基于实例的汉语语义超常搭配的自动发现 *)

杨 芸[1] 李剑锋[1] 周昌乐[1,2] 黄孝喜[2,3]
(厦门大学人工智能研究所 厦门 361005)[1] (浙江大学语言与认知研究中心 杭州 310028)[2]
(浙江大学计算机科学与技术学院 杭州 310027)[3]

摘 要 搭配在语言信息处理中具有重要的应用价值,通常我们主要关注符合语法规则的常规搭配。实际上,语言中还存在着大量的语法上符合规则而语义上不符合常规认知的语义超常搭配现象,而这样的现象与语言的隐喻表达和思维有着密切的联系,对自然语言理解将产生重要的影响。本文面向汉语隐喻理解来研究文本中语义超常搭配的自动发现方法,从汉语语义超常搭配判断的心理机制出发,提出了基于实例的汉语语义超常搭配识别的量化计算方法。实验以动词为中心的搭配语料为测试集,语义超常搭配识别的召回率为 80.7%,准确率为 81.5%。实验结果表明本文所给出的基于实例语义超常搭配判断的办法是切实可行的。

关键词 义超常搭配,语义搭配超常度,隐喻

Example-based Discovery of Unconventional Chinese Semantic Collocations

YANG Yun[1] LI Jian-feng[1] ZHOU Chang-le[1,2] HUANG Xiao-xi[2,3]
(Institute of Artificial Intelligence, Xiamen University, Xiamen 361005, China)[1]
(Center for the Study of Language and Cognition, Zhejiang University, Hangzhou 310028, China)[2]
(School of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China)[3]

Abstract Semantic collocation is the tendency for lexical words to occur together. It is constrained by the syntactic structure and semantic cohesion between its lexical items. Besides conventional semantic collocations, there are also unconventional semantic collocations whose semantic combination is far beyond people's conventional cognitive sense. They are pervasive in all types of writing and considered to be the focus of rhetorical and metaphorical expressions which could be a real difficulty in Natural Language Understanding. In this paper, the characteristics of unconventional semantic collocations and their psychological triggering process are first analyzed and an example-based model is then proposed to automatically discover unconventional semantically collocated word pairs. The experiment is carried out on Chinese verbal collocations. The precision is 81.5% and recall is 80.7%. The experimental results show that the example-based model is effective in detecting unconventional semantic collocations.

Keywords Unconventional semantic collocations, Unconventionality, Metaphor

1 引言

语言的长期发展与积累往往使词语之间的组合形成一种基本的制约,这就是词语之间的搭配。这样的搭配不是任意的,而是受到词性和语法的制约,也受到词义、语境等的制约,这就是"为什么我们说'穿衣'戴帽'而不说'穿帽'戴衣'"[1]的原因。除了"穿衣"戴帽"这种合乎用语习惯的搭配之外,还有一种合乎常规认知的语义搭配,比如我们说"喝矿泉水"这样的动宾搭配短语语义就是合适的。然而在日常语言中,我们还常会遇到另一种搭配现象,如在"他喝过很多墨水"这样的表达中"喝墨水"这样的搭配。这样的语义搭配存在一种偏离常规认知的不和谐,但是我们并不认为"喝墨水"是错误搭配,而是会尽力在非常规的隐含层面上,如修辞或隐喻表达层面上寻找其可协调和可兼容的语义特征。这样的搭配我们称之为语义的超常搭配。由于常规搭配往往不能满足丰富的语言表达的需要,人们往往会通过一些"超常搭配"的手段来形成生动写意的表达。只要这种超常规的搭配没有超过人们正常认知推理的范围,则能使句子生动、形象,而且不抵触语法规范,这也是修辞的核心所在。

当前的自然语言处理领域对词语搭配的研究主要集中在从大规模语料库中抽取正确词语搭配[1-5]。虽然从语言学角度有关于超常搭配修辞效果与认知功能的研究[6],但是却鲜有从汉语信息处理的角度讨论对超常搭配词语的发现问题。然而,由于超常搭配是语言修辞的核心,词语超常搭配的自动发现对从文本中发现修辞表达如隐喻表达有着积极意义,而修辞与隐喻表达的发现与提取对深化自然语言理解内容也有一定意义[7],因此本文提出一种判断汉语词语语义超常搭配的量化计算方法,为处理汉语语篇中的隐喻或其它修辞现象做一些探索性和基础性的工作。

2 语义超常搭配

词语超常搭配主要包括词语间语义的超常搭配和语法的

*)本项目得到国家自然科学基金资助(项目号:60373080)和浙江大学 985 工程资助。杨 芸 博士研究生,主要研究方向为计算语言学、自然语言处理和汉语隐喻计算;李剑锋 硕士研究生,主要研究方向为自然语言处理;周昌乐 教授,博士生导师,主要从事人工智能基础研究,包括隐喻认知计算、理论脑科学和认知逻辑学等;黄孝喜 博士研究生,主要研究方向为自然语言处理和汉语隐喻计算。

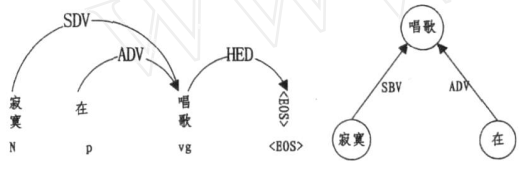
超常搭配。通常情况下,语法的超常搭配主要通过词性在用法上的转变来实现。这样的超常搭配情况在词语超常搭配中占有的比例不多,在语法规则的约束的框架下,容易通过句法分析和语法检查得到。相比之下,语义超常搭配的灵活性更大,从普通语言表达中发现语法合理但语义超常的搭配则要复杂得多,因为语义超常搭配在句法结构上完全符合规则,它的超常性主要受到语义认知和逻辑范畴的制约,与人们的认知知识系统有着密切的联系。例如,"寂寞在唱歌"中,名词"寂寞"与"唱歌"形成主谓搭配,主谓搭配的句法规则为"名词/代词+动词",这样的搭配符合句法,但是"唱歌"的主语从常规认识来讲是人,而不是抽象的概念"寂寞"。不过,可以引入隐喻隐含层面对该短语进行释义,比如将寂寞具体化并赋予它生命,让"寂寞"具有活跃感,从而突出"寂寞"这种状态的强烈等,这就形成了语义上的超常搭配。

2.1 类型及表示

语义的超常搭配的识别与其语法搭配类型有关。典型的搭配结构有主谓搭配、动宾搭配、定中搭配、状中搭配、动补搭配。语义超常搭配表现在表层语义和深层语义的不一致,这里的表层语义是词语在某种搭配类型中的字面组合意义,而深层语义是由超常搭配构成的修辞的隐含真实意义。

为形式化表示每一个搭配,我们借助依存句法分析来对搭配进行结构化表示。根据依存句法,我们将由两个词语构成的搭配关系表示成依存关系,即一个为中心词,另一个词由中心词支配,受支配词以某种搭配类型从属于其支配词。

例如,短语"寂寞在唱歌"表示成依存结构,如图1所示,其中SBV为主谓关系标记,ADV为"状中"搭配关系标记,HED为句子中心词。



(a)依存句法结构　　(b)依存句法树

图1　依存结构示例

"寂寞在唱歌"短语实际包含两个词语搭配,"寂寞"与"唱歌"组成主谓搭配,"在"与"唱歌"形成状中搭配。在超常搭配"寂寞"和"唱歌"中,"唱歌"是中心词,支配"寂寞"。由于语义超常搭配符合句法规则,因此可以通过依存句法分析器[8]将每个搭配表示成结构化的依存关系对。

3　基于实例的语义超常搭配计算

3.1 基本思想

由于语义超常搭配是在合理的语法约束框架之内的语义的动态生成,然而却很难为语义定义规则。从心理语言学角度,我们做如下假设:所有的搭配不可能被全部记录和掌握下来,人们通常会掌握一部分符合常规认知的搭配实例。当遇到新的搭配时,采取将新搭配与同类型的搭配实例进行比较,通过比较新搭配与搭配实例间的相似性或可接受度判断搭配属正常搭配还是超常搭配。人们通过发现搭配单元中形成搭配关系的双方之间表层语义关系(如动作-对象关系、动作-施事关系等)与常规认知之间产生的矛盾来掌握对语义超常搭配的识别。

基于此,我们提出基于实例的语义超常搭配发现方法:首先构建结构化的搭配实例库,确定当前搭配属于实例库中的哪类实例范畴,最后结合实例的搭配关系以及概念知识库(如《同义词词林》[9])进行相关参数的计算,从数值与阈值的比较中发现超常搭配。

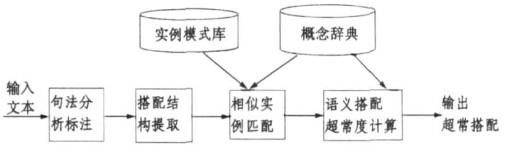


图2　基于实例的超常搭配判断结构

3.2 实例模式库举例

由于以动词为中心的搭配占据搭配现象中的大部分,因此我们将以动词为例构建相应实例库。本文以《常用汉语动词搭配实例库》为蓝本来构建动词搭配语料库,其中搭配关系包括了主要的动词用法,分为主、宾、状、补4种。实例库约70万字,收录1273个常见动词的语法语义搭配信息[5]。

表1　搭配类型标注

搭配类型	标记	搭配类型	标记
[主～]	SBV	[～补]	CMP
[～宾]	VOB	[状～]	ADV

每个动词的搭配结构表示方法为:每一个动词以不超过5行的信息来描述,第一行为中心动词,并用一对大括号标记;第二行为主谓搭配,由标记[SBV]引导;第三行为述宾搭配,由标记[VOB]引导;第4行为动补搭配,由标记[CMP]引导;第五行为状中关系,由标记[ADV]引导。搭配实例词语紧跟在每行标记之后,词语之间以",",号相隔。由于动词具有及物或不及物的不同类型,因此不一定具有全部所列举的4种搭配情况,每个动词包括几种搭配则根据词典语料所提供的信息来进行设置。若不足4种搭配情况的,则按照SBV,VOB,CMP和ADV的先后进行排列,如图3所示。

```
{蛰}
[SBV]蜈子,马蜂,蜜蜂
[VOB]手,脸,人,头,你
[ADV]突然,被,几乎,差点儿
{阻止}
[SBV]我们,他,他们,她,政府,公司
[VOB]前进,前进,发展
[CMP]不了,半天,过
[ADV]一再,拼命,别,不必,何必,竟然,居然,毫不客气地偏偏
```

图3　汉语动词搭配语料库文本示例

搭配库涵盖动词的4种搭配关系,搭配关系保证语义搭配在语法上的正确性。实例库存在的问题是所提供的搭配信息相对有限,而且无法达到涵盖所有常规搭配实例的规模。因此,我们采取了词语语义范畴属性相似度计算的方法,利用现有的知识词典(如《同义词词林》)与实例库结合,从而动态地扩大常规搭配实例的规模。例如,实例库中有[喝 白开水]的实例,则结合同义词词典,我们可以得到更多的常规搭配实例,如[喝 矿泉水][饮 白开水]等。因此,在算法中,常规语义搭配实例将结合概念词典,从而提供丰富的常规语义搭配信息。

3.3 语义搭配超常度量

定义1(语义搭配超常度)　语义搭配超常度是指同一语

境下共现并形成搭配关系的两个词语偏离常规认知的程度，定义为一个取值大于等于 1 的实数 $UnColloca(w_h, w_d)$，其中 w_h 为搭配的中心词语，w_d 为依存结构中的被支配词语。取值越高，则认为越偏离常规认知的语义搭配，如果取值高于常规搭配阈值，则判断为超常搭配：

$$UnColloca(w_h, w_d) = \frac{1}{\max\limits_{i=1,\cdots,n}[\mathrm{Sim}_h(w_h, s_i)\max\limits_{j=1,\cdots,m}\mathrm{Sim}_d(e_{s_ij}, w_d)]} \quad (1)$$

其中，$s_i \in S_h$，S_h 为实例库中与搭配中心词 w_h 最相似的词语集合，$S_h = \{s_i : \mathrm{Sim}(w_h, s_i) > \theta\}$ $(0 < \theta \leq 1)$，本文根据实验数据将 θ 取值为 0.9，变量 n 为集合 S_h 的元素个数。

$e_{s_i} \in E_{s_i}$，E_{s_i} 为与中心词 s_i 形成当前搭配关系的实例集合，变量 m 为集合 E_{s_i} 的元素个数。

$\mathrm{Sim}(a,b)$ $(0 \leq \mathrm{Sim}(a,b) \leq 1)$ 为计算两个词语的语义相似度函数，由于本文强调各属性词语之间的类别相似度，因此，采用《同义词词林扩展版》进行计算，采用一般相似度计算公式：

$$\mathrm{Sim}(a,b) = \frac{\beta}{Dis(a,b) + \beta} \quad (2)$$

其中 β 是可调节参数，本文根据实验需要取值 $\beta = 1.1$；$Dis(a,b)$ 为《词林》中词语 a 和 b 的距离。公式(2)不完全取节点之间的路径长度，而会考虑节点的层次的影响因素。例如，两对节点 (a,b) 和 (c,d) 具有相同的路径长度，但是在词林中所处层次更深的那对节点则具有更大的相似度。

语义搭配超常度的计算过程以“喝矿泉水”和“喝墨水”两个搭配的判断为例：

(1)句法分析得到两个搭配为动宾搭配，搭配中心语都是动词“喝”；

(2)对于搭配中心“喝”，通过《同义词词林》和《常用汉语动词搭配实例库》的双重查询，计算得到 Sim_h(喝，喝) = Sim_h(喝，饮) = 1.000，从而得到实例库中与“喝”最相似的词语集合 $S_{喝}$ = {喝，饮}；

(3) s_1 =“喝”时，s_1 的动宾搭配(VOB)实例集合 $E_{喝}$ = {茶，汽水，白酒，白开水，牛奶，咖啡，药，汤，果汁}；s_2 =“饮”时，其动宾搭配实例集合 $E_{饮}$ = {酒，茶}。计算得：

$\max\limits_{j=1\cdots9}\mathrm{Sim}_d(e_{s_1j}, 矿泉水) = 0.73$

$\max\limits_{j=1\cdots2}\mathrm{Sim}_d(e_{s_2j}, 矿泉水) = 0.61$，

进而根据公式(1)得到：

$UnColloca(喝, 矿泉水) = 1.37$ (小于设定阈值)

计算“喝墨水”搭配，得到：

$\max\limits_{j=1\cdots9}\mathrm{Sim}_d(e_{s_1j}, 墨水) = 0.07$，

$\max\limits_{j=1\cdots2}\mathrm{Sim}_d(e_{s_2j}, 墨水) = 0.01$，进而得到，

$UnColloca(喝, 墨水) = 14.29$ (大于设定阈值)

4 实验结果及讨论

实验从《读者》语料库[5]中抽取 400 个动宾搭配短语，进行语义搭配超常度计算以及对超常搭配和正常搭配的区分判断。

对阈值的设置是在超常搭配发现的准确率与召回率之间寻找一个折衷。一般来讲，阈值越高，准确率越高，召回率越低；阈值越低，准确率越低，召回率越高。本实验根据语料实际测试结果选择了一个比较合适的取值，阈值 = 3.3。表 2 给出了部分实验数据。

表 2 语义搭配超常度计算实验示例

超常搭配				常规搭配			
序号	中心词 w_h	受支配词 w_d	$Uncolloca(w_h, w_d)$	序号	中心词 w_h	受支配词 w_d	$Uncolloca(w_h, w_d)$
1	编织	梦想	32.00	1	编织	花篮	1.82
2	阅读	人生	3.64	2	阅读	书籍	1.00
3	捕捉	歌声	14.54	3	治愈	创伤	1.00
4	摧毁	健康	20.00	4	展示	风景画	1.36
5	撕破	天	3.64	5	流露	气质	1.82
6	酿造	黑暗	32.00	6	建立	帝国	1.36
7	雕镌	人生	14.54	7	失去	目标	1.82
8	敞开	心扉	3.64	8	需要	养料	1.00
9	拥抱	大地	14.54	9	感到	失望	1.00
10	劈开	海浪	1.09 (判断错误)	10	挖掘	宝石	3.63 (判断错误)

从实验结果进一步统计得到语义超常搭配自动发现的召回率与准确率：

$$召回率 = \frac{系统识别的语义超常搭配数}{实际的语义超常搭配总数} = 80.7\%$$

$$准确率 = \frac{系统正确识别的语义超常搭配数}{系统识别的语义超常搭配总数} = 81.5\%$$

(说明：该实验结果基于两个前提，第一个前提是假定依存句法分析器的句法分析结果是准确的，因此依存句法分析器的输出结果进行了必要的修正以保证算法不受算法本身以外的因素的影响；第二个前提是算法仅针对来自正式出版物的语言表达没有问题的句子，而不考虑非正式的错误语句)

结果符合我们语义超常搭配度量的预期，也表明从已有认知实例推断和发现新搭配的性质的方法具有一定的可行性和合理性。实验结果的主要影响因素在于概念词典和实例词典所提供的知识量的大小，这是由于所有对语言的处理系统归根结底都是基于知识的。严格来讲，计算机对语义超常搭配的发现所需要的知识将不能少于人完成同样任务所需要的全部知识。超常的语义搭配是一种普遍但是重现率低的语言现象，是单纯基于概率的方法所无法处理的，因此，处理这类问题需要从一些语言机制中寻找方法。本文从词语语义超常搭配识别的机制入手，提出了一种语义超常搭配的发现方法，利用的是现有的知识与语言机制相结合的办法来发现非常规的语言现象，对于探索语言中隐喻与修辞表达的自动发现有一定的基础意义，以此为基础的相关汉语隐喻表达的自动发现正在开展之中。

参 考 文 献

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

[2] 车万翔，刘挺，等. 面向依存文法分析的搭配抽取方法研究 //自然语言理解与机器翻译 ——全国第六届计算语言学联合学术会议论文集. 中国太原，2001：153-159

[3] Bolshakov I A. Gelbukh A. Heuristics-based Replenishment of Collocation Databases //Ranchhod E M，Mamede N J，eds. PorTAL 2002，LNAI 2389. 2002：25-32

[4] Li W Y，Lu Q，Xu R F. Similarity Based Chinese Synonym Collocation Extraction. Computational Linguistics and Chinese Language Processing，2005，1(1)：123-144

[5] 李剑锋，杨芸，等. 面向隐喻计算的语料库研究和建设. 心智与计算，2007，1(1)：142-146

[6] 张新红，刘锋. 从修辞看词语超常搭配. 伊犁师范学院学报，2003，9(3)：36-39

[7] Zhou C L，Yang Y，Huang X X. Computational Mechanisms for Metaphor in Languages：A Survey. Journal of Computer Science and Technology，2007，22(2)：308-319

[8] 刘挺，马金山，等. 基于词汇支配度的汉语依存分析模型. 软件学报，2006，17(9)：1876-1883

[9] HIT-IRLab. TongYiCi CiLin (Extended_Version). 2005. http://ir.hit.edu.cn/

[10] Schrfe H. C G Representations of Non-literal Expressions //Priss U，Corbett D，Angelova G，eds. ICCS 2002，LNAI2393. 2002：166-176