CN 43-1258/TP
ISSN 1007-130X

计算机工程与科学
Computer Engineering & Science

第 37 卷第 12 期 2015 年 12 月
Vol. 37, No. 12, Dec. 2015

文章编号:1007-130X(2015)12-2294-06

现代汉语通感的自动抽取及映射方向性*

刘洪超,Francesca Striklievers,黄居仁
(香港理工大学中文及双语学系,香港)

摘　要:主要介绍现代汉语中通感(Synaesthesia)句子的自动抽取和感觉域之间的映射规律。通过构建各个感觉领域的词表和词性匹配的方式抽取语料库中的通感句子,采取了两种方法,一种是单纯的多领域感觉词匹配,准确率为 20.78%;第二种方法加入了词性匹配,准确率为 46.37%。主要难点在于五种感觉领域词表中词的选取和收集以及词性分布规则的总结上。最后统计了抽取句子通感源域到目标域的映射情况,检查了其映射方向是否与其他语言相同。

关键词:现代汉语;通感;感觉词;自动抽取
中图分类号:TP391.1　　**文献标志码**:A
doi:10.3969/j.issn.1007-130X.2015.12.015

Automatic extraction and mapping directionality of synaesthetic sentences of modern Chinese

LIU Hong-chao, Francesca Striklievers, HUANG Chu-ren
(CBS, The Hong Kong Polytechnic University, Hong Kong, China)

Abstract: This paper focuses on the extraction and mapping tendencies of synaesthetic sentences in modern Chinese. The extraction applies two kinds of methodologies both based on the perception related word lists. We have constructed five sense word lists of touch, taste, smell, hearing and vision respectively. By checking each list and extracting the sentences with two or more kinds of perception related words, the accuracy of this methodology is 20.78%; by introducing POS distributing tendencies checking, the accuracy rises to 46.37%. The difficulty lies in collecting and further selecting the perception related word and also in observing the POS distributing rules of each perception related word. Finally, we check the mapping directionality of one domain of sense to another one.

Key words: modern Chinese; synaesthesia; perception related word; automatic extraction

1　引言

通感(Synaesthesia),又称连觉,是一种生理现象,主要是指"一种非自觉的跨感觉领域的联系体验"[1]。同时,在语言学上,它又是一种特殊的隐喻,是指用一种感觉词去描述另外一种感觉,如"噪音污濁的空氣"(本文所使用的语料库为台湾中央研究院现代汉语语料库,以下简称 sinica,因此所抽取的句子为繁体中文,为了保持句子的真实可靠,未进行繁简转换),"噪音"是属于听觉的感觉词,但是"污濁"一般只用来形容视觉现象,该句用视觉词"污濁"来描述听觉"噪音"就是一种通感,从隐喻上来说,这里的视觉是源域(Source Domain),听觉是目标域(Target Domain)。之所以称之为特殊的隐喻,是因为其源域和目标域都属于同一个范畴,因此也有人称之为准隐喻[2]。本文主要关注的是语言中的通感句子,而不是认知上的通感疾病或

* 收稿日期:2015-08-07;修回日期:2015-10-19
基金项目:Word Chinese and Their Grammatical Variations: Empirical Studies based on Comparable Corpora (GRF project 543512)
通信地址:香港九龙红磡红荔道 1 号香港理工大学学生宿舍 1909 室
Address: R1909, Student Halls of Residence, the Hong Kong Polytechnic University, 1 Hunglai Rd, Hong Hum, Hong Kong, P. R. China

通感体验。

国外对通感的研究一直是一个比较热门的话题,通感所体现出来的映射方向性被当做一个通用的假设而在各个语言中进行检验,研究领域跨越了神经科学、认知科学和语言学等学科;研究方法也包括了神经科学实验、认知实验和语言学观察及统计等,如文献[1,3,4]。但是,国内对于通感的研究大部分仍然仅限于文学赏析上。

另一方面,对于汉语通感的研究不论是国内还是国际都比较缺乏,只有文献[5,6]等有限的几篇。汉语通感映射方向性研究的一个前提就是要有大量的实际数据作为统计上的支撑,但是文献[5,6]等文章中分析的汉语例子都非常有限。由于通感句子在语料库中的分布比较稀疏,因此采用人工寻找的方法不可取。最好是采用自动抽取的方法。关于通感自动抽取,本文只找到了文献[7],而关于汉语通感自动抽取的研究,目前还未发现。因而,本文力图基于文献[7]的方法实现语料库中汉语通感句子的自动抽取。同时分别构建触觉(Touch)、味觉(Taste)、嗅觉(Smell)、听觉(Hearing)和视觉(Vision)的词库,帮助进行通感句子的抽取和语言学研究。

本文第1节是引言,介绍本文的研究背景和意义;第2节是相关研究的简介,对相关文献进行简单的综述;第3~4节介绍本文的研究方法和结果,其中第3节介绍感觉词的选择标准和感觉词表的呈现,第4节介绍抽取方法和抽取结果,对抽取句子进行统计,寻找映射规律;最后对全文进行总结,指出本文的思路、不足和未来的研究工作重点。

2 相关研究

2.1 感觉域

由于通感是将一个感觉域的词映射到另外一个感觉域,因而就涉及到对感觉域的定义。根据是否将"情感(Emotion)"考虑为一个感觉域,基本上可以分为两派。

大部分的研究,包括文献[2],未将"情感"考虑为一个感觉域,只考虑同时分别构建触觉(Touch)、味觉(Taste)、嗅觉(Smell)、听觉(Hearing)和视觉(Vision)这五种感觉,认为"这是通用的英美文化标准"[8]。有的研究,进一步将这五类进行了细化,如文献[4]把视觉细化为"颜色(Color)"和"维度(Dimension)"。

另外一派,包括文献[6],将"情感"考虑进了研究对象。

本文在构建感觉词表时并未设立"情感"类,主要是由于对于"情感"的定义比较抽象,难以划定一个非常明确的标准。但是,在抽取的例子中,本文发现了大量的其他五类感觉词和情感域之间的映射例子,如:

還是這個心情[target]感受是讓我很低沉[source]的……

"低沉"中"低"属于视觉,"沉"属于触觉,而"心情"则是典型的情感,因此这里是用视觉和触觉来描述心情。这样的例子很多,情感这个抽象类在借助相比之下较为具体的其他五类感觉词来表达具有很强的倾向性,因此本文也将之列入考察范围。

2.2 方向性

文献[2]通过对英语、法语和匈牙利语通感句子的考察发现,这些句子的通感映射通常遵循下面的几条规律:

(1)感觉词从源域到目标域的映射一般都按照下面的方向进行:

触觉→味觉→嗅觉→听觉→视觉

(2)处于映射等级最低端的触觉词被用来表示其他感觉的频次最高;

(3)处于映射等级第二高位的听觉是借助其他感觉词来表达的最高频次的感觉域。

后来的研究,如文献[3,7]等又相继考察了希伯来语、汉语等语言中的情况,用不同的方法力图证明在这些语言中也存在同样的规律。

对该规律做了更好证明的是文献[4],通过考察英语中65个感觉形容词的185例词义演变情况,发现83%以上的词都按照上面的规律进行映射,而其他例外也都做了合理的解释。其绘制的感觉域之间的映射次序如图1所示。

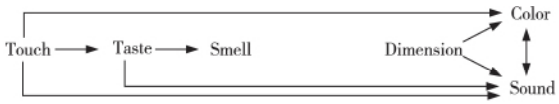

Figure 1 Mapping drectionality of sense-related words in reference[4]

图1 文献[4]感觉词映射路线

由于没有大量的数据进行支撑,因此得出的结论都不是很可靠。如文献[7],其对汉语的考察只通过15个句子就得出汉语通感也是按照图1的路线进行映射转换就不是很有说服力。

本文构建了总量为1 759个的感觉词的词表,抽取了1 452个句子,去除重复后为940个例句,

力图找出汉语映射的规律。

3 感觉词表的构建

本文关于通感的抽取主要就是基于感觉词,因此本文的中心就是要为触觉、味觉等五个感觉域创建尽量完整的词表或词典。在确定收词标准时,本文兼顾了语言学定义和语言工程应用两方面。而本文在收词时主要借助了《现代汉语语义信息词典》和从 sinica 语料库抽取的词表,两者合计共约 20 余万词条,通过半自动的方式,按照我们制定的收词标准从中收集了 1 700 余个感觉词,分别归入了五个感觉词表。

首先是需要对感觉域进行定义,本文参照了文献[8]对几个感觉域的定义标准,即:

(1)触觉|听觉:通过物理性途径刺激形成的感觉,一般说来,触觉的生成都是对皮肤感受器的物理刺激引起的,如"冰凉"是对皮肤温度感受的刺激;而听觉一般是声波对耳膜的震动形成的,这也是一种物理刺激。

(2)味觉|嗅觉:通过化学途径刺激形成的感觉,一般说来,味觉和嗅觉都是鼻腔粘膜或口腔粘膜的感觉细胞对相应的化学分子的刺激反应所产生的感觉。

(3)视觉:通过光线刺激形成的感觉。一般说来,视觉的形成都是光线从物体反射进入眼睛后刺激视神经所产生的感觉。

按照以上的标准,本文利用《现代汉语语义信息词典》中的语义类信息对词典中的词进行了筛选,初步将相关的感觉词筛选出来,本文按照表 1 中的映射关系进行筛选。

Table 1 Mapping between sense domain and the SKCC's semantic domain

表 1 感觉域与抽取的《现代汉语语义信息词典》语义类的对应关系

	形容词	名词	动词
触觉	温度、硬度、湿度、松紧(部分)、触感	生理、人性	
味觉		食物(部分)	
嗅觉	味道	食物(部分)、抽象事物、材料、物性	
听觉	音质	可听现象、乐器、创作物(部分)	五官感觉
视觉	长度、高度、宽度、深度、厚度、大小、视感、外形、颜色、样貌、浓度(部分)、松紧(部分)	可视现象、颜色、外形、创作物(部分)	

然后根据表 1 所抽出的候选感觉词制作关键词(及语素)表,通过关键词和关键语素在 sinica 语料库中抽取出的词表中进行查找,最终两者取合集,去除重复词项之后进行人工校对。

由于《现代汉语语义信息词典》与感觉域不是一一对应的关系,因此,也需要对各个领域的候选感觉词通过手工方法一一归类。

在针对具体的词进行人工校对时,我们也采取了以下的语言学标准:

(1)按照词义进行类别划分,如"滚烫",其中"滚"是属于视觉的语素,"烫"是属于触觉的语素,但是整个词义是表示"非常热",跟视觉并没有关系。由于该词本身是通感词,即在构词层面上看,其词义形成过程中有通感参与,这种通感词进行了特别标注,抽取时含有该词的句子应该直接抽出,因为含有通感词的句子都是通感句。

(2)单纯词的词义主要参照本义,本义无法确定时参照常用义,如"聞",本义属于听觉,在现代汉语中发展出嗅觉的意义,但是本文将之归入听觉。

(3)复合词词义也主要是参照本义,如果词义跨多个感觉域,如"平滑",表示"既平又滑",前者属于视觉,后者属于触觉,则主要参照其搭配对象,一般"平滑"是用于修饰视觉现象,因此,将之归入视觉。

其中数量最多、最难处理的就是第(3)种情况,但是由于本文采取的方法主要是根据句子中是否同时含有多个不同感觉域的感觉词,所以事实上无论将"平滑"这样的词放入哪一类,最终都不会影响抽取的句子总数,只是抽取的具体例句属于哪一类映射上会存在差错。

总体原则就是除了本身就是通感词,如"粗話""冷笑""滚烫"等,需要打上标记(事实上就是兼类标记)之外,其它所有的词都要保持排他性,不处理为兼类,否则就会大大影响句子抽取的准确率。表 2 是对建立的感觉词库的示例(由于采用的语料库为繁体中文,因此本文构建的词库也都是繁体中文版)。

Table 2 A sample of sense-related word database

表 2 感觉词库示例

詞語	意義	語義類	感覺域	構詞情況
鬱鬱	香氣濃厚	境況|味道|性質	Smell	鬱的本義為木叢生也;通感:Vision→Smell
酸	像醋的氣味或味道	境況|味道|品格	Smell	
醇厚	(氣味、滋味等)純正濃厚	性質	Smell	厚:Vision;通感:Vision→Smell

表 3 是本文对感觉词库中各感觉域感觉词分

布的统计结果。

Table 3 Distribution of sense-related words

表 3 感觉词分布情况

感觉词	数量
触觉	154
味觉	85
嗅觉	64
听觉	253
视觉	1 203
总数	1 759

从表 3 可以看出,视觉词占了绝大多数,听觉词次之,最少的是嗅觉词。本文在收集的过程中发现与嗅觉相关的大部分都是形容词,名词和动词都比较少,而且收集难度比较高。嗅觉和味觉有很多词都是词义共通的,即可以同时用于两个领域,只是本文按照本义划定了归属,比如"味道"既可以用于嗅觉,又可以用于味觉,但是其本义是味觉。而且,《现代汉语语法信息词典》中的形容词语义类对嗅觉和味觉根本就没有做区分,两者合在"味道"这个语义类下,可见两者关系比较紧密。

4 抽取实验及结果

在构建了感觉词库的基础上进行通感句子的抽取,本文采用了两种方法,两种方法的基本思路都是基于一个通感句子至少应当含有两个以上感觉域的感觉词这个基本原则进行的。

4.1 方法一

第一种方法的基本步骤如下:

(1)首先对语料库中所有句子进行一遍扫描,将至少含有一种感觉域的感觉词的句子抽取;

(2)然后对这个句子列表进行第二次扫描,如果含有另外一种或多种感觉域的感觉词就将整个句子作为候选句子输出。

表 4 是第一种方法的抽取结果。

Table 4 Result of method 1

表 4 第一种方法实验结果

	抽出的句子总数	通感句子(Token)	通感句子(Type)	准确率/%(Token/总数)
数量	5 073	1 054	554	20.78%

这种方法的准确率比较低,本文对抽出的句子进行分析,发现以下一些原因导致准确率比较低:

(1)未检查词性问题。有的词有不同的词性,不同词性的词有不同的分布规律,但是单纯的关键词匹配不会考虑这一点,如:

例 1 北風正凜冽。

"正"在词表中属于视觉,但是词表中的"正"是分布在谓语位置上的形容词,表示"纯、不杂",如"模样很正"等。而例 1 中的"正"是处于状语位置上的副词"正",表示正在进行。对于这一类问题的解决办法就是考察每一个词的词性分布情况,在抽取时进行检查,如果符合其分布规律就将句子抽出,如果不符合就淘汰。

(2)联合结构问题。有的句子中确实含有两种以上的感觉域的感觉词,但是分别处在联合结构的前项和后项上,两者之间不形成通感,只是单纯的并列,如:

例 2 這種又甜又冷的冰淇淋作風全行不通

"甜"和"冷"分别属于味觉和触觉,但是两者并列,并没有形成通感映射关系。对于这一类问题的解决方法是加入联合结构标志词,如"又…又…""和""并且"等词的检查,但是这样一来又可能去掉一些通感的句子,如:

例 3 周治平依然以清亮[source]而感性的[source]溫柔[source]歌聲[target]……。

"清亮"和"感性"是并列关系,但是两者都与"歌聲"构成了通感关系。

同时,很多没有明显的标志词的并列结构无法用这种方式排除,如:

例 4 炒出來的青菜會軟爛而不好吃。

这里的"軟爛"是并列结构,两者之间并不形成通感,应该被排除,但是仅仅加显性并列结构标志词判断并不能解决这个问题,本文未来将要引入其他句法分析或语义特征的方法将之排除。

4.2 方法二

在对抽出例子分析的基础上,本文着重解决了第一个问题,逐个考察了每个感觉词的词类分布情况,将明显不可能出现在通感句子中的词性从感觉词库中剔除。表 5 是触觉类感觉词的词性分布情况。

Table 5 Touch-related words' distribution in different POS categories

表 5 触觉类感觉词词性分布情况

词性	数量	词性	数量
Nv	25	VHC	5
VA	2	Na	28
VG	1	VJ	1
A	1	VC	6
VH	80	VB	1
DE	1	D	3
总数			154

限于篇幅问题,本文不再列出其他感觉域的词性分布情况,可以看出,大部分的触觉词是 VH(状态不及物动词)(每个符号具体的意义可以访问 sinica 语料库,网址:http://app.sinica.edu.tw/cgi-bin/kiwi/mkiwi/kiwi.sh),加入了词性检查之后,本文将两次结果一并展示如表 6 所示。

Table 6 Results comparison of method 1 and method 2

表 6 两种方法实验结果对比

	抽出的句子总数	通感句子(Token)	通感句子(Type)	准确率/%(Token/总数)
方法一	5 073	1 054	554	20.78
方法二	3 131	1 452	940	46.37

由于本文无法得知 sinica 语料库中通感句子总数的多少,因而无法直接计算召回率等值,但是从表 6 中的具体抽取结果对比可以看出,方法二的效果获得了明显的提升:抽取的句子数在 Type 值上几乎是原来的两倍,准确率是方法一的两倍多。即便是 *TTR*(Type/Token)值也获得了提升(0.53 vs 0.65),说明提取出的句子的类型丰富度,方法二也大大地高于方法一。

4.3 汉语通感映射情况及映射路线

在抽取的大规模通感例句的基础上,本文可以对第二部分提出的映射等级进行检验,检查汉语通感句子是否也遵循了文献[2]提出的映射规律,表 7 是本文对 940 个通感句子的映射情况的统计。

Table 7 Mappings distribution between different sense domains

表 7 各感觉域映射分布

Source	Target						
	Touch	Taste	Smell	Hearing	Vision	Emotion	Total
Touch	0	17	32	129	56	24	258
Taste	6	0	8	20	30	93	157
Smell	0	22	0	3	7	15	47
Hearing	0	0	78	0	0	7	85
Vision	40	49	73	187	0	35	384
Emotion	0	0	1	7	2	0	10
Total	46	88	192	346	95	174	941

需要注意的是,很多句子含有多种通感,因而在计算的时候这些句子都是要分别计算的,如:

例 5 似有似無的(幽$_{[vision/source]}$香$_{[smell/target]}$)$_{[smell/target]}$(飄曳如絲)$_{[vision/source]}$。

例 6 兩種(口味)$_{[taste/source]}$(間雜交錯)$_{[vision/source]}$、(甜$_{[taste/target]}$香$_{[smell/source]}$)$_{[taste/target]}$有味。

在例 5 中,“幽香”在词汇层面形成通感,通感的方向是用属于视觉的“幽”修饰“香”这种嗅觉,源域是视觉,目标域是嗅觉;同时整个“幽香”属于嗅觉,再用属于视觉的“飄曳如絲”描述“幽香”,这又构成了另外一种通感,源域是视觉,目标域是嗅觉。例 6 情况也类似。

因此,最终的通感总数与抽取的句子总数并不是一一对应的。

从表 7 可以观察到以下几个映射的倾向性:

(1)映射几乎可以在任意两个感觉域之间进行,即通感几乎可以在任意两个感觉域之间产生。

(2)听觉域感觉词只能向嗅觉和情感感觉词进行映射。

(3)听觉域接受的通感映射最多,视觉域提供的通感映射最多。换句话说,用其他感觉词来表示听觉的频次最高;视觉词表示其他感觉的频次最高。

这个结论除了第二点与文献[2]得出的结论 3 有相似之处外,其他都不相同。

5 结束语

从语言学上说,本文要研究的问题是汉语通感句子是否存在一定的映射方向性?映射规律是什么?对于这个问题,目前的汉语研究很少,即便有,结论也是建立在少量的例句之上,因此要回答这个问题就要获得大量的通感句子,手工办法不可取,本文采取自动抽取的方法。

为了解决句子的抽取问题,本文采用了两种方法,两种方法都是基于感觉词匹配,因此本文首先构建了感觉词库,这个工作产生两个结果:一方面明确了六种感觉(包括“情感”)的定义,另一方面构建了一个较为全面的感觉词库,方便语言学研究。

从抽取实验结果来看,引入了词性检查的抽取效果明显好于没有词性检查的效果,但是准确率仍然不高,还可以通过引入句法语义分析的方式进一步提高,这是本文下一步要进行的工作。

通过对抽取的大量句子的统计结果来看,汉语通感的映射方向性并不是很明显,但是映射也存在一定的规律,在这里不再重复,下一步本研究要进

行的工作是解释这些规律和找到这些规律形成的原因。

参考文献:

[1] Cytowic R E. Synesthesia: A union of the senses [M]. New York:Springer Verlag, 2002.

[2] Ullmann S. The principles of semantics [M]. Oxford: Basil Blcakwell,1957.

[3] Shen Y, Eisenamn R. "Heard melodies are sweet, but those unheard are sweeter": Synaesthesia and cognition [J]. Language and Literature, 2008, 17(2):101-121.

[4] William J. Synaesthetic adjectives: A possible law of semantic change [J]. Language, 1976, 52(2):461-478.

[5] Sean D. Synaesthesia and synaesthetic metaphors [J]. PSYCHE, 1996,32(2):1-16.

[6] Yen-Han Lin, Shelley Ching-Yu Hsieh. Synaesthetic metaphors of television food commercial ads in mandarin Chinese [J]. TMUE Journal of Language and Literature, 2011(1.6):1-16.

[7] Strik Lievers F. Synaesthesia: A corpus-based study of cross-modal directionality [J]. Functions of Language, 2015, 22(27):69-95.

[8] Yu Ning. Synaesthetic metaphor: A cognitive perspective [J]. Journal of Literary Semantics, 2003, 32(1):19-34.

作者简介:

刘洪超(1987-),**男,山东博兴人,博士生,研究方向为现代汉语语法和计算语言学**。**E-mail**:jiye12yuran@126.com

LIU Hong-chao, born in 1987, PhD candidate, his research interests include Chinese syntax, and computational linguistics.

Francesca StrikLievers(1981-),**女,意大利人,博士生,研究方向为词汇语义学**。**E-mail**:francesca.striklievers@gmail.com

Francesca StrikLievers, born in 1981, PhD candidate, her research interest includes lexical semantics.

黄居仁(1958-),**男,台湾人,博士,教授,研究方向为计算语言学、词汇语义学、知识本体和语料库语言学,汉语语言学**。**E-mail**:churen.huang@polyu.edu.hk

HUANG Chu-ren, born in 1958, PhD, professor, his research interests include computational linguistics, lexical semantics, ontology, corpus linguistics, and Chinese linguistics.