文章编号：1003-0077（2011）00-0000-00

汉语语义选择限制知识自动获取研究*

贾玉祥[1]，王浩石[1]，昝红英[1]，俞士汶[2]，王治敏[3]

（1. 郑州大学信息工程学院，河南省 郑州市 450001；

2. 北京大学计算语言学教育部重点实验室，北京市 100871；

3. 北京语言大学汉语学院，北京市 100083）

摘要：语义选择限制刻画谓语对论元的语义选择倾向，是一种重要的词汇语义知识，对自然语言的句法、语义分析具有重要作用。本文研究汉语语义选择限制知识的自动获取，提出基于 HowNet 和基于 LDA（Latent Dirichlet Allocation）的两种知识获取方法，对方法进行了实验对比与分析。实验表明，前者所获取的知识可理解性更好，后者所获取的知识应用效果更好。两种方法具有很好的互补性，我们提出了一个二者的融合方案。

关键词：语义选择限制；知识获取；HowNet；LDA（Latent Dirichlet Allocation）

中图分类号：TP391　　**文献标识码：**A

Research on Chinese Selectional Preferences Acquisition

JIA Yuxiang[1], WANG Haoshi[1], ZAN Hongying[1], YU Shiwen[2], WANG Zhimin[3]

(1. School of Information Engineering, Zhengzhou University, Zhengzhou, Henan 450001, China;

2. MOE Key Laboratory of Computational Linguistics, Peking University, Beijing 100871, China;

3. College of Chinese Studies, Beijing Language and Culture University, Beijing 100083, China)

Abstract: Selectional preference describes the semantic preference of the predicate for its arguments. It is an important lexical knowledge which can be applied to syntactic and semantic analysis of natural languages. This paper studies the automatic acquisition of Chinese selectional preferences and proposes a HowNet based method and a LDA (Latent Dirichlet Allocation) based method. A comparative study shows that the former method acquires better understood knowledge while the latter achieves better performance in applications. The two methods are complementary and we propose a combination strategy for them.

Key words: selectional preference; knowledge acquisition; HowNet; LDA (Latent Dirichlet Allocation)

1 引言

一个句子不是词语的随意组合，除了要满足语法约束外，还需要满足语义约束。比如 Chomsky 著名的例子“Colorless green ideas sleep furiously.”，在语法上是正确的，但（从常识上讲）却不符合语义，因此没有意义。因为 sleep 的主语要求是人或动物，green 修饰的应该是具体的事物，colorless 修饰 green 及 furiously 修饰 sleep 都存在矛盾。

语义约束的一个主要体现是谓语（如，动词）对充当其句法成分（如，主语、宾语等）或语义角色（如，施事、受事等）的词语（论元）在语义上具有选择性，称为语义选择限制（Selectional Restriction / Selectional Preference, SP）。比如，动词“吃”（eat）的主语（或施事）更倾向于选择“人或动物”类的词语，宾语（或受事）更倾向于选择“食物”类的词语。

SP 知识可以用一个四元组<*p*,*r*,*a*,*sp*>或函数 *sp*(*p*,*r*,*a*)来表示，*p* 表示谓语（或谓词，predicate，可以是词或语义类），*r* 表示句法成分（或语义角色，role），*a* 表示名词论元（argument，可以是词或语义类），*sp* 是选择优先度，为一个实数值，表示谓语 *p* 选择 *a* 充当其句法成分（或语义角色）*r* 的倾向性，*sp* 越大，越倾向于选择 *a*。根据 *sp* 值从大到小对

* **收稿日期：**　　**定稿日期：**

基金项目：国家博士后科学基金（2011M501184）；河南省博士后科研资助（2010027）；计算语言学教育部重点实验室（北京大学）开放课题（201301）；国家自然科学基金（60970083, 61170163, 61272221）；国家社会科学基金（14BYY096）；国家 863 计划项目（2012AA011101）；河南省科技厅科技攻关计划项目（132102210407）。

论元名词或语义类进行排序，排在前面的可用于构建 SP 知识库。

SP 知识获取就是对任意的<*p*,*r*,*a*>，给出其对应的 *sp* 值。对于 *r* 来说，选择句法还是语义只是分析深度的区别，语义角色比句法成分更接近语义本质，手工创建的 SP 知识库一般选语义角色，但是从目前知识自动获取的实际和在自然语言处理中的应用角度出发，一般选择句法成分。

语义选择限制是重要的词汇语义知识[1]，除了可以用来判断句子的合法性之外，还具有数据平滑和消歧作用，因此被用于自然语言处理的很多任务，包括句法分析[2]、语义角色标注[3]、词义消歧[4]、指代消解[5]、隐喻计算[6]等，在信息抽取、问答系统、机器翻译等方面也有潜在的应用。

汉语研究者在语义选择限制知识库建设方面做了很多工作，也开展了一些语义选择限制规律的探索[7,8]，但语义选择限制知识自动获取方面的研究还相对较少[9]。本文研究汉语语义选择限制知识的自动获取，对比考察了基于语义分类体系的方法 HowNet-SP 和基于分布的方法 LDA-SP，并对两种方法的融合提出了一个可行的方案。本文的章节安排如下：第 2 节介绍相关研究工作，第 3 节介绍两种知识获取方法，第 4 节给出实验结果与分析，第 5 节提出一个知识获取方法的融合方案，第 6 节给出总结和展望。

2 相关研究

语义选择限制是词汇知识库的重要组成部分。剑桥大学等构建的综合语言知识库描述了动词对名词的语义选择限制，规定了动词主体和客体的语义类。VerbNet 为每一类动词涉及的相关语义角色描述选择限制。北京大学现代汉语语义词典以义项为单位描述了实词的配价信息和多种语义组合限制。清华大学等构建的现代汉语述语动词机器词典以义项为单位，描述每一个义项涉及的论旨角色的典型语义类。HowNet[10]描述的语义关系中的施事/经验者/关系主体-事件关系、受事/内容/领属物-事件关系等也体现了语义选择限制。柏晓鹏[11]在建立现代汉语词义分类体系时，把选择限制作为词语描述的属性之一。

Resnik[4]最先提出语义选择限制的自动获取，结合 WordNet 和真实语料获得英语动词对宾语语义类的选择限制。继英语之后，德语、法语、拉丁语、荷兰语、汉语、日语、韩语、泰语等多种语言都开展了 SP 自动获取的研究。除面向语言学方面的研究之外，SP 在自然语言处理方面也得到了广泛应用。

SP 获取的关键是论元扩展，即基于已知的论元实例<*p*,*r*,*a*>，计算未知论元（没有与 *p* 在语料中共现）*a'*的 *sp* 值，即 *sp*(*p*,*r*,*a'*)。根据论元扩展中是否使用语义分类体系，可以将 SP 获取方法分为两类。

第一类是基于语义分类体系的方法。该方法借助语义分类体系（如 WordNet），计算谓语对论元语义类的 *sp* 值，那么对于未知论元，只要它出现在某一个语义类中，就可以给它一个 *sp* 值。这里的关键是语义类 *sp* 值的计算，Resnik 使用一个基于相对熵的统计指标，Li 和 Abe[12]基于最小描述长度模型，Clark 和 Weir[13]基于假设检验。对于面临的一词多义问题，Judea 等[14]通过只考虑没有歧义的 Wikipedia 中的实体论元加以规避，Ciaramita 和 Johnson[15]则结合贝叶斯模型来加以处理。这类方法的优点是学习出的知识是关于语义类的排序，而不是简单的词语排序，易于人类理解，便于集成到词汇知识库中。缺点是需要一个语义分类体系，且由于词典收词有限会导致论元覆盖率比较低。这类方法主要面向语言学研究和词汇知识库构建。

第二类是基于分布的方法。该方法不需要语义分类体系，而是利用词语在语料中的分布情况来实现论元的扩展，具体模型包括基于概率的模型、基于向量空间的模型、基于机器学习的模型等。基于概率的模型把 *sp* 值定义为一个关于 *p*、*r*、*a* 的概率，计算 *sp* 就是估计概率值。其中最常用的是隐变量模型[16]（如 Latent Dirichlet Allocation, LDA），隐变量可以看成一个个隐含的语义类，把谓语和未知论元联系起来。基于向量空间的模型[17]利用大规模语料构建一个向量空间，通过在该空间里计算未知论元和已知论元的相似度，把谓语和未知论元联系起来。基于机器学习[5]的方法直接对论元进行二分类：合适的论元和不合适的论元，把分类器给论元的打分作为 *sp* 值。Tian 等[18]通过在谓语论元搭配图上的随机游走算法来解决未知论元问题和 *sp* 值的计算。基于分布的方法优点是不依赖语义分类体系，论元覆盖率高，对一词多义问题能更好地处理，易于和其他自然语言处理任务结合。缺点是学习出的知识是词语的列表，与语义类列表相比，不易于人类理解。这类方法主要面向自然语言处理，

也是 SP 获取的主流方法。

SP 知识获取方法的评价可以有三种途径：一是与人的判断进行一致性比较，由人制定标准测试集。二是伪消歧（pseudo-disambiguation）[19]，自动构建测试集。三是嵌入自然语言处理任务。从以往研究可以看出，基于分布的方法一般要好于基于语义分类体系的方法，基于分布的方法的各种具体模型的表现各有优劣。

3 汉语 SP 获取方法

对于汉语语义选择限制知识的获取，对比考察基于语义分类体系的方法和基于分布的方法。基于语义分类体系的方法采用 Resnik[4]的统计指标和 HowNet 的分类体系，基于分布的方法采用 LDA 模型。

3.1 基于语义分类体系的方法——HowNet-SP

假设谓语动词 v，论元角色 r，名词语义类 c，定义谓词的选择优先强度（selectional preference strength, SPS）为论元语义类的后验概率分布和先验概率分布之间差异，用相对熵表示，体现谓词对论元语义类的选择性，值越大选择性越强，如“吃”（SPS = 0.585318）对宾语的选择性要比“想”（SPS = 0.185432）对宾语的选择性强。

$$SPS(v,r) = D(P(c \mid v,r) \,\|\, P(c \mid r)) = \sum_{c} P(c \mid v,r) \log \frac{P(c \mid v,r)}{P(c \mid r)} \tag{1}$$

谓词 v 的论元角色 r 选择语义类 c 的优先度（selectional preference, sp）即选择关联度（selectional association, SA）定义如下，

$$sp_{HowNet}(v,r,c) = SA(v,r,c) = \frac{1}{SPS(v,r)} P(c \mid v,r) \log \frac{P(c \mid v,r)}{P(c \mid r)} \tag{2}$$

即该语义类对谓词选择优先强度的贡献，体现了该语义类用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，如“edible|食物”（SA= 0.313351）作为“吃”的宾语的选择关联度大于“stone|土石”（SA = 0.000482528）。

谓词 v 的论元角色 r 选择某一名词 n 的优先度定义为 v 对 n 所属的所有语义类的选择优先度的最大值：

$$sp_{HowNet}(v,r,n) = \max_{n \in c} sp_{HowNet}(v,r,c) \tag{3}$$

使用最大似然估计方法来估计概率 $P(c|r)$及 $P(c|v,r)$，如公式 4、5 所示。

$$\hat{P}(c \mid r) = \frac{freq(r,c)}{\sum_{c'} freq(r,c')} \tag{4}$$

$$\hat{P}(c \mid v,r) = \frac{freq(v,r,c)}{freq(v,r)} \tag{5}$$

文本中出现的是词 w，不是语义类 c。用词频 $freq(r,w)$（统计 w 作为角色 r 出现的次数，比如 w 作为动词宾语出现的次数）或共现词频 $freq(v,r,w)$来估计语义类出现的频率 $freq(r,c)$或共现频率 $freq(v,r,c)$，需要借助语言知识本体（语义分类体系），这里使用 HowNet。一个词可能有多个义项，每个义项对应于 HowNet 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$|classes(w)|$。

$$freq(r,c) = \sum_{w \in c} \frac{1}{|classes(w)|} freq(r,w) \tag{6}$$

$$freq(v,r,c)=\sum_{w\in c}\frac{1}{|classes(w)|}freq(v,r,w) \quad (7)$$

3.2 基于分布的方法——LDA-SP

概率主题模型 LDA（Latent Dirichlet Allocation）[20]是一种有效的文档表示模型，把文档看做隐含主题的随机混合，隐含主题看做词的分布。该模型既可以挖掘文本中潜在的语义信息，又可以降低文档表示的维度。

这里把描述文档词项共现的 LDA 模型迁移到谓词论元共现的描述，把谓词（如动词）看做文档，把论元（如做动词宾语的名词）看成词项，把论元的语义类看成隐含主题。这样基于 LDA 的语义选择限制表示模型称为 LDA-SP，图模型表示如图 1 所示。

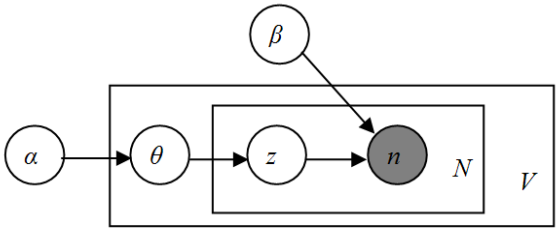

图 1 LDA-SP 的图模型表示

空心点表示隐含随机变量或参数，实心点表示可观察值，箭头代表依赖关系。矩形表示重复过程，右下角是重复次数。大矩形表示从 Dirichlet 分布中为每个谓词 v 反复抽取语义类分布 Θ_v，共 V 个谓词。小矩形表示从语义类分布中反复抽样产生谓词的论元名词，共 N 个名词。

LDA-SP 的具体生成过程如下：

1）对每一个谓词 v，选择隐含语义类上的一个多项式分布 Θ_v，Θ_v 是参数为 α 的 Dirichlet 分布。

2）对每一个语义类 z，选择论元名词上的一个多项式分布 Φ_z，Φ_z 是参数为 β 的 Dirichlet 分布。

3）生成一个谓词 v 的论元名词 n，先以分布 Θ_v 从隐含语义类中选择一个语义类 z，再以分布 Φ_z 从论元名词中选择一个论元名词 n。

模型生成的结果可以用公式（8）表示：

$$P(n|v,r)=\sum_{z}P(n|z)P(z|v,r) \quad (8)$$

在此基础上可以定义谓词对论元语义类和论元名词的选择优先度：

$$sp_{LDA}(v,r,z)=P(z|v,r) \quad (9)$$

$$sp_{LDA}(v,r,n)=P(n|v,r) \quad (10)$$

LDA-SP 两个重要的参数是各语义类下论元名词的概率分布 $P(n|z)$和各谓词的语义类概率分布 $P(z|v,r)$。参数估计可以采用期望最大化（Expectation Maximization ,EM）算法和 Gibbs 采样等方法。给定参数 α，β，语义类个数 T 及谓词论元搭配集，就可以出训练参数 $P(n|z)$及 $P(z|v,r)$。

4 实验与分析

我们选择动宾关系、主谓关系来对语义选择限制知识获取进行评价。对人民日报 2000

年全年语料使用哈工大语言技术平台进行依存句法分析，抽取动词-名词宾语对 935319 对、动词-名词主语对 459913 对。对于 LDA-SP 模型，忽略只出现一次的动词，使用 GibbsLDA++来实现，主题（语义类）数量设为 200，迭代次数设为 2000，其他参数为缺省设置。

4.1 优选语义类

基于语义分类体系的方法可以获取动词优选语义类的列表，基于分布的方法一般获得的是词语的列表。LDA-SP 方法中的隐含变量 z 是词语的聚类，相当于语义类。表 1 给出“吃”、“喝”、“写”、“唱”四个动词的宾语最优先选择的语义类的情况，即 SA 最大的 Class 及 $P(z|v,r)$ 最大的 z。可见，HowNet-SP 与 LDA-SP 方法所获取的优选语义类与人的认知基本一致。比较而言，以语义类表示的前者要比后者更清楚更易于理解。给隐含变量标注语义类标签将是提高 LDA-SP 方法所获取知识的可理解性的手段。

表 1 优选语义类比较

Verb	Class	SA	z	$P(z\|v,r)$
吃	edible|食物	0.313351	鱼 饭 肉 羊 菜 牛 油 猪 鸡 动物 …	0.792391
喝	drinks|饮品	0.298229	水 酒 泉 段 啤酒 咖啡 配偶 汤 袋 茶 …	0.736486
写	text|语文	0.300648	作品 节目 书 电视 电影 诗 文章 广告 …	0.487836
唱	music|音乐	0.530643	消息 歌 声 声音 主旋律 国歌 音乐会 …	0.641984

文献[7]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“attribute|属性”，“举行”的宾语语义类是“fact|事情”。本文对文献[7]中的所有动词，从语料库中自动获取对宾语的语义优选，得到动词对各层次所有语义类的选择优先度。表 2 给出每一个动词选择关联度 SA 最大的宾语语义类。可见，大部分的语义类都是符合常识的。但是结果还是受一些因素的影响：（1）语料的规模。受语料库规模影响，一些动宾搭配的频率比较小，比如“改掉”只有一个宾语“陋习”、“建筑”只出现 8 次、“震惊”出现 20 次、“改正”出现 21 次，这些都可能影响所获取语义类的质量。（2）语料的领域。本文是新闻领域语料，某些搭配的分布很不平衡，比如“附”的宾语基本都是“图片”，因此优选的语义类是“image|图像”。（3）文本自动分析的错误。分词、词性标注、句法分析等的错误会导致搭配抽取的错误，如“计算机爱虫病毒”这句话里把“虫”分析成“爱”的宾语，由于这样的分析出现了 49 次，直接导致“爱”最优选的语义类是“InsectWorm|虫”。（4）HowNet 中的词汇知识没有充分利用。HowNet 中名词出现在多个语义分类体系中，除“entity|实体”外，还有“attribute|属性”等，这里只用了“entity|实体”，导致不少名词成了未登录词因而被忽略。另外，这里使用词语定义中的第一义原来表示词语所属的语义类，在有些情况下，第一义原并不明确反映词语的语义类，真正有用的义原是其他义原，这一问题也有待解决。

表 2 动词宾语优选语义类

Verb	Class	SA	Verb	Class	SA
改掉	aspiration|意愿	0.688822	建	building|建筑物	0.298582
附	image|图像	0.548819	震撼	human|人	0.267717
采取	method|方法	0.512735	改革	component|部分	0.283378
震惊	place|地方	0.255565	来自	place|地方	0.307283
发挥	purpose|目的	0.53406	开放	InstitutePlace|场所	0.4799
举行	fact|事情	0.519061	表现	emotion|情感	0.415085

召开	fact|事情	0.411358	包含	component|部分	0.479924
修改	regulation|规矩	0.300159	发展	affairs|事务	0.292651
建筑	InstitutePlace|场所	0.210283	使用	artifact|人工物	0.305338
改正	result|结果	0.48526	爱	InsectWorm|虫	0.287053
制定	regulation|规矩	0.339154	表示	emotion|情感	0.57535
解决	problem|问题	0.309949	支持	organization|组织	0.489517
发生	phenomena|现象	0.516249	维持	organization|组织	0.202872
伤害	emotion|情感	0.442184	保持	document|文书	0.210524
会见	human|人	0.334778	形成	regulation|规矩	0.289593
具有	information|信息	0.549563	服务	part|部件	0.258473
成立	organization|组织	0.400433	提高	human|人	0.327911
举办	fact|事情	0.579078	发现	phenomena|现象	0.293797
建造	building|建筑物	0.340912	开展	affairs|事务	0.536131
进入	time|时间	0.421497	说明	reason|道理	0.277248
克服	phenomena|现象	0.538424	改变	mental|精神	0.423832
完成	affairs|事务	0.608286	成为	place|地方	0.292817
解释	phenomena|现象	0.16597	包括	physical|物质	0.387996

语言中的隐喻表达可以看做是一种搭配异常，比如“编织梦想”“嫁接资本”等就是由动宾搭配异常而形成的语言创新用法。获得动词的优选语义类，进而获得动词字面用法下的优选语义类（字面语义类，如“嫁接”的宾语字面语义类“plant|植物”），对隐喻的判别和理解都有重要的作用。我们选择 10 个常用于隐喻用法的动词，考察其宾语或主语优选语义类（选择关联度 SA 最大的语义类）的获取情况（见表 3）。

可见，“嫁接”、“提炼”等给出了准确的字面语义类。“编织”的宾语有多个“网”（有一个语义类是“internet|因特网”），其实“网”前有修饰词，如“关系”，这里就形成了动词隐喻和名词隐喻的嵌套（如编织关系网），比较难以处理。“medicine|药物”泛滥、“fund|资金”流入，本身是隐喻用法，已成为动词最常用的搭配。解剖“part|部件”和“part|部件”瘫痪，这里的“part|部件”就是一个不能准确反映词语语义类信息的义原，比如“身体”这个词的定义是“DEF=part|部件,%AnimalHuman|动物,body|身”，第一义原“part|部件”不如义原“body|身”更能反映“身体”的语义类信息。要准确获得动词的字面语义类，可以结合概念的抽象和具体信息，具体的概念更易于成为字面语义类。比如，作为“滑坡”的主语，“stone|土石”的 SA 值小于“experience|感受”，但是前者是具体概念，后者是抽象概念，可以过滤掉后者，而得到字面语义类“stone|土石”。

表 3 隐喻动词优选语义类

Verb	Object	SA	Verb	Subject	SA
编织	internet|因特网	0.138548	泛滥	medicine|药物	0.203829
嫁接	plant|植物	0.255273	滑坡	experience|感受	0.209901
解剖	part|部件	0.367822	流入	fund|资金	0.218189
净化	InstitutePlace|场所	0.359427	碰撞	time|时间	0.213532
提炼	material|材料	0.25639	瘫痪	part|部件	0.485951

4.2 伪消歧

语义选择限制获取的一个标准评价方法是伪消歧（pseudo-disambiguation）。伪消歧最初是用来评价词义消歧的，词义消歧评价的一个难点就是需要人工来标注标准测试集。为了减少人工标注的工作量，提出了伪消歧这种自动构建测试集的方法。具体做法是（以动宾搭配为例）：从语料中自动抽取动宾搭配集，认为都是正确的搭配。对每一个搭配<*v*,*r*,*n*>，基于某一种策略自动选择另一个名词 *n'*来代替 *n*，形成伪搭配<*v*,*r*,*n'*>，即错误的搭配，然后判断哪一个搭配是原搭配哪一个是伪搭配。假定如果原搭配的选择优先强度 *sp*(*v*,*r*,*n*)大于伪搭配的选择优先强度 *sp*(*v*,*r*,*n'*)即为判断正确。

评价指标采用覆盖率（coverage）和正确率（accuracy），定义如公式 11 和 12 所示。四元组<*v*,*r*,*n*,*n'*>形成一个测试样本。如果<*v*,*r*,*n*>和<*v*,*r*,*n'*>都有 *sp* 值，那么称该测试样本被覆盖（covered）。如果 *sp*(*v*,*r*,*n*) > *sp*(*v*,*r*,*n'*)，则判断正确（correct）；如果 *sp*(*v*,*r*,*n*) = *sp*(*v*,*r*,*n'*)，则强度相等（tie）；否则为判断错误。

$$coverage = \frac{\#covered\ samples}{\#all\ samples} \times 100\% \quad (11)$$

$$accuracy = \frac{\#correct + \#tie * 0.5}{\#covered\ \ samples} \times 100\% \quad (12)$$

测试数据使用 1998 年 1 月的人民日报语料（使用哈工大语言技术平台进行依存句法分析），从中抽取动词和名词宾语搭配，要求：（1）动词和名词的频率在 20 和 300 之间。（2）动宾搭配频率大于 2。（3）动词和名词都是二字词。这样得到 1952 个不同的动宾搭配，通过人工校对最后确定搭配 1329 对，包含 373 个动词和 386 个名词。

给每一个搭配中的名词，选择一个替代词。替代词的选择可以有不同的策略，比如随机选择、选择相近词频的词等[19]。我们选择一个更加严格的策略，先对名词按词频从大到小降序排列，然后用直接前驱词替代目标名词。目标词和替代词一起形成一个测试样本。

从训练数据中去掉测试样本中的所有搭配，包括原搭配和伪搭配，这样来保证所有的测试样本对模型来说都是没有见过的，更能反映所获取的语义选择限制知识的泛化能力和数据平滑能力。

表 4 总体结果

	covered	correct	tie	coverage	accuracy
HowNet-SP	831	472	209	62.53%	69.37%
LDA-SP	1329	1086	0	100%	81.72%

表 5 HowNet-SP 模型结果举例

v	*n*	*n'*	*c*	*c'*	*sp*(*v*,*r*,*n*)	*sp*(*v*,*r*,*n'*)	*result*
把握	机遇	观念	time|时间	entity|实体	0.274294	0.254048	correct
摆脱	危机	意见	thing|万物	mental|精神	0.141892	0.226636	wrong
表明	决心	合同	thing|万物	thing|万物	0.174385	0.174385	tie

表 6 LDA-SP 模型结果举例

v	*n*	*n'*	*sp*(*v*,*r*,*n*)	*sp*(v,r,*n'*)	*result*
把握	机遇	观念	0.00318787	0.00349969	wrong
摆脱	危机	意见	0.00261737	0.000580109	correct

总体结果如表 4 所示。可见，LDA-SP 模型在覆盖率和正确率上都比 HowNet-SP 模型好。LDA-SP 模型的覆盖率是 100%，而 HowNet-SP 模型的覆盖率是 62.53%。一个原因是我们只使用了分类体系“entity|实体”（HowNet 2000 版的名词语义分类体系“entity|实体”包括 142 个义原，涵盖 27267 个词），而有些名词则属于其他的语义分类体系，如“主题”“困境”“内容”“局面”等词都属于“attribute|属性”。考虑更多的语义分类体系可能会提高覆盖率。对于被覆盖的样本，HowNet-SP 模型的正确率也比 LDA-SP 模型低很多。

表 5 给出了 HowNet-SP 模型的几个例子，*c* 是包含 *n* 且 *sp*(*v*,*r*,*c*)最大的语义类。表 6 给出 LDA-SP 模型的两个例子。可见，LDA-SP 错误的例子在 HowNet-SP 中是正确的。总体上，79 个样本（约占样本总数的 5.94%）在 LDA-SP 是错的，但 HowNet-SP 是正确的，所以两个模型的结合可以进一步提高实验结果。

5 方法的融合

基于语义分类体系和基于语料库分布的方法有很强的互补性。从理论上说，二者的结合可以充分利用词汇语义知识和语料库分布信息，从而获得更理想的语义选择限制知识。从实验结果看，二者的结合也会使知识获取的质量得到提升。这里尝试为两种方法的融合提出一个可行的方案。

把 SP 知识获取分成两个步骤。第一步是获取基础论元搭配，形成基础搭配库。第二步是论元扩展。基础搭配可以从一个较小规模的语料中自动获取，也可以融合各种知识源，比如搭配词典、树库 treebank 等。通过计算已知论元和未知论元之间的相似度来实现论元的扩展并得到选择优先度 *sp*。论元相似度计算可以把词汇语义知识和语料库分布信息融合起来。

谓语对一个论元的选择优先度 *sp* 定义为该论元与基础搭配库中该谓语的所有已知论元的相似度的加权组合[17]，如公式 13 所示。

$$sp_{sim}(p,r,a_0)=\sum_{a\in Seen(p,r)} weight(p,r,a)\cdot sim(a,a_0) \qquad (13)$$

权值 *weight*(*p*,*r*,*a*)可以用来区分不同的论元类型，若设为 1，表示所有类型的论元统一看待；也可以根据基础搭配的数据来源设置不同的权值，如搭配词典高于树库、树库高于自动获取的搭配等。相似度 *sim*(*a*,a_0)的计算可以基于词汇知识库与语料库。基于词汇知识库的方法利用词典中的信息建立词语之间的关联并计算相似度，如英语基于 WordNet，汉语基于 HowNet。语料库方法基于分布性假设，即语义相似的词语通常有着相似的上下文，具体实现有基于向量空间的模型和基于概率的模型，基于深度学习的词语表示方法也可以用于计算词语相似度。

这里把论元相似度定义为两种方法计算所得相似度的线性组合，如公式 14 所示。

$$sim(a,a_0)=\alpha\cdot sim_{LKB}(a,a_0)+\beta\cdot sim_{DIST}(a,a_0) \qquad (14)$$

其中，$\alpha+\beta=1$，$0\le\alpha\le1$，$0\le\beta\le1$。sim_{LKB} 表示基于词汇知识库的方法，sim_{DIST} 表示基于语料库分布的方法，两个相似度都归一化到[0,1]。这样就会给每一个<*p*,*r*,*a*>计算一个 *sp* 值，对每一个<*p*,*r*>，把论元 *a* 按照 *sp* 从大到小排序值，得到一个论元列表，即语义选择限制知识。

6 总结与展望

本文研究汉语语义选择限制知识的自动获取，分别基于 HowNet 和 LDA 模型实现了基于语义分类体系和基于分布的知识获取方法，对知识获取的结果进行了比较与分析。基于语义分类体系的方法所获得的优选语义类易于人类理解，而基于分布的方法所获取的知识在自然语言处理中有更好的应用效果。两种方法有很好的互补性，我们提出了一个二者的融合方案。本文的实验结果还比较初步，下一步将对方法进行改进和优化，扩大数据规模，考察更多的谓语论元类型，考察句法分析等数据预处理中的错误对结果的影响。实现方法融合，对不同方法进行更深入的对比研究。

参考文献

[1] Y. Wilks. A Preferential Pattern-seeking Semantics for Natural Language Inference [J]. Artificial Intelligence, 1975, 6: 53-74.

[2] Guangyou Zhou, Jun Zhao, Kang Liu, et al. Exploiting Web-Derived Selectional Preference to Improve Statistical Dependency Parsing [C]//Proceedings of ACL2011, 2011, 1556-1565.

[3] 邵艳秋, 穗志方, 吴云芳. 基于词汇语义特征的中文语义角色标注研究[J]. 中文信息学报, 2009, 23(6): 3-10.

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

[5] Shane Bergsma, Dekang Lin, Randy Goebel. Discriminative Learning of Selectional Preference from Unlabeled Text [C]//Proceedings of EMNLP2008, 2008, 59-68.

[6] Yuxiang Jia, Shiwen Yu. Unsupervised Chinese Verb Metaphor Recognition Based on Selectional Preferences [C]//Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation (PACLIC 22), 2008, 207-214.

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

[8] 李斌. 现代汉语动宾搭配的语义分析和计算[D]. 南京师范大学博士学位论文, 2009.

[9] 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[C]//第四届全国学生计算语言学研讨会(SWCL 2008), 2008, 90-96.

[10] 董振东. HowNet [EB/OL]. http://www.keenage.com.

[11] 柏晓鹏. 现代汉语词义分类体系的建立和自动标注[D]. 新加坡国立大学博士学位论文, 2012.

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

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

[14] Alex Judea, Vivi Nastase, Micheal Strube. Concept-based Selectional Preferences and Distributional Representations from Wikipedia Articles [C]//Proceedings of LREC2012, 2012, 2985-2990.

[15] M. Ciaramita, M. Johnson. Explaining away ambiguity: Learning verb selectional preference with Bayesian networks [C]//Proceedings of COLING2000, 2000, 187-193.

[16] Diarmuid ´O S´eaghdha. Latent variable models of selectional preference [C]//Proceedings of ACL2010, 2010, 435-444.

[17] Katrin Erk, Sebastian Pado, Ulrike Pado. A Flexible, Corpus-driven Model of Regular and Inverse Selectional Preferences [J]. Computational Linguistics, 2010, 36(4): 723-763.

[18] Zhenhua Tian, Hengheng Xiang, Ziqi Liu, et al. A Random Walk Approach to Selectional Preferences Based on Preference Ranking and Propagation [C]//Proceedings of ACL2013, 2013, 1169-1179.

[19] Nathanael Chambers and Dan Jurafsky. Improving the use of pseudo-words for evaluating selectional preferences[C]//Proceedings of ACL2010, 2010, 445-453.

[20] D. Blei, A. Ng, M. Jordan. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research, 2003, 3:993-1022.

贾玉祥（1981—），男，博士，讲师，主要研究领域为自然语言处理。Email: ieyxjia@zzu.edu.cn

王浩石（1994—），男，本科生，主要研究领域为人工智能。Email: 439303605@qq.com

咎红英（1966—），女，博士，教授，主要研究领域为自然语言处理。Email: iehyzan@zzu.edu.cn