

# 基于 MDL 和 LSC 的语义优选方法

李东明<sup>1</sup>, 张丽娟<sup>2</sup>, 赵 伟<sup>1</sup>, 石 晶<sup>2</sup>

(1. 吉林农业大学信息技术学院, 长春 130118; 2. 长春工业大学计算机科学与工程学院, 长春 130012)

**摘 要:** 为实现谓语动词对论元的自动选择, 提出基于最小描述长度(MDL)和潜在语义聚类(LSC)的语义优选方法。基于 MDL 原则计算与动词搭配的名词的  $\delta sc$  值, 根据 LSC 模型的 EM 算法求取动词、名词的搭配概率  $P(v, n)$ , 并针对每一对动词、名词计算  $\delta sc$  和  $P(v, n)$  之和, 将其作为衡量两者语义关联度的标准。实验结果表明, 该方法的  $F1$  值达到 85.26%, 优于单独使用 MDL 或 LSC 方法。

**关键词:** 语义优选; 最小描述长度; 潜在语义聚类; 无指导学习; 期望极大化

## Semantics Preference Method Based on MDL and LSC

LI Dong-ming<sup>1</sup>, ZHANG Li-juan<sup>2</sup>, ZHAO Wei<sup>1</sup>, SHI Jing<sup>2</sup>

(1. College of Information Technology, Jilin Agricultural University, Changchun 130118, China;

2. College of Computer Science and Engineering, Changchun University of Technology, Changchun 130012, China)

**【Abstract】** To solve automatic predicate-verb choosing for argument, this paper gives semantics preference method based on Minimum Description Length(MDL) and Latent Semantic Clustering(LSC). MDL is used to calculate  $\delta sc$  of each verb-noun pair. The probabilities of a verb preferring for a noun  $P(v, n)$  is computed based on LSC model and EM is used to evaluate the parameters. For the same verb-noun pair, the sum of  $\delta sc$  and  $P(v, n)$  is considered to represent the association between the verb and the noun. Experiments show the  $F1$  reaches 85.26%, and it is better than MDL or SCL methods.

**【Key words】** semantics preference; Minimum Description Length(MDL); Latent Semantic Clustering(LSC); unsupervised learning; Expectation Maximization(EM)

DOI: 10.3969/j.issn.1000-3428.2011.17.004

### 1 概述

语义优选通常指谓语动词对其论元的限制性选择, 特定的谓语动词倾向于选择特定的名词作为宾语, 反之亦然。这一点在汉语的语言学中早就有所涉猎<sup>[1]</sup>, 但从计算机处理的角度进行研究相对较晚<sup>[2-3]</sup>。语义优选在自然语言处理(NLP)中通常被作为提高语言分析效率的有利工具<sup>[4-12]</sup>。目前语义优选的获取方法大致分为 3 种: 本体方法, 统计模型方法<sup>[13-16]</sup>以及两者结合的方法<sup>[6]</sup>。3 种方法的对比研究可参见文献<sup>[17]</sup>。绝大部分关于语义优选的研究集中在英语<sup>[14-17]</sup>, 对汉语的关注相对较少<sup>[18-20]</sup>。

本文提出一种基于最小描述长度(Minimum Description Length, MDL)原则的无指导学习汉语语义优选的方法, 并将该方法与潜在语义聚类(Latent Semantic Clustering, LSC)方法结合起来。本文采用无指导策略, 所用语料库仅需分词和词性标注, 不涉及深层次语法标注, 也不需要本体的支持, 符合汉语目前语言资源匮乏的现状。

### 2 MDL 原则

#### 2.1 原则简介

给定数据序列  $x^m = x_1 x_2 \dots x_m$  及固定的概率模型  $M$ ,  $x^m$  相对于  $M$  的随机复杂度为  $sc(x^m : M)$ , 表示利用  $M$  对  $x^m$  进行编码所需的最小代码长度。MDL 原则<sup>[21]</sup>认为  $sc$  值越小的模型, 越有可能生成数据序列  $x^m$ 。

#### 2.2 原则应用

MDL 应用于语义优选时, 一般作为本体库的概念层选择标准, 见文献<sup>[7]</sup>。本文以一种新的方式利用 MDL, 无需本体库而直接从语料库中学习语义优选知识。

对于数据序列  $(v_1, n_1), (v_2, n_2), \dots, (v_m, n_m)$ , 其中,  $(v_i, n_i)$  表示动词  $v$  和名词  $n$  在第  $i$  个句子中的共现状况,  $v$  和  $n$  的取值为 0 或者 1, 1 表示存在, 0 表示不存在。假定  $v^m = v_1 v_2 \dots v_m$ ,  $n^m = n_1 n_2 \dots n_m$ , 若模型  $I$  表示  $n$  的存在与否独立于  $v$ , 则  $n^m$  相对于  $I$  的  $sc$  值计算如下:

$$sc(n^m : I) = total \times H\left(\frac{freq_n}{total}\right) + \frac{1}{2} \times \ln \frac{total}{2 \times \pi} + \ln \pi \quad (1)$$

若模型  $D$  表示  $n$  存在与否依赖于  $v$ , 则  $n^m$  相对于  $D$  的  $sc$  值计算如下:

$$sc(n^m : D) = (freq_v \times H\left(\frac{freq_{v,n}}{freq_v}\right) + \frac{1}{2} \times \ln \frac{freq_v}{2 \times \pi}) + (freq_{-v} \times H\left(\frac{freq_{-v,n}}{freq_{-v}}\right) + \frac{1}{2} \times \ln \frac{freq_{-v}}{2 \times \pi} + \ln \pi) \quad (2)$$

于是:

$$\begin{aligned} \delta sc &= \frac{1}{total} (SC(n^m : I) - SC(n^m : D)) = \\ &= H\left(\frac{freq_n}{total}\right) - \frac{freq_v}{total} H\left(\frac{freq_{v,n}}{freq_v}\right) - \frac{freq_{-v}}{total} H\left(\frac{freq_{-v,n}}{freq_{-v}}\right) - \\ &\quad - \frac{1}{2 \times total} \ln \left( \frac{freq_v \times freq_{-v} \times \pi}{2 \times total} \right) \end{aligned} \quad (3)$$

**基金项目:** 吉林省科研发展计划科技支撑基金资助重点项目(20100214); 吉林省科技发展计划青年基金资助项目(20100155)

**作者简介:** 李东明(1979—), 男, 讲师、硕士, 主研方向: 智能信息处理, 信息论; 张丽娟, 讲师、硕士; 赵 伟, 教授、博士; 石 晶, 讲师、博士

**收稿日期:** 2011-04-07 **E-mail:** crystal1087@126.com

其中,  $H(z) = -z \times \lg z - (1-z) \times \lg(1-z)$ ,  $0 < z < 1$ , 若  $z=0$  或  $z=1$ , 则  $H(z) = 0$ ;  $total$  表示语料库中所有句子的数目;  $freq_n$  表示名词  $n$  出现的句子数目;  $freq_v$  表示动词  $v$  出现的句子数目;  $freq_{v,n}$  表示动词  $v$  和名词  $n$  共同出现的句子数目;  $freq_{\neg v}$  表示不出现动词  $v$  的句子数目;  $freq_{\neg v,n}$  表示出现名词  $n$  但不出现动词  $v$  的句子数目。根据 MDL 原则,  $\delta_{sc}$  值越大, 名词  $n$  越依赖于动词  $v$ 。所以, 从统计学角度讲,  $\delta_{sc}$  可以很好地表示名词与动词的语义关联。

### 3 LSC 模型

#### 3.1 模型简介

潜在语义聚类(LSC)模型最早由文献[22]提出用于语法结构消歧, 文献[23]基于该模型实现词义消歧、句法结构聚类。

假定动词集合  $V = \{v_1, v_2, \dots, v_m\}$ , 名词集合  $N = \{n_1, n_2, \dots, n_k\}$ , 定义选择模式为  $\langle V', N' \rangle$ , 其中,  $V' \subseteq V$ ;  $N' \subseteq N$ , 选择模型即为这些选择模式的集合, 这是一个概念模型。LSC 将其转为概率模型, 即用概率分布式替换特征函数集合。据此, 将选择模式  $C = \{c_1, c_2, \dots, c_l\}$  定义为一对分别基于动词和名词的离散分布值, 表示如下:

$$\begin{aligned} & \langle \lambda v P(v|c), \lambda n P(n|c) \rangle \\ & \sum_{v \in V} P(v|c) = 1 \\ & \sum_{n \in N} P(n|c) = 1 \quad c \in C \end{aligned} \quad (4)$$

其中, 函数  $\lambda v P(v|c)$  把动词  $v$  映射为区间(0,1)上的一个值, 且同一模式内所有动词的值之和满足等于 1 的限制, 名词类似。

假定  $P(c)$ 、 $P(v|c)$  与  $P(n|c)$  互相独立, 则基于某一个特定的模式  $c$ , 构造  $V \times N$  上的概率分布式如下:

$$P(v, n|c) = P(v|c) \times P(n|c) \quad v \in V, n \in N, c \in C \quad (5)$$

还可以构造  $C \times V \times N$  上的概率分布式为:

$$P(c, v, n) = P(c) \times P(v|c) \times P(n|c) \quad v \in V, n \in N, c \in C \quad (6)$$

式(6)在所有模式上求和, 即得:

$$\begin{aligned} P(v, n) &= \sum_{c \in C} P(c, v, n) = \\ & \sum_{c \in C} P(c) \times P(v|c) \times P(n|c) \\ & v \in V, n \in N, c \in C \end{aligned} \quad (7)$$

LSC 模型是软聚类的方法, 意味着某动、名词对不是绝对属于或不属于某模式, 而是在某种程度上属于某模式, 这个程度就是概率  $P(v, n|c)$ 。在实际的模型拟合时, 并不以  $v$ 、 $n$  的联合概率  $P(v, n|c)$  直接作为参数, 而是以边缘概率  $P(v|c)$ 、 $P(n|c)$  作为参数, 然后利用期望极大化(EM)算法估算参数值。

#### 3.2 基于 EM 的参数估算

LSC 模型的可视数据(incomplete)为动词  $v$  和名词  $n$  的共现频率  $freq(v, n)$ , 该值可以从语料库中获得; 相应的不可视数据(complete)为三元组  $(c, v, n)$ , 参数分别是  $P(c)$ 、 $P(v|c)$ 、 $P(n|c)$ 。于是 EM 算法交替于如下 2 个步骤, 直到收敛:

(1)E-步, 利用当前估计的参数值计算三元组  $(c, v, n)$  的数学期望  $E(c, v, n)$ 。

对于给定模型, 在模式  $c$  中生成可视数据  $(v, n)$  的概率表示为  $\frac{P(c, v, n)}{P(v, n)}$ , 于是事件  $\langle c, v, n \rangle$  发生的数学期望为:

$$E(c, v, n) = freq(v, n) \times \frac{P(c, v, n)}{P(v, n)} \quad (8)$$

其中,  $P(c, v, n)$  和  $P(v, n)$  根据式(6)、式(7)计算。

(2)M-步, 基于数学期望  $E(c, v, n)$ , 按照式(9)~式(11)更新参数值:

$$P(v|c) = \frac{E(c, v)}{E(c)} \quad (9)$$

$$P(n|c) = \frac{E(c, n)}{E(c)} \quad (10)$$

$$P(c) = \frac{E(c)}{\sum_{v \in V, n \in N} freq(v, n)} \quad (11)$$

其中,  $E(v, c)$ 、 $E(n, c)$  及  $E(c)$  的计算一目了然:

$$\begin{aligned} E(c, v) &= \sum_{n \in N} E(c, v, n) \\ E(n, c) &= \sum_{v \in V} E(c, v, n) \\ E(c) &= \sum_{v \in V, n \in N} freq(v, n) \end{aligned} \quad (12)$$

### 4 语义优选的获取

获取语义优选的过程分为 3 大步骤:

(1)根据式(3)计算动词  $v$  和名词  $n$  的  $\delta_{sc}$  值, 以此代表其语义关联程度。

(2)利用基于 LSC 模型的 EM 算法求取  $P(v, n)$  的值, 如下:

1)对 3 个模型参数  $P(v|c)$ 、 $P(n|c)$ 、 $P(c)$ , 随机赋予初始值, 满足条件  $\sum_{v \in V} P(v|c) = 1$ ,  $\sum_{n \in N} P(n|c) = 1$ ,  $\sum_{c \in C} P(c) = 1$  (模式的标准化处理不是必需的)。

2)根据式(8)计算数学期望  $E(c, v, n)$ 。

3)利用式(12)和式(9)~式(11)分别重新计算参数值  $P(v|c)$ 、 $P(n|c)$ 、 $P(c)$ 。

4)基于式(7)计算动词  $v$  和名词  $n$  的语义搭配概率  $P(v, n)$ 。

5)按照  $P(v, n)$  从大到小排序取  $\langle v, n \rangle$  组合。

6)循环执行步骤(2)~步骤(5), 直到收敛。

(3)将同一个  $\langle v, n \rangle$  组合的  $\delta_{sc}$  与  $P(v, n)$  的值相加, 结果按照从大到小的顺序排列, 取前  $k$  个作为动词  $v$  语义优选的名词集合。

### 5 实验结果与分析

因为汉语的语义优选目前没有“黄金标准”, 所以根据所选词汇手工构造语义搭配表。测试时以每个动词为单位进行, 最后取所有动词的平均值。

本文以 1998 年人民日报手工标注的语料库(分词并标注词性)为建模对象, 涉及到的词汇大约 18 049 个。实验取 2 个集合: 一个集合包括 34 个动词, 60 个名词(集合 1); 另一个集合包括 49 个动词, 97 个名词(集合 2)。

#### 5.1 基于 MDL 方法的测试

利用集合 2 对基于 MDL 的方法进行测试。针对每一个动词, 根据式(3)计算名词的  $\delta_{sc}$  值, 结果如表 1 所示。

表 1 12 个动词语义搭配的  $\delta_{sc}$  值

加大	建设	把握	保持	存在	改善
力度	工作	认识	经济	体制	生活
0.020 647	0.017 584	0.000 340	0.001 501	0.000 686	0.001 473
工作	国家	计划	思想	经济	条件
0.005 016	0.007 787	0.000 204	0.001 128	0.000 500	0.001 169
经济	经济	机遇	目标	机构	能力
0.003 548	0.007 500	0.000 190	0.000 745	0.000 433	0.000 644
-	思想	思想	局势	目标	工作
-	0.005 824	0.000 175	0.000 285	0.000 185	0.000 621
-	服务	剧本	产量	机遇	机制
-	0.005 406	0.000 155	0.000 267	0.000 096	0.000 595
提高	安排	保护	出现	形成	爱好
水平	工作	国家	信息	格局	体育
0.017 250	0.000 783	0.000 937	0.001 552	0.001 026	0.000 120
能力	会议	会议	机构	时机	生活
0.010 547	0.000 647	0.000 717	0.001 055	0.000 502	0.000 077
质量	铁路	计划	思想	局势	技术
0.010 348	0.000 515	0.000 467	0.000 887	0.000 285	0.000 069
工作	国家	价格	农村	版面	-
0.007 729	0.000 375	0.000 439	0.000 389	0.000 116	-
经济	公司	力度	格局	机会	-
0.006 341	0.000 323	0.000 373	0.000 262	0.000 105	-

表 1 是随机取 12 个动词的语义搭配表, 按照  $\delta_{sc}$  值从大到小的顺序取  $\delta_{sc} > 0$  的前 5 个名词。从表中可见, 每个动词对名词的语义选择基本合理。

## 5.2 基于 LSC 方法的测试

### 5.2.1 模式数量的确定

对于固定的动词概念群组 and 名词概念群组, 模式数量过多会导致很多无关联的语义搭配出现, 但数量过少又会丢失某些动、名词搭配组合。一般文献都是人为主观地固定模式数量, 实验结果的理想性很难从理论上验证<sup>[22-23]</sup>。本文通过计算  $P(v, n)$  直接获取动词对名词的选择, 可以不必过多考虑模式的数量。

如果以  $change_{prob}$  作为变化率, 则其计算公式为:

$$change_{prob} = \frac{change_{num}}{total_{num}} \times 100\% \quad (13)$$

其中,  $change_{num}$  表示本次迭代与下次迭代组合中有所变化的数量;  $total_{num}$  表示总的组合数。

以集合 1 为实验对象, 取总的组合数为 10、30、50 (按照  $P(v, n)$  从大到小的顺序排列的前 10 个、30 个、50 个), 以式(13)计算变化率, 结果如图 1 所示。实验说明模式数量超过某值后, 对于动名词的搭配结果影响不大。

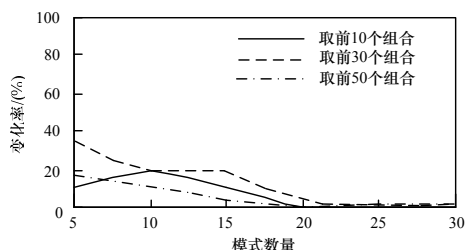


图 1 稳定语义搭配组合的实验结果

### 5.2.2 初始值与结果的关系

根据本文方法, 初始值是随机给出的, 因此实验中要讨论的下一个问题是初始值是否影响实验结果。以集合 1 为实验对象, 模式数量取 25 个, 分别从 5 次不同的初始值运行 EM 算法, 迭代 1 000 次, 取动词、名词的语义搭配 (按照  $P(v, n)$  从大到小前 10 个), 结果如表 2 所示。从表 2 中可见, 虽然  $P(v, n)$  的值有些差别, 但搭配结果基本不变。这说明优选结果独立于初始值。

表 2 不同初始值的  $P(v, n)$  值

初始值 1	初始值 2	初始值 3	初始值 4	初始值 5
发展经济	发展经济	发展经济	发展经济	发展经济
0.054 247	0.054 133	0.054 311	0.054 267	0.054 706
加大力度	加大力度	加大力度	加大力度	加大力度
0.035 812	0.035 927	0.036 188	0.035 815	0.035 832
提高水平	提高水平	提高水平	提高水平	提高水平
0.029 857	0.029 922	0.029 952	0.029 911	0.029 902
提高能力	提高能力	提高能力	提高能力	提高能力
0.018 933	0.018 761	0.018 779	0.018 760	0.018 746
提高质量	提高质量	提高质量	提高质量	提高质量
0.018 006	0.017 978	0.018 011	0.017 974	0.017 987
发展水平	发展水平	发展水平	发展水平	发展水平
0.015 055	0.015 060	0.014 974	0.015 036	0.015 173
提高工作	提高工作	提高工作	提高工作	提高工作
0.013 516	0.013 402	0.013 389	0.013 403	0.013 395
发展工作	发展工作	发展工作	发展工作	发展工作
0.013 170	0.013 316	0.013 223	0.013 186	0.013 280
发展技术	发展技术	发展技术	发展技术	发展技术
0.012 613	0.012 614	0.012 649	0.012 616	0.012 810
发展生活	发展生活	发展生活	发展生活	发展生活
0.011 846	0.011 836	0.011 766	0.011 835	0.012 628

### 5.2.3 收敛的确定

本文以动词和名词的语义组合是否改变作为收敛的依据。这样做的好处是: 迭代次数少, 节省时间; 更科学, 符

合语义优选的目的。对 2 个集合分别进行实验, 模式数量取 25 个, 按照  $P(v, n)$  从大到小的顺序取前 100 个, 根据式(13)计算变化率, 结果如图 2 所示。

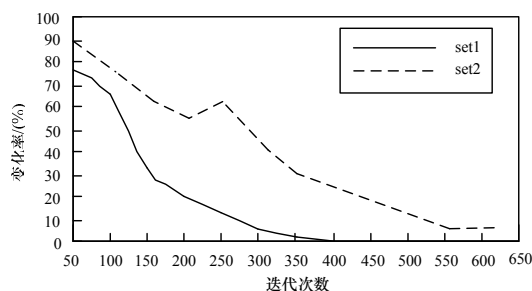


图 2 本文算法的收敛结果

从图 2 中可以看出, 尽管 2 个集合的动词、名词数目不同, 但迭代 600 次, 其组合基本固定。

## 5.3 度量标准

权衡准确率与召回率, 以  $F1$  作为度量标准检测实验结果, 令  $correct$  表示正确的语义搭配组合 (本文方法识别出的在“黄金标准”中的搭配) 的识别数量;  $algorithm$  表示本文方法识别出的语义搭配总的组合数量;  $gold$  表示“黄金标准”中手工给出的语义搭配数量, 则  $F1$  定义如下:

$$F1 = \frac{2 \times correct}{algorithm + gold} \times 100\% \quad (14)$$

## 5.4 结果与讨论

理论上讲,  $P(v, n)$  和  $\delta_{sc}$  都能代表动词、名词间的语义关联, 只是出发的统计角度不同, LSC 模型着重于动、名词的共现, 而 MDL 原则不仅考虑共现, 而且考虑非共现, 甚至不出现的情况, 就是说, MDL 原则对背景及上下文的考虑更全面。

从测试结果看, 2 种方法各有短长。表 3 和表 4 给出 6 个动词实例, 表 3 中的 6 个动词 MDL 方法的结果好些, 表 4 中的 3 个动词 LSC 方法的结果好些。可见, 2 种方法适合的动词不同, 实际上, 结果差别取决于动词、名词出现的上下文背景状况。

表 3 最小描述长度原则方法适合的动词

建设		爱好		把握	
MDL	LSC	MDL	LSC	MDL	LSC
经济	工作	体育	人才	认识	工作
0.006 064	0.017 473	0.000 120	0.000 204	0.000 340	0.001 536
工作	经济	生活	国家	计划	情况
0.003 822	0.008 167	0.000 077	0.000 058	0.000 204	0.000 564
农村	国家	技术	学生	机遇	思想
0.002 323	0.007 819	0.000 069	0.000 055	0.000 190	0.000 561
铁路	思想	-	经济	思想	国家
0.001 237	0.005 757	-	0.000 048	0.000 175	0.000 552
国家	服务	-	工作	剧本	服务
0.001 144	0.005 367	-	0.000 045	0.000 155	0.000 509

表 4 语义聚类模型方法适合的动词

安排		保护		出现	
LSC	MDL	LSC	MDL	LSC	MDL
生活	工作	工作	国家	情况	信息
0.002 024	0.000 783	0.002 138	0.000 937	0.003 818	0.001 552
工作	会议	经济	会议	经济	机构
0.001 332	0.000 647	0.002 000	0.000 717	0.003 174	0.001 055
计划	铁路	国家	计划	工作	思想
0.000 868	0.000 515	0.001 932	0.000 467	0.001 093	0.000 887
时间	国家	计划	价格	机会	国家
0.000 578	0.000 375	0.000 604	0.000 439	0.000 861	0.000 787
技术	公司	财产	力度	技术	时间
0.000 441	0.000 323	0.000 584	0.000 373	0.000 817	0.000 657

为了有更理想的结果, 考虑将 2 种方法结合起来, 以  $value(v, n) = P(v, n) + \delta_{sc}$  的值衡量动词对名词的选择。若利用

集合 2 进行测试, 3 种方法的  $F1$  结果如表 5 所示, 其中, LSC 方法取 25 个模式, 迭代 1 000 次。每种方法都仅考虑所有正数大于算术平均值的语义搭配组合, 即:

$$\begin{aligned} \text{LSC 方法: } P(v, n) &> \frac{\sum_{k=1}^{i-1} P_i(v, n)}{k} \quad P_i(v, n) > 0 \\ \text{MDL 方法: } \delta_{sc} &> \frac{\sum_{m=1}^{i-1} \delta_{sc_i}}{m} \quad \delta_{sc_i} > 0 \\ \text{LSC+MDL 方法: } value(v, n) &> \frac{\sum_{z=1}^{i-1} value_z(v, n)}{z} \quad value_i(v, n) > 0 \end{aligned} \quad (15)$$

若正值数目不足 5, 则取结果值为正的组。先针对每个动词计算  $F1$  值, 然后求所有动词  $F1$  值的平均值作为结果。

表 5 3 种方法的  $F1$  值 (%)

LSC 方法	MDL 方法	LSC 结合 MDL 方法
74.21	82.87	85.26

表 5 说明 MDL 方法比 LSC 方法的  $F1$  值高, 而结合 2 种方法的结果更好。很多错误是由于缺乏语法信息所致, 比如“建设-工作”、“面对-经济”、“帮助-生活”等, 如果针对标注语法信息的语料库进行学习, 结果会有大幅度提高。

## 6 结 束 语

本文利用 MDL 原则从生活语料库获取语义优选信息, 并结合 LSC 方法进一步提高实验结果。实验说明本文方法有较理想的  $F1$  值。今后的工作将围绕 3 个方面做进一步研究: (1)寻找更理想的结合策略; (2)考虑不采用固定的动、名词组合, 而利用语料库抽取相应词汇; (3)将本体库结合进来。

## 参 考 文 献

- [1] 邵敬敏. 汉语语法的立体研究[M]. 北京: 商务印书馆, 2007.
- [2] 徐 波, 孙茂松, 靳光瑾. 中文信息处理若干重要问题[M]. 北京: 科学出版社, 2003.
- [3] 俞士汶. 现代汉语语法信息词典详解[M]. 2 版. 北京: 清华大学出版社, 2003.
- [4] McCarthy D, Sussex F E, Joshi V S, et al. Detecting Compositionality of Verb-object Combinations Using Selectional Preferences[C]//Proc. of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Prague, Czech: [s. n.], 2007.
- [5] McCarthy D, Carroll J. Disambiguating Nouns, Verbs, and Adjectives Using Automatically Acquired Selectional Preferences[J]. Computational Linguistics, 2003, 29(4): 639-654.
- [6] Wagner W, Schmid H, Schulte S. Verb Sense Disambiguation Using a Predicate-argument-clustering Model[C]//Proc. of CogSci Workshop on Distributional Semantics Beyond Concrete Concepts. Amsterdam, Holland: [s. n.], 2009: 23-28.
- [7] Schulte S, Hying C, Scheible C, et al. Combining EM Training and the MDL Principle for an Automatic Verb Classification Incorporating Selectional Preferences[C]//Proc. of the 46th Annual Meeting of the Association for Computational Linguistics. Columbus, USA: [s. n.], 2008: 496-504.
- [8] Sun Lin, Korhonen A. Improving Verb Clustering with Automatically Acquired Selectional Preferences[C]//Proc. of Conference on Empirical Methods in Natural Language Processing. Beijing, China: [s. n.], 2009.
- [9] Zanzotto F M, Pennacchiotti M, Pazenza M T. Discovering Asymmetric Entailment Relations Between Verbs Using Selectional Preferences[C]//Proc. of ACL'06. Sydney, Australia: [s. n.], 2006: 849-856.
- [10] Mason Z J. Cornet: A Computational, Corpus-based Conventional Metaphor Extraction System[J]. Computational Linguistics, 2004, 30(1): 23-44.
- [11] Zapirain B, Agirre E, Marquez L, et al. Improving Semantic Role Classification with Selectional Preferences[C]//Proc. of Annual Conference of the North American Chapter of the Association for Computational Linguistics. Los Angeles, USA: [s. n.], 2010.
- [12] Young A C. The Effect of Selectional Preferences on Semantic Role Labeling[D]. [S. l.]: The University of Texas at Austin, 2009.
- [13] Katrin E, Padó S, Padó U. A Flexible, Corpus-driven Model of Regular and Inverse Selectional Preferences[EB/OL]. (2010-10-14). [http://www.mitpressjournals.org/doi/abs/10.1162/coli\\_a\\_00017](http://www.mitpressjournals.org/doi/abs/10.1162/coli_a_00017).
- [14] Bergsma S, Lin Dekang, Goebel R. Discriminative Learning of Selectional Preference from Unlabeled Text[C]//Proc. of Conference on Empirical Methods in Natural Language Processing. Morristown, USA: [s. n.], 2008: 59-68.
- [15] Mausam R A, Etzioni O. A Latent Dirichlet Allocation Method for Selectional Preferences[C]//Proc. of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden: [s. n.], 2010.
- [16] Erk K. A Simple, Similarity-based Model for Selectional Preferences[C]//Proc. of the 45th Annual Meeting of the Association for Computational Linguistics Association for Computer Linguistics. Michigan, USA: [s. n.], 2007.
- [17] Schulte S. Comparing Computational Models of Selectional Preferences-second-order Co-occurrence vs. Latent Semantic Clusters[C]//Proc. of International Conference on Language Resources and Evaluation. Valletta, Malta: [s. n.], 2010.
- [18] Zheng Xuling, Zhou Changle, Li Tangqiu, et al. Automatic Acquisition of Chinese Semantic Collocation Rules Based on Association Rule Mining Technique[J]. Journal of Xiamen University: Natural Science, 2007, 46(3): 331-336.
- [19] Wu Yunfang, Duan Huiming, Yu Shiwen. Verb's Selectional Preference on Object[J]. Spoken and Written Language in Practice, 2005, 21(2): 121-128.
- [20] Jia Yuxiang, Yu Shiwen. Automatic Acquisition of Selectional Preference and Its Application to Metaphor Processing[C]//Proc. of the 4th National Student Conference on Computational Linguistics. Taiyuan, China: [s. n.], 2008.
- [21] Li Hang, Yamanishi K. Topic Analysis Using a Finite Mixture Model[J]. Information Processing & Management, 2003, 39(4): 521-541.
- [22] Mats R. Two-dimensional Clusters in Grammatical Relations[J]. Inducing Lexicons with the EM Algorithm, 1998, 4(3): 7-24.
- [23] Wagner A. Learning Thematic Role Relations for Lexical Semantic Nets[D]. [S. l.]: Tubingen University, 2004.

编辑 任吉慧