词“发挥”和“举行”的宾语语义类，并按选择关联度从大到小排序。可见，第一，宾语语义类是弹性多层次的，有高层抽象的，也有低层具体的，能很好解决数据稀疏的问题。第二，具体且准确的语义类排在前面，更好地反映了真实的认知情况，如<00606005, 作用>排在其上位<00017487, 行为>的前面，<06126145, 社交集会>排在其上位<00017954, 团体>前

面。

表 2 动词宾语语义类举例

发挥		举行	
<00606005, 作用>	0.0392447	**<06126145, 社交集会>**	0.0678182
<04546552, 价值>	0.0381189	<06126536, 聚会>	0.0510053
<00112303, 作用>	0.036244	<05978845, 集合>	0.0488007
<04044314, 效果>	0.0356859	<06128050, 宴会>	0.0482897
<04043948, 影响>	0.0356329	<06128171, 宴会>	0.0482897
<04041746, 威力>	0.0354619	<06128375, 招待会>	0.0467096
……		……	
<04546018, 意义>	0.0311583	<05319740, 听觉交流>	0.0174465
<04544110, 感想>	0.0289656	**<00017954, 团体>**	0.0156234
<00018604, 属性>	0.0138881	<06167318, 集会>	0.0147478
<00017487, 行为>	0.00215883	<05538537, 事情>	0.0141116
……		……	

4 隐喻处理

获取动词对主语或宾语的语义选择限制，对隐喻的处理是有帮助的。可以把选择关联度最大的那个（或前几个）语义类作为正常用法下的语义类，而其他语义类认为是隐喻用法的语义类。这样，通过考察动词主语或宾语所在的语义类，就可以识别出动词是否为隐喻用法。

本文考察了 10 个常用作隐喻的动词对宾语的语义选择限制情况（如表 3 所示，动词按其选择优先强度从高到低排序），并给出其最倾向于选择的宾语语义类，即选择关联度最大的语义类。简单假设该语义类为正常语义类，其他语义类为隐喻语义类，就可以识别出前 7 个动词的如下隐喻用法：透支生命、浇灌和平之花、播撒爱心、酿造悲剧、提炼经验、点燃激情、编织梦想。

相对而言，正常语义类一般为具体概念，而隐喻语义类常为抽象概念。“兜售、培植、兑现”的宾语抽象语义类（<现象>、<财源>、<诺言>，见表 3）已经超过具体语义类（<商品>、<植物>、<钱财>）成为最常用的语义类，这也说明动词隐喻义成了最常用的义项。这恰恰反映了词语义项的发展规律，即先有基本义，再有隐喻义，隐喻义用得多了，就超越基本义成为最常用义项。这时，可以通过考察语义类是具体还是抽象来识别下列隐喻用法：兜售歪理邪说、培植亲信、兑现诺言。

表 3 隐喻动词的宾语语义类

动词	$S_r(p)$	宾语语义类	$A_r(p,c)$
透支	2.48766	<09633105, 信用卡>	0.0638644
浇灌	2.26237	<06240750, 植物>	0.033384
播撒	2.24628	<05839679, 种子>	0.0359649
酿造	2.07474	<05910986, 酒类饮料>	0.0442283

提炼	2.05898	<10603250, 硫磺>	0.0589559
点燃	2.00696	<03372456, 光源>	0.0933905
编织	1.71787	<03454043, 毛衣>	0.0591498
兜售	1.64352	<00020461, 现象>	0.0420158
培植	0.993831	<09616555, 财源>	0.0820914
兑现	0.988036	<05396507, 诺言>	0.027107

5 总结与展望

本文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汉语动词对宾语的语义选择限制，并将其用于隐喻的识别。实验表明，自动获取的语义选择限制与人工判断具有较高的一致性。本文方法可以推广到其他谓词论元关系，如动词与主语、形容词与名词中心语等。

下一步工作主要包括：改进动词宾语中心词抽取方法，考虑使用依存分析工具或块分析工具。比较几种语义选择限制自动获取的计算模型，看哪一个模型更适于汉语的情况。获取动词对主语的语义选择限制，联合宾语语义选择限制的知识，共同用于动词隐喻的处理。

参 考 文 献

[1] J. Katz and J. Fodor. The structure of a semantic theory. Language, 39(2). 1963. 170-210.

[2] D. McCarthy and J. Carroll. Disambiguating nouns, verbs, and adjectives using automatically acquired selectional preferences. Computational Linguistics, 29(4). 2003. 639-654.

[3] D. Hindle and M. Rooth. Structural ambiguity and lexical relations. Computational Linguistics, 19(1). 1993. 103-120.

[4] D. Gildea and D. Jurafsky. Automatic labeling of semantic roles. Computational Linguistics, 28(3). 2002. 245-288.

[5] M. Rooth, S. Riezler, D.Prescher, G. Carroll, and F. Beil. Inducing a semantically annotated lexicon via EM-based clustering. In Proceedings of ACL 1999, Maryland. 104-111.

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

[7] H. Li, and N. Abe. Generalizing case frames using a thesaurus and the MDL principle. Computational Linguistics, 24(2). 1998. 217-244.

[8] S. Clark and D. Weir. Class-based probability estimation using a semantic hierarchy. Computational Linguistics. 28(2). 2002. 187-206.

[9] S. Abney and M. Light. Hiding a semantic hierarchy in a Markov model. In Proc. of the ACL Workshop on Unsupervised Learning in Natural Language Processing. 1999. 1-8.

[10] M. Ciaramita and M. Johnson. Explaining away ambiguity: Learning verb selectional preference with Bayesian networks. In proc. of the COLING 2000. 2000.187-193.

[11] 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.

[12] 吴云芳, 段慧明, 俞士汶. 动词对宾语的语义选择限制. 语言文字应用. 2005 年 5 月第 2 期. 121-128.

[13] 俞士汶. 自然语言理解研究与文学表现手法. 第二届文学与信息技术国际研讨会. 2005. 2-13.

[14] Z. J. Mason. CorMet: A computational, corpus-based conventional metaphor extraction system. Computational Linguistics, 30(1). 2004. 23-44.