语义空间的研究方法*

鲁忠义　孙锦绣
(河北师范大学 教育学院,石家庄 050091)

摘　要:对于语义空间的研究一直是认知心理学研究的一个热点。由于对词汇语义系统的不同观点,科学家们试图从不同的角度采用不同的方法来进行研究。目前,有代表性的语义空间研究方法主要有两种:潜在语义分析(LSA)和语言的多维空间类比(HAL)。潜在语义分析是指利用奇异值分解的方法来探索文章中潜在的语义关系的方法;语言的多维空间类比则是利用多维量表(MDS)的方法来提取语义信息。

关键词:语义空间;潜在语义分析;语言的多维空间类比
中图分类号:B842.5　　文献标识码:A　　文章编号:1003 - 5184(2007)03 - 0022 - 07

1　引言

语义空间(semantic space)是指由概念与概念、概念与其特征之间的关系所构成的心理上的网络或区域。在 Collins 与 Quillian 关于语义记忆的第一个信息表征的心理学模型——层次网络模型中,可以说表示概念的结点和表示概念与概念之间、概念与其特征之间的连线就构成了一个语义空间。后来,Collins 与 Loftus 对层次网络模型进行了修正,提出了激活扩散模型。在该模型中,用语义距离(semantic distance)代替概念的层次结构来作为组织语义网络的基本原则。语义距离的远近反映了概念之间的联系程度或紧密程度。在 Smith 等人 1974 年的特征比较模型(feature comparison model)中,语义空间也是它研究的一个重要问题。

语义空间在心理表征研究中占有重要地位,问题是用什么样的方法来建构它。心理学家为构建语义空间并进行语义空间的研究采用了一些心理学的方法。比如,层次网络模型中概念与概念、概念与其特征之间的关系就是靠反应时的方法得到的。在涉及语义距离的心理模型中,语义距离的测量也有一些方法,如让被试确定一对一对概念之间的联系程度;评价某一总括概念的各个成员的典型性;还有就是根据列出的某一范畴的各个成员的频率来确定它们之间的密切程度。

可以说,上述这些研究方法是认知心理学中研究语义空间的一些直接的方法。但是对于从一个大的语料库(corpus)中去确定两个概念的关系,不仅相当困难,而且往往会掩盖概念间的潜在关系。Deerwester 等人就认为,在信息检索中词语选择的任意性会部分地掩盖数据当中的一些潜在的结构,但使用某种统计技术可以估计出这种潜在的结构并去掉模糊的噪声(noise)[1]。这样一些信息科学专家另辟蹊径提出了语义空间研究的新方法,如 Deerwester 等人的潜在语义分析(latent semantic analysis, LSA)[1],其中的合作者 Landauer 和他的同事们还将这种方法应用到了认知心理学的研究中[2,3];Burgess 等人的语言的多维空间类比(hyperspace analogue to language, HAL)[4-6]。随着这种研究思路不断被人们接受,又有很多人提出自己的研究方法,如 Rohde 等人将这两种方法综合起来提出了词汇语义的相关共现类比(correlated occurrence analogue to lexical semantic, COALS)[7]等等。在目前的语义空间的研究方法中,潜在语义分析与语言的多维空间类比是两种得到广泛认可的方法。

这些语义空间的研究方法的总思路是:首先要构建一个语义空间,即选择一个大的语料库,然后选用合适的数学模型来简化它,从而得到一个能够测量词语相似性的语义空间。然后利用这个语义空间来模拟人的各种心理现象,并进行统计学的推论。

2　潜在语义分析

潜在语义分析探索了一种知识获得与表征的新理论,提供了从一个大语料库中提取的词语与文章或文章片断相似性的表征方法。Landauer 等人认为对相似性的估计不是依赖简单的接近率,或者共现

* 基金项目:国家社会科学基金(04BYY008)。

的数目和相关，而是要依赖一项能够正确地、有力地推论出非常深刻关系的数学分析。潜在语义分析所使用的数学分析叫做奇异值分解(singular value decomposition，SVD)，这是一种类似于因素分析的数学矩阵的分解技术。利用奇异值分解技术就可以简化大的语料库，从而生成一个较小的表征相似性的语义空间，这个过程就叫做降维(dimension reduction)。降维的过程就是将表面的浅层信息与内部深层的抽象信息发生联系的过程，并且捕捉到了词与文章或文章片断所共有的隐含信息。因此，运用该技术的一个重要的部分就是为最后的表征确定一个最佳的维度，即语义空间。这个最佳的维度可以类比于通常所说的词的基本意义的语义特征[1-3]。

潜在语义分析的具体方法是：根据一个词语在句子中出现的频率作出一个 $m \times n$ 的长方形原始矩阵 {A}，m 是行，n 是列，通过奇异值分解可以形成三个矩阵：{U}、{V} 和 {S}。{U}、{V} 代表两个相互正交矩阵，{S} 代表一个对角矩阵，即奇异值矩阵。需要说明的是在做奇异值分解之前要将原始矩阵 {A} 中的各词出现的频数进行加权处理，这个过程可以称之为标准化(normalization)，常用的有 TF-IDF (term frequency-inverse document frequency)方法和熵(Entropy)方法[8]等等，然后再对矩阵 {A} 进行奇异值分解，得到 {U}、{V}、{S} 三个矩阵，这三个矩阵与原矩阵的关系为：

$$SVD(A) = [U * S * V]$$

利用奇异值分解的方法分解完标准化后的矩阵 {A} 后，再从对角矩阵 {S} 中取排在前面的 n 个值(根据经验选取)，这样对角矩阵 {S} 就变成了 {S'}，然后再进行奇异值分解的逆运算，就得到原矩阵的近似矩阵 {A'}。在所得的近似矩阵 {A'} 中，通过分析就可以得出原始矩阵 {A} 中没有的信息[3,9]。

下面就举例[3]来说明潜在语义分析的具体过程。

以下是九个技术备忘录的题目，前五个是关于人和计算机交互作用的题目，后四个是关于数学图论的题目，两个题目之间是没什么关系的。每一题目作为列，每一题目中至少出现 2 次的实词(斜体字标出)作为行，这样就构成了一个 9 列 12 行的原始矩阵 {A}。

C1：*Human* machine *interface* for ABC *computer* applications

C2：A *survey* of *user* opinion of *computer system response time*

C3：*The EPS user interface* management *system*

C4：*System* and *human system* engineering testing of *EPS*

C5：Relation of *user* perceived *response time* to error measurement

M1：The generation of random，binary，ordered *trees*

M2：The intersection *graph* of paths in *trees*

M3：*Graph minors* Ⅳ：Widths of *trees* and well-quasi-ordering

M4：Graph minors：A *survey*

{A} =

	C1	C2	C3	C4	C5	M1	M2	M3	M4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

$r_{(human,user)} = -0.38$

$r_{(human,minors)} = -0.29$

{U} =

0.22	-0.11	0.29	-0.41	-0.11	-0.34	0.52	-0.06	-0.41
0.20	-0.07	0.14	-0.55	0.28	0.50	-0.07	-0.01	-0.11
0.24	0.04	-0.16	-0.59	-0.11	-0.25	-0.30	0.06	0.49
0.40	0.06	-0.34	0.10	0.33	0.38	0.00	0.00	0.01
0.64	-0.17	0.36	0.33	-0.16	-0.21	-0.17	0.03	0.27
0.27	0.11	-0.43	0.07	0.08	-0.17	0.28	-0.02	-0.05
0.27	0.11	-0.43	0.07	0.08	-0.17	0.28	-0.02	-0.05
0.30	-0.14	0.33	0.19	0.11	0.27	0.03	-0.02	-0.17
0.21	0.27	-0.18	-0.03	-0.54	0.08	-0.47	-0.04	-0.58
0.01	0.49	0.23	0.03	0.59	-0.39	-0.29	0.25	-0.23
0.04	0.62	0.22	0.00	-0.07	0.11	0.16	-0.68	0.23
0.03	0.45	0.14	0.01	-0.30	0.28	0.34	0.68	0.18

{S} =

3.34	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	2.54	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	2.35	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	1.64	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	1.50	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	1.31	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.85	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.56	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.36

{V} =

0.20	0.61	0.46	0.54	0.28	0.00	0.01	0.02	0.08
-0.06	0.17	-0.13	-0.23	0.11	0.19	0.44	0.62	0.53
0.11	-0.50	0.21	0.57	-0.51	0.10	0.19	0.25	0.08
-0.95	-0.03	0.04	0.27	0.15	0.02	0.02	0.01	-0.03
0.05	-0.21	0.38	-0.21	0.33	0.39	0.35	0.15	-0.60
-0.08	-0.26	0.72	-0.37	0.03	-0.30	-0.21	0.00	0.36
0.18	-0.43	-0.24	0.26	0.67	-0.34	-0.15	0.25	0.04
-0.01	0.05	0.01	-0.02	-0.06	0.45	-0.76	0.45	-0.07
-0.06	0.24	0.02	-0.08	-0.26	-0.62	0.02	0.52	-0.45

在对角矩阵 {S} 中保留最大的两个奇异值，即 3.34 和 2.54，然后作奇异值分解的逆运算：

$$\{A'\} = [U] * [S'] * [V]$$

得到原始矩阵 {A} 的近似矩阵 {A'}：

{A'} =

	C1	C2	C3	C4	C5	M1	M2	M3	M4
human	0.16	0.04	0.38	0.47	0.18	-0.05	-0.12	-0.16	-0.09
interface	0.14	0.37	0.33	0.40	0.16	-0.03	-0.07	-0.10	-0.04
computer	0.15	0.51	0.36	0.41	0.24	0.02	0.06	0.09	0.12
user	0.26	0.84	0.61	0.70	0.39	0.03	0.08	0.12	0.19
system	0.45	1.23	1.05	1.27	0.56	-0.07	-0.15	-0.21	-0.05
response	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
time	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
EPS	0.22	0.55	0.51	0.63	0.24	-0.07	-0.14	-0.20	-0.11
survey	0.10	0.53	0.23	0.21	0.27	0.14	0.31	0.44	0.42
trees	-0.06	0.23	-0.14	-0.27	0.14	0.24	0.55	0.77	0.66
graph	-0.06	0.34	-0.15	-0.30	0.20	0.31	0.69	0.98	0.85
minors	-0.04	0.25	-0.10	-0.21	0.15	0.22	0.50	0.71	0.62

$r_{(human,user)} = 0.94$

$r_{(human,minors)} = -0.83$

无论是原始矩阵 {A} 在经过标准化后的矩阵中，还是在经过奇异值分解所得到的近似矩阵 {A'} 中，任意两行词语间的相关关系可以用余弦值来表示：

$r = \cos\theta = \alpha \cdot \beta / (\| \alpha \| \cdot \| \beta \|)$，其中 $\alpha = (a_1, a_2, a_3, \cdots a_n)$，$\beta = (b_1, b_2, b_3, \cdots b_n)$，$\alpha \cdot \beta = (a_1b_1 + a_2b_2 + a_3b_3 + \cdots a_nb_n)$，$\| \alpha \| = (a_1^2 + a_2^2 + a_3^2 + \cdots + a_n^2)^{1/2}$，$\| \beta \| = (b_1^2 + b^2b_2 + b_3^2 + \cdots b_n^2)^{1/2}$

以上就是潜在语义分析的过程，从结果中可以看出 在原始矩阵 {A} 标准化后的矩阵中 $r_{(human,user)} = -0.38$，而在奇异值分解后得到的近似矩阵 {A'} 中 $r_{(human,user)} = 0.94$，表明在原始矩阵中低度相关（-0.38）的 human 与 user，经过了奇异值分解后，却成了高相关（0.94）；trees 在 M4 列中没有出现，trees 在 M3 中出现一次，经过奇异值分解后，在 M4 中 trees 由 0 变成了 0.66。这一过程说明潜在语义分析将其中的潜在关系发掘了出来[3]。

值得一提的是，Landauer 等人对一个大的语料库，包括美联社的新闻（Associated Press news wire）、

美国大百科辞典(Grolier 's Academic American Encyclopedia)和有代表性的学生读物,做奇异值分解,得到了一个较小的语义空间(潜在语义空间),然后根据这个较小的语义空间来完成托福考试(TOEFL)的同义词测试。为了模拟人的心理相似性表征,他们把测试词与每个选项之间的余弦值(相似性)都计算出来,然后,再算出潜在语义空间中的该测试词与各个选项之间的余弦值,将两者进行比较,发现答案的正确选择率为 65%,相当于在美国申请大学的非英语国家的学生的平均成绩[3]。

潜在语义分析是以创建文章词语的矢量表征(vector representation)为基础来捕捉文章的语义信息的。潜在语义分析的过程就是将高维度语义空间中的词语矢量、文章矢量投射到低维度的潜在语义空间中,使得原来在高维度语义空间中的稀疏矩阵(parse matrix)在经过了奇异值分解所得的低维度语义空间中变得不再稀疏,原来看起来没有任何共同词汇的两篇文章,经过奇异值分解之后,就可以找到它们之间在意义上的相似性[3,10]。潜在语义分析的主要功能是通过比较矢量的表征来计算相似性,利用奇异值分解这种矩阵的代数方法在本质上使得文章中的词语在矢量空间上重新排列和重新定位。

3　语言的多维空间类比的研究方法

Burgess 等人认为,语言的多维空间类比是一个记忆表征的模型,是在一个大的语料库中获得词语的表征的。记忆不是一个静止的过程,而是一个动态的过程,环境与心理表征之间的动态关系提供了系统的基础,这个系统的一个重要的特征就是自组织(self – organization)。在语言的多维空间类比中,概念的获得过程是通过词语全部共现(global co – occurrence)的方式,将上下文中简单的联想整合成概念的表征的过程[11]。

Burgess 等人从 Usenet newspaper 中选择了一个 3 亿 2 千万词的语料库,利用窗口长度等于 10 个词共现(co – occurrence)的方式来呈现这个语料库。Burgess 等人认为庞大的矢量空间是能够通过保留较大变异的列来达到降维的。这个降维的过程就是要去掉一些词,可以设定一个频率的数值作为标准,小于此值的频率都要去掉。另外,诸如" the "" of "之类的对于语义信息贡献不大的词(stop words)也要去掉[12]。总之,通过一些手段将原始矩阵处理,使之标准化,之后才能够在这个标准化的矩阵中作两词之间相似性的测量。Burgess 等人据此得到了一个 140,000 维度的语义空间,然后使用多维量表法(multidimensional scaling, MDS)对数据进行分析。多维量表法是一种数据分析技术,以几何图片的形式展现距离相近的两数据点的结构关系。这种方法假设每一个物体或事件都可以在高维空间表征为一个点,这些点就排列在高维空间中,这些点之间的距离与它们所表征的事物的相似性之间一定有很强的相关。所以,表征两个相似的物体的点就会离得近一些,反之,表征两个不相似的物体的点就会离得远一些[11,13,14]。相似性的测量使用的是明考斯基(Minkowski)距离的测量方法,其中包括欧氏距离(Euclidean)的测量[5,6,14]。

语言的多维空间类比的基本的方法是在一篇文章中设定窗口长度为 n 个词(n – word)作为共现词,n 的取值最小是 1,最大不要超过工作记忆的容量太多,比如说 n 取 10 个词,那么共现词的呈现是从 1 个词开始逐一递增到 10 个词,也就是说是通过移动共现窗口,使词一个一个地增加并记录窗口中的权重值(见下例和矩阵 {E} 与矩阵 {F} 中的数字),这些权重值(weights)就组成了一个共现矩阵。对于一篇文章中的每一个词语都可以作为靶词(target vocabulary),那么共现词的出现会有两种情况,即共现词出现在靶词之前与共现词出现在靶词之后,共现词出现在靶词之前的权重值组成了共现矩阵的横行,共现词出现在靶词之后的权重值组成了纵列,横行与纵列成对地连接在一起,就形成了一个 $n \times n$ 的共现矩阵,同时对应每一个词语也得到了一个共现长度为 $2n$ 的向量。这个长度为 $2n$ 的向量可以被认为表征了一个处在 $2n$ 高维空间的词语[5,6,11,13]。

下面以" The horse raced past the barn fell "为例[5,6]来说明语言的多维空间类比的研究过程。在这个例子中窗口长度等于 5 个词,例子中数字表明的是权重值,括号内的词为共现词。共现词的呈现分两种情况:

1)共现词出现在靶词之前:

(The) **horse** raced past the barn fell.
　5

(The horse) **raced** past the barn fell.
　4　5

(The horse raced) **past** the barn fell.
　3　4　5

(The horse raced past) **the** barn fell.
　2　3　4　5

(The horse raced past the) **barn** fell.
　1　2　3　4　5

The(horse raced past the barn)**fell**.
1 2 3 4 5

2)共现词出现在靶词之后：

The horse raced past the **barn** (fell).
5

The horse raced past **the**(barn fell).
5 4

The horse raced **past**(the barn fell).
5 4 3

The horse **raced** (past the barn fell).
5 4 3 2

The **horse**(raced past the barn fell).
5 4 3 2 1

The(horse raced past the barn) fell.
5 4 3 2 1

将 1)所得的权重值作为横行，将 2)所得的权重值作为纵列，得到矩阵 {E}。

{E} =

	barn	Horse	past	raced	the	fell
barn	0	2	4	3	6	0
horse	0	0	0	0	5	0
past	0	4	0	5	3	0
raced	0	5	0	0	4	0
the	0	3	5	4	2	0
fell	5	1	3	2	4	0

将横行与纵列联合在一起得到原始矩阵 {F}。

{F} =

barn	0	2	4	3	6	0	0	0	0	0	0	5
horse	0	0	0	0	5	0	2	0	4	5	3	1
past	0	4	0	5	3	0	4	0	0	0	5	3
raced	0	5	0	0	4	0	3	0	5	0	4	2
the	0	3	5	4	2	0	6	5	3	4	2	4
fell	5	1	3	2	4	0	0	0	0	0	0	0

一旦原始矩阵形成，就要对这个原始矩阵进行降维处理，使之标准化，然后涉及到词语间的相似性的测量，使用明考斯基距离的测量方法，一般是能够测量矩阵中所有词对的。Burgess 等人所计算出来的词语之间的 D 值是在以 3 亿 2 千万的语料库为背景所得到的一个 140,000 维度的语义空间中计算出来的。需要注意的是共现窗口大小不同，计算出来的 D 值是不一样的。

语言的多维空间类比模型的过程就是通过全部共现的方式将较大的语料库中的词语逐一进行编码，即形成不同的权重值，形成一个原始矩阵，然后将原始矩阵降维，得到一个较小语义空间，在这个语义空间中每一个词语都表征了一个语义符号，每一个符号都是移动窗口中背景构成的一部分，然后利用多维量表法计算两个以及多个词语之间的相似性，或者利用这个语义空间进行模拟研究。Burgess 等人证明了词语矢量的语义相似性与词汇判断作业的启动效应有关，他们还使用词语矢量对分类问题成功地作了模拟。

4 综合讨论

潜在语义分析与语言的多维空间类比是要试图解答诸如“环境输入是如何转化成表征信息的”、“意义是如何从实际经验中获得的”、“人类知识表征的相似性是如何获得的”、“人类是如何在他们所得到的少量的信息的基础上获得那么多知识的”等等问题[2,11]。

潜在语义分析就是要从一个较大的语料库中获得词与词、词与文章或文章片断、文章与文章或文章片断与文章片断之间的相似性的表征，体现了人类获得知识的一般学习机制。这个模型借助于奇异值分解的方法来诱导出这种相似性[2,3]。

Burgess 等人认为意义在语言、认知和经验间架起了一座桥梁，构成了语言理解的系统。语言的多维空间类比要选取一个大的语料库，在词语的全部共现后，利用多维量表法这种心理测量学的方法来发展语义记忆的模型[11,13]。

这两种理论在研究方法的思路方面相同点还是很多的 都是要选择一个大的自然语料库，呈现这个语料库，形成一个高维语义空间，然后借助数学的方法来简化高维语义空间的维数，最后得到一个低维的相似性语义空间，这样就可以利用这个低维语义空间来进行各种心理现象的模拟研究了。在语义空

间的研究中都用到了大的语料库。因为大的语料库包括的内容广,不会将呈现的文章内容局限在一个特定的领域,而且能够反映不同类别的词语的广泛分布,可以达到统计学所要求的水平[15]。不同点在于具体的操作过程:所选的大的自然语料库是不同的,呈现语料库的方式不同,构建矩阵的方法不同,所使用的降维的数学模型不同,最后得到的低维语义空间的维数是不同的。但这些都不会影响它们的研究。

目前,潜在语义分析已经广泛地应用到预测心理现象,做心理语言学实验,范围从认知词语发展性的获得词语的分类,句子 – 词语的语义启动,篇章理解,文章相似性的研究,文章质量的判断以及训练学生写作并进行写作成绩的评估等等。语言的多维空间类比也应用到研究大的认知现象当中,比如说:联结语义启动、语义语法分类、隐含定义、语义判断、深度诵读困难、概念获得、决策等等。

尽管潜在语义分析与语言的多维空间类比在语义空间的研究中产生了重大的影响,并且在认知心理学的应用研究方面取得了不少成果,但是,潜在语义分析与词汇语义的多维空间类比仍然面临着许多难题[16]:

1)不能解决由于上下文背景变化所带来的语义空间的变化。语义空间理论假设词是静止的,固定在语义空间中的一个位置。这样就完全忽视了词的位置对上下文背景的依赖,忽视了词在语义空间中变化。

2)不能解决语法抽象结构问题,尤其是这些抽象结构被分离呈现时。这是共现技术本身存在的缺陷,所以,只能提取表层的信息表征,却不能提取深层的信息表征。

3)缺乏对世界本质的认识。人类的知识信息是通过与世界的直接接触获得的,当人们作判断时,一定是使用了大量的非共现所得的信息,所以知识获得过程的深层次内容是无法用共现的程序考察的,这需要从学习和直接的经验中获得。

4)词语不是原子实体(atomic entities)。这里他们举了一个例子:

Flugly as the name of glamorous Hollywood actress.

Flugly as the name of an accountant in a W. C. Fields movie.

从字面上看,Flugly 只是个名字,但是,对于一名好莱坞的女明星来说,Flugly 绝对不是一个好名字,而对于会计师来说,它就是一个很好的名字。这与组成这个词的字母有关:把它当作不佳的名字时,会听到一个不太好听的喉音"g",而且 Flugly 中还包含了"ugly"(丑陋);当认为它是会计师的好名字时,人们会把它听成"Flugleeee"。

这说明词语中所含有的一部分字母所包含的信息对整个词义的判断起着关键的作用。共现的程序对这些信息是不敏感的,即使共现到字母与音节的水平也还是解释不清楚的。

除此之外,两种方法还有其自身的一些具体问题。Landauer 本人就认为潜在语义分析没有考虑到词的顺序、句法关系、逻辑关系和词法[3],语言的多维空间类比虽然考虑了词的顺序问题,但是,同潜在语义分析一样仍然没有解决句法关系、逻辑关系和词法等问题[9]。

另外,就操作方法来说,一个多大的自然语料库,或者说高维空间的维数是多少,就能够包含或类比人脑中的所有信息,两种方法对这一点看法也是不一致的。最后所得到的低维语义空间的维数是多少才能够进行确定意义的心理表征的模拟研究,两种方法得出的研究结果也不相同。潜在语义分析认为最后的维数大约是 300 个,而语言的多维空间类比则将维数定为 140,000。所以,还需对这类方法进行深入研究。

总之,人类从环境中获得的信息经过人脑的加工形成意义表征,这个过程决不是一个简单的过程。人脑所存储的知识,绝不是用一个大的语料库能够代替的,人脑对信息的加工过程也绝不是用一种数学模型就能够分析清楚的。环境与意义表征之间所联结的中间过程是一个精细的、动态的、交互的过程。任何单一的模型都不足以充分的解释这个过程。语义空间对于此的研究即使存在这样那样的问题,但它提供了一种研究的角度与方法,人们还是能从这个新视角中获得新的认识。揭示人脑对环境信息的加工所形成的意义表征的过程是一个浩大的工程,应该将语言学、心理学、认知科学、计算机科学、生物学、神经生理学以及社会科学联系在一起,本着多学科多方法多水平的思路继续进行研究。

参考文献

1　Deerwester S, Dumais S T, Furnas G W, et al.. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 1990, (41): 391 – 407.

2　Landauer T K, Dumais S T. A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and

representation of knowledge. Psychological Review, 1997, (104): 211 – 240.

3 Landauer T K, Foltz P W, Laham D. An introduce to latent semantic analysis. Discourse Process, 1998, (25): 259 – 284.

4 Burgess C, Cottrell G. Symposium at the cognitive science society conference: using high – dimensional semantic spaces derived from large text corpora. In: Proceedings of the Cognitive Science Society. Hillsdale, NJ: Erlbaum Publishers, 1995. 13 – 14.

5 Lund K, Burgess C. Producing high – dimensional semantic spaces from lexical co – occurrence. Behavior Research Methods, Instrumentation, and Computers, 1996, (28): 203 – 208.

6 Burgess C. From simple associations to the building blocks of language: Modeling meaning in memory with the HAL model. Behavior Research Methods, Instruments, & Computers, 1998, (30): 188 – 198.

7 Rodhe D L T, Gonnerman L M, Plaut D C. An improved method for deriving word meaning from lexical co – occurrence. Cognitive Science, (in press).

8 复旦大学. 概率论基础. 高等教育出版社, 1983. 184 – 197.

9 Rehder B, Schreiner M E, Wolfe M B, et al.. Using latent semantic analysis to assess knowledge: some technical considerations. Discourse Processes, 1998, (25): 337 – 354.

10 刘云峰, 齐欢, 代建民, 等. 中文信息的潜在语义分析. 华南理工大学学报(自然科学版增刊), 2004, (32): 107 – 111.

11 Burgess C, Lund K. The dynamics of meaning in memory. In: Conceptual and representational change in humans and machines. Lawrence Erlbaum Associates, 2000. 117 – 156.

12 Song D, Bruza P D, Cole R J. Concept learning and information inferencing on a high dimensional semantic space. In: ACM SIGIR 2004 Workshop on Mathematical/Formal Methods in Information Retrieval (MF/IR 2004). Sheffield, UK, 2004.

13 Burgess C. Representing and resolving semantic ambiguity: A contribution from High – dimensional memory modeling. In: On the consequences of meaning selection: Perspectives on resolving lexical ambiguity. Lawrence Erlbaum Association, 2002. 233 – 261.

14 Young F W. Multidimensional Scaling. Encyclopedia of Statistical Sciences, 1985. 5.

15 Wettler M, Rapp R. Computation of word associations based on the co – occurrences of words in large corpora. In: Proceedings of the workshop on very large corpora: Academic and industrial perspectives. Columbus/ Ohio, 1993. 83 – 84.

16 French R M, Labiouse C. Four problems with extracting human semantics from large text corpora. In: Proceedings of the 24th Annual Conference of the Cognitive Science Society. NJ: LEA, 2002.

Research Methods in Semantic Space

Ru Zhongyi Shun Jinxiu

(Education College, Hebei Normal University, Shijiazhuang 050091)

Abstract: The research on semantic space has always been regarded as a hot area. Because of the different standpoints in this area, scientists try to adopt a variety of methods to study it. Currently, the most influential methods in semantic space are latent semantic analysis (LSA) and hyperspace analogue to language (HAL). LSA makes use of the singular value decomposition (SVD) and HAL resorts to multidimensional scaling (MDS) to investigate the semantic similarity relationship in a large corpus.

Key words: semantic space; latent semantic analysis; hyperspace analogue to language