



# 目 录

---

<b>绪论</b>	<b>1</b>
0.1 研究背景	1
0.2 本研究的理论及实践意义	4
0.2.1 理论意义	4
0.2.2 实践意义	6
0.3 本研究概述	9
0.3.1 研究目的	9
0.3.2 研究问题及研究对象	9
0.3.3 研究步骤	10
0.3.3.1 语料及预处理	11
0.3.3.2 特征集的建立及其转换	11
0.3.3.3 权重计算	13
0.3.3.4 索引行聚类	14
0.3.3.5 型式提取	15
0.4 本书结构	15
0.5 小结	16
<b>第一章 型式与型式语法</b>	<b>17</b>
1.1 型式	17
1.1.1 型式研究的缘起与发展	17
1.1.2 型式的定义	20
1.1.3 型式实例	21
1.1.4 型式元素及其编码	23
1.1.5 本研究中判别型式的六个标准	26
1.2 型式语法	26



x 基于索引行聚类的英语动词型式自动识别与提取研究

1.2.1 短语学思想与习语原则	27
1.2.2 词汇与语法关系及各自地位	28
1.2.2.1 词汇语法不可分	28
1.2.2.2 词汇为中心的研究	30
1.2.3 型式与意义的关系	32
1.3 型式语法的优缺点	35
1.3.1 型式语法与传统语法的差异	35
1.3.2 型式语法的优点	36
1.3.3 型式语法的不足	37
1.4 以型式语法为理论基础的相关研究	40
1.5 小结	41

---

**第二章 型式识别方法与相关应用研究 42**

2.1 型式识别标准	42
2.2 型式总结的必要性	43
2.3 型式的识别	44
2.3.1 型式的人工识别	45
2.3.2 型式人工识别辅助工具的开发	48
2.3.3 型式的自动识别	48
2.3.3.1 型式自动识别的理据	48
2.3.3.2 型式的自动识别研究	51
2.4 现有的语言型式网络平台数据库	52
2.4.1 基于机器处理的网络数据库	53
2.4.2 基于人工处理的网络数据库	54
2.5 小结	56

---

**第三章 聚类分析 57**

3.1 文本表示	58
3.2 特征选择及其权重	59
3.2.1 特征选择	59
3.2.2 权重计算	60
3.3 相似度计算	61
3.3.1 相似度计算的源起及理据	61
3.3.2 相似度计算方法	61
3.3.3 相似度分析在语言研究中的应用	63



3.4 聚类算法	64
3.4.1 划分聚类	65
3.4.2 层次聚类	66
3.5 聚类质量评价指标	67
3.6 聚类在本研究的应用理据	68
3.7 小结	69

---

## **第四章 文本预处理与特征集的建立及转换** **70**

4.1 研究概述	70
4.2 语料选取	70
4.3 研究工具	72
4.3.1 语料预处理工具	72
4.3.2 自主开发的模块及脚本	72
4.4 语料预处理流程	73
4.5 动词型式中的必要元素及其转换方法	75
4.5.1 型式列表中元素的总体特征	75
4.5.2 具体词形的处理方法	77
4.5.2.1 右侧搭配词处理方法	77
4.5.2.2 左侧搭配词处理方法	90
4.5.2.3 两侧搭配词处理方法	90
4.5.3 相邻单词组合的处理方法	90
4.5.4 词类标签及语义类标签的转换方法	98
4.5.4.1 词类标签的转换方法	98
4.5.4.2 语义类标签转换方法	102
4.5.5 转换顺序及步骤	104
4.6 小结	105

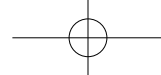
---

## **第五章 索引行聚类及型式自动提取方法** **106**

5.1 特征表示方法、特征权重与位置权重的计算	106
5.1.1 特征及特征权重	106
5.1.1.1 功能词处理方法	107
5.1.1.2 特征标记方法	107
5.1.1.3 型式边界的界定	108
5.1.1.4 索引行跨距的设定	110
5.1.1.5 特征权重计算方法	112



5.1.2 位置权重	112
5.1.3 特征—索引行矩阵的生成	113
5.2 聚类算法	114
5.2.1 相似度计算	114
5.2.2 K均值算法	115
5.2.2.1 K均值聚类	115
5.2.2.2 K值的选择标准	115
5.3 型式自动提取	116
5.4 小结	117
<b>第六章 型式自动提取模型测试</b>	<b>118</b>
6.1 模型调试数据集及模型验证集的构建	118
6.1.1 调试阶段语料集合的构成	119
6.1.2 验证集的创建过程	121
6.1.2.1 验证集中词项的选择	121
6.1.2.2 验证集中索引行的抽取及处理方法	122
6.2 配置文件的处理及参数设置与调整	123
6.2.1 配置文件处理顺序	123
6.2.2 参数设置与调整	126
6.2.2.1 特征权重计算方法调试过程与解决方法	126
6.2.2.2 位置权重计算方法调试过程及解决方法	126
6.2.2.3 每个特征的总体权重计算方法	127
6.2.2.4 跨距设定调试过程	128
6.3 测试数据的评价指标	128
6.3.1 聚类内部效度评估指标在本研究中的应用	129
6.3.2 聚类外部效度评估指标在本研究中的应用	134
6.4 数据结果报告	135
6.4.1 索引行中型式及型式元素分布特征	135
6.4.1.1 型式元素总体分布特征	135
6.4.1.2 与动词高频共现的特征及其频数	136
6.4.1.3 不同动词型式列表中特征的异同	140
6.4.2 K值确定下基于现有人工标签的型式自动识别效度分析	143
6.4.2.1 验证集中型式自动识别外部效度评估结果总体描述	143

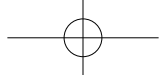


6.4.2.2	实验动词的型式自动识别准确率及召回率	144
6.4.2.3	K值确定下型式自动识别后的再思考	153
6.4.3	K值不确定下基于现有的人工标签的型式自动识别 效度分析	153
6.4.3.1	K值不确定下型式自动识别的外部效度测量 结果总体描述	154
6.4.3.2	K值不确定下型式自动识别的准确率及召回率	155
6.4.3.3	K值未知情况下模型验证的再思考	170
6.5	分析和讨论	171
6.5.1	数据背后的语言学思考	171
6.5.1.1	印证了分布假设	171
6.5.1.2	印证了词汇语法不可分	172
6.5.1.3	型式元素间的横组合关系	173
6.5.1.4	常规型式与非常规型式	173
6.5.2	影响部分型式自动识别错误的原因	175
6.5.2.1	型式元素间的插入成分	175
6.5.2.2	无引导词THAT标志的从句识别	176
6.5.2.3	赋码错误	177
6.5.2.4	名词短语的识别问题	177
6.5.3	关于聚类外部效度测量结果的再思考	178
6.5.4	与前人研究结果的对比	180
6.6	小结	181

---

## **第七章 英语动词型式自动提取模型的应用设想 183**

7.1	应用范围	183
7.1.1	型式自动识别与提取在语言教学领域的应用	184
7.1.1.1	型式自动识别与提取在教学大纲制定中的应用	184
7.1.1.2	型式自动识别与提取在语言课堂教学中的应用	185
7.1.1.3	型式的自动识别与提取可以服务于学习者 自主学习	185
7.1.2	型式自动识别与提取在语言研究中的应用	186
7.1.2.1	型式自动提取在词典编纂中的应用	186
7.1.2.2	型式自动提取模型对非常规用法的识别	186
7.1.2.3	型式自动识别在语言学研究中的应用	187
7.1.3	型式自动识别与提取在其他领域的应用可能	190



7.2 应用时需要注意的问题	191
7.2.1 语料及语料的赋码	191
7.2.2 聚类时参数的设置	192
7.3 小结	194

---

## **第八章 研究发现及价值 195**

8.1 研究发现	195
8.1.1 索引行聚类的影响因素及分组数量的选择	196
8.1.1.1 索引行聚类的影响因素	196
8.1.1.2 索引行聚类时簇数量的选择	196
8.1.2 型式自动识别的准确率及其影响因素	197
8.1.2.1 型式自动识别准确率的提高	197
8.1.2.2 影响部分型式自动识别错误的因素	198
8.1.3 不同动词索引行中型式元素的分布存在差异	198
8.2 研究价值及创新之处	200
8.2.1 理论方面	201
8.2.1.1 对型式语法的反哺	201
8.2.1.2 对其他语言学理论的贡献	205
8.2.2 方法论方面	206
8.2.3 实践方面	207
8.3 本研究的不足之处	208
8.4 未来研究方向	209
8.5 小结	210

---

## **参考文献 211**

---

## **附录 226**