



上海外国语大学博士后流动站
战略支援部队信息工程大学

原伟
著

基于知识本体的俄汉 可比语料库建设与应用研究

JINYU ZHISHI BEN TI DE E-HAN KEBI YULIAOKU
JIAN SHE YU YINGYONG YANJIU





上海外国语大学博士后流动站
战略支援部队信息工程大学

原伟
著

基于知识本体的俄汉 可比语料库建设与应用研究



世界图书出版公司

广州·上海·西安·北京

图书在版编目 (CIP) 数据

基于知识本体的俄汉可比语料库建设与应用研究 /
原伟著. —广州: 世界图书出版广东有限公司, 2019.4
ISBN 978-7-5192-6145-0

I. ①基… II. ①原… III. ①语料库—建设—研究
IV. ①H0

中国版本图书馆 CIP 数据核字 (2019) 第 067857 号

-
- 书 名 基于知识本体的俄汉可比语料库建设与应用研究
JIYU ZHISHI BENTI DE E-HAN KEBI YULIAOKU JIANSHE YU YINGYONG YANJIU
- 著 者 原 伟
- 策划编辑 刘正武
- 责任编辑 张东文
- 出版发行 世界图书出版广东有限公司
- 地 址 广州市海珠区新港西路大江冲 25 号
- 邮 编 510300
- 发行电话 020-84451969 84459539
- 网 址 <http://www.gdst.com.cn/>
- 邮 箱 wpc_gdst@163.com
- 经 销 新华书店
- 印 刷 虎彩印艺股份有限公司
- 开 本 787 mm × 1092 mm 1/16
- 印 张 16
- 字 数 265 千字
- 版 次 2019 年 4 月第 1 版 2019 年 4 月第 1 次印刷
- 国际书号 ISBN 978-7-5192-6145-0
- 定 价 48.00 元
-

版权所有 侵权必究

咨询、投稿: 020-84460251 gzzlw@126.com
(如有印装错误, 请与出版社联系)

序 言

可比语料库作为近年来语料库研究的热点方向之一，可广泛应用于语言学研究 and 自然语言处理领域。本体是一种基于语义网技术的知识表示方法，它与可比语料库的融合式研究是对可比语料获取、组织和应用方法的一次革新，将更好地发挥可比语料库的效能并扩展其应用领域。经过前期调查分析，国内外鲜有俄汉可比语料库相关研究，未见基于本体的俄汉可比语料库相关成果。本研究首先在理论层面分析了可比语料库研究的现存问题，提出将本体引入可比语料库研究的思路，并以此为基础提出了基于本体的可比语料库理论构想。随后在实践层面将该理论构想运用到了面向俄汉可比语料库的乌克兰事件多语言复合型本体构建、基于该本体的俄汉新闻及维基百科可比语料获取、语料库构建和语料库应用等核心问题的研究中。

本研究所取得的成果：第一，本研究所提出的基于本体的可比语料库理论构想由“一对关系、三个模型和四个问题”组成，阐释了本体与语料的层次关系，建立了基于本体的单语料、语料库和语料可比关系模型，在理论上讨论了面向可比语料库的本体构建、基于本体的可比语料获取、语料库构建和语料库应用问题。第二，本研究所构建的面向俄汉可比语料库的乌克兰事件多语言复合型本体（МОПКУС）

由表征领域知识和描述存储语料的 2 大类及 8 个子类组成, 包含数据属性 46 个, 对象属性 9 个, 实例 60733 个 (领域知识类实例 4525 个, 语料描述类实例 56207 个)。第三, 本研究所构建的基于 МОПКУС 的俄汉可比语料库包含新闻原始语料 3554 篇, 维基百科原始语料 1670 篇, 其中包括中文语料 163 万字, 俄文语料 132 万词; 以领域相同、发布时间相似对齐俄汉新闻可比语料文本 50148 对, 以俄汉维基语言链接对齐维基百科可比语料文本 835 对, 并对俄汉语料各层面的可比程度进行了评估计算, 为语料库的应用奠定了坚实的数据基础。第四, 在所构建的基于本体的俄汉可比语料库基础上探讨了三个应用问题: 复杂语料调用与语义查询、基于多维度特征的可比度评估以及跨语言文本推荐与信息整合, 并研制了相关软件系统。

本研究的创新之处在于: 第一, 通过对现有可比语料库研究理论的阐述与分析, 对可比语料的定义和分类问题进行了重新整理与论证, 分析了可比语料库研究现存问题, 提出了基于本体的可比语料库研究思路。第二, 提出并建立了基于本体的可比语料库理论构想, 并将该理论在基于本体的俄汉可比语料库构建和应用中进行了验证。第三, 尝试提出一种基于多维特征的语料可比度评估方法, 并通过软件系统进行了验证。第四, 尝试提出一种使用非翻译手段获取跨语言文本信息的方法——基于可比语料的跨语言文本推荐和信息整合, 并通过软件系统进行了验证。

在本书选题、撰写和修订过程中, 我的老师王松亭教授给予了我悉心的指导, 启迪了我的学术思想, 倾注了大量的精力和心血, 激励我在学术道路上不断探索, 对他的帮助我感恩不敢忘怀, 希望在今后成长的道路上不辜负老师的谆谆教诲和殷切希望。

感谢我的老师易绵竹教授多年以来对我的悉心指导和无私帮助, 他广博的知识和深刻的学术洞察力对我的学习和科研产生了巨大影

响，在困惑时总能为我提出宝贵的意见与建议，此刻心中更多的是感谢与感恩。

感谢我的家人多年以来给予我的宽容、理解、支持与帮助，任何时候他们都是我最坚实的后盾和力量源泉，希望我的点滴成绩能为他们带来一些欣慰与快乐，祝愿他们永远健康平安。

作 者

2019年2月于洛阳

目 录

绪 论	1
0.1 研究背景	2
0.2 研究意义	4
0.3 研究现状	5
0.4 研究内容	8
0.5 研究方法	9
0.6 创新点和难点	10
0.7 研究架构	11
第一章 可比语料库研究综论	13
1.0 本章引言	14
1.1 可比与类比的术语界定	15
1.2 定义重构与分类研究	18
1.2.1 定义的分析与重构	18
1.2.2 分类的标准与方法	22
1.3 可比语料库构建方法	24
1.3.1 基于现存语料库的构建方法	24

1.3.2	基于网络资源的构建方法	25
1.3.2.1	基于新闻网站的可比语料库构建	25
1.3.2.2	基于维基百科的可比语料库构建	27
1.3.2.3	基于网络的领域可比语料库构建	28
1.3.3	基于混合数据的构建方法	28
1.4	语料的可比度及其计算	30
1.4.1	单语种语料的可比度计算	30
1.4.2	多语种语料的可比度计算	31
1.5	俄语可比语料库研究现状	32
1.6	本章小结	34

第二章 基于本体的可比语料库理论构想

37

2.0	本章引言	38
2.1	本体和语料库融合式研究的理论前提	39
2.1.1	本体的定义	39
2.1.2	本体的构建方法	40
2.1.3	多语种本体的构建	44
2.1.3.1	衍生拓展法	45
2.1.3.2	中介语映射法	46
2.1.3.3	关系注释法	48
2.1.4	基于语料的本体研究	48
2.1.4.1	基于语料的概念抽取	50
2.1.4.2	基于语料的概念关系抽取	51
2.1.5	基于本体的语料库研究	52
2.2	基于本体的可比语料库理论体系	54
2.2.1	语言信息层面语料与本体的理论关系	54

2.2.2	面向单语料知识描述的理论模型.....	58
2.2.3	基于本体的可比语料库结构模型.....	60
2.2.4	面向可比语料库的本体构建问题.....	63
2.2.4.1	构建目标.....	63
2.2.4.2	构建方法.....	65
2.2.4.3	其他重要问题.....	66
2.2.5	基于本体的可比语料获取问题.....	67
2.2.5.1	基于本体获取领域种子词.....	68
2.2.5.2	基于本体获取可比语料.....	70
2.2.6	基于本体的可比语料库构建问题.....	73
2.2.7	基于本体的可比语料库应用问题.....	77
2.3	本章小结.....	78
第三章 面向俄汉可比语料库的乌克兰事件复合型本体 (МОПКУС)		
		79
3.0	本章引言.....	80
3.1	МОПКУС 概览与总体设计.....	81
3.1.1	МОПКУС 的定义与结构.....	81
3.1.2	МОПКУС 的构建目标.....	84
3.1.3	МОПКУС 的构建方法.....	85
3.1.4	МОПКУС 其他重要问题.....	88
3.2	МОПКУС 领域知识类.....	90
3.2.1	领域知识类需求分析与结构设计.....	90
3.2.2	基于乌克兰事件领域语料的知识获取.....	92
3.2.3	МОПКУС 人物子类.....	95
3.2.4	МОПКУС 地点子类 (place).....	99

- 3.2.5 МОПКУС 组织子类 (organization) 102
- 3.2.6 МОПКУС 时间子类 (time) 106
- 3.2.7 МОПКУС 客体子类 (object) 108
- 3.2.8 МОПКУС 行为子类 (action)111
- 3.3 МОПКУС 语料描述类 113
 - 3.3.1 语料描述类需求分析与结构设计 114
 - 3.3.2 МОПКУС 单语料描述方法 115
 - 3.3.3 МОПКУС 可比语料描述方法..... 116
 - 3.3.4 МОПКУС 原始语料子类 (original_corpora) 118
 - 3.3.5 МОПКУС 可比语料子类 (comparable_corpora) 121
- 3.4 МОПКУС 中俄实例关联..... 123
- 3.5 本章小结 128

第四章 基于 МОПКУС 的俄汉可比语料获取与语料库构建 129

- 4.0 本章引言 130
- 4.1 基于 МОПКУС 的俄汉可比语料获取..... 131
 - 4.1.1 基于 МОПКУС 的俄汉种子词获取 132
 - 4.1.1.1 基于 Jena 的 本体解析方法 132
 - 4.1.1.2 基于本体获取种子词的 程序实现..... 134
 - 4.1.1.3 基于 МОПКУС 的 种子词获取结果..... 135
 - 4.1.2 基于 МОПКУС 的 俄汉维基语料获取 139
 - 4.1.2.1 核心问题 分析与讨论..... 139
 - 4.1.2.2 俄汉维基可比语料获取的方法 143
 - 4.1.2.3 俄汉维基可比语料获取的程序实现 146
 - 4.1.2.4 维基百科语料获取结果..... 148
 - 4.1.3 基于 МОПКУС 的 俄汉新闻语料获取 149

4.1.3.1	面向新闻语料获取的种子词选取	149
4.1.3.2	新闻语料获取工具与过程	151
4.1.3.3	俄汉新闻语料获取结果	153
4.2	基于 МОПКУС 的俄汉可比语料库构建	155
4.2.1	МОПКУС 俄汉原始语料实例的处理与导入	156
4.2.1.1	中文原始语料的处理方法	157
4.2.1.2	俄文原始语料的处理方法	159
4.2.1.3	语料处理与导入程序实现	160
4.2.1.4	中俄文语料处理与导入结果	162
4.2.2	МОПКУС 俄汉可比语料实例的处理与导入	163
4.2.2.1	核心问题的分析讨论	163
4.2.2.2	可比语料实例的关系建立	166
4.2.2.3	可比语料实例的数据属性	170
4.2.2.4	俄汉跨语言相似度计算	174
4.2.2.5	可比语料实例处理导入结果	175
4.3	本章小结	177
第五章	基于 МОПКУС 的俄汉可比语料库应用	179
5.0	本章引言	180
5.1	应用一: 语料复杂查询与语义检索系统	181
5.1.1	基于词汇的检索	181
5.1.2	基于句子的检索	185
5.1.3	跨语言语料检索	187
5.1.4	语料可比性推理	191
5.1.5	程序实现	196
5.2	应用二: 基于多维特征的语料可比度评估系统	198

5.2.1	思路来源	198
5.2.2	概念定义	198
5.2.3	实施方法	199
5.2.4	具体算法	200
5.2.5	多维可比度示例	201
5.2.6	多维可视化分析	203
5.2.7	程序实现	204
5.2.8	应用优势	206
5.3	应用三：跨语言文本推荐与信息整合系统	206
5.3.1	思路来源	206
5.3.2	概念定义	207
5.3.3	实施方法	208
5.3.4	程序实现	212
5.3.5	应用优势	213
5.4	本章小结	213
结 论		215
参考文献		221

图表目录

图 0.1 论文研究内容示意图	9
图 1.1 第一章内容示意图	14
图 2.1 第二章内容示意图	38
图 2.2 语言信息层面语料与本体关系	57
图 2.3 单文本语料知识的信息层次模型示例	59
图 2.4 基于本体的可比语料库模型一：数据独立型	61
图 2.5 基于本体的可比语料库模型二：本体独立型	62
图 2.6 基于本体的可比语料库模型三：本体复合型	63
图 2.7 简单生物本体（中文命名示例）	69
图 2.8 基于本体从传统文书中获取双语可比语料	71
图 2.9 基于本体获取维基百科中获取双语可比语料	72
图 2.10 基于本体方法从新闻中获取双语可比语料	73
图 2.11 传统语料表示与本体语料单文本表示区别示例	74
图 2.12 传统语料表示与本体语料多文本组织区别示例	74
图 2.13 基于本体的语料导入处理	75
图 2.14 基于本体的可比语料对齐方法示例	76
图 3.1 第三章研究内容示意图	80

图 3.2	МОПКУС 的结构示意图.....	81
图 3.3	МОПКУС 本体框架图	82
图 3.4	МОПКУС 中的语义关系图示	83
图 3.5	МОПКУС 结构图	84
图 3.6	МОПКУС 的构建步骤	86
图 3.7	МОПКУС 领域知识类结构图	91
图 3.8	语料信息中的概念类及术语获取流程	94
图 3.9	人物子类与语料描述类对象属性关系	97
图 3.10	人物子类部分实例（为程序自动排序，中俄实例已建立内部 关联）	97
图 3.11	人物子类中文实例属性示例	98
图 3.12	人物子类俄文实例及属性示例	99
图 3.13	地点子类特有对象属性关系	101
图 3.14	地点子类部分中俄文实例（为程序自动排序，中俄实例已建 立内部关联）	101
图 3.15	地点子类中文实例示例	102
图 3.16	地点子类俄文实例示例	102
图 3.17	组织子类中俄文部分实例（为程序自动排序，中俄实例已建 立内部关联）	103
图 3.18	组织子类特有对象属性关系	105
图 3.19	组织子类俄文实例的数据及对象属性示例	105
图 3.20	组织子类中文实例的数据及对象属性示例	106
图 3.21	时间子类中俄文部分实例（为程序自动排序，中俄实例已建 立内部关联）	107
图 3.22	时间子类中文实例示例	108
图 3.23	时间子类俄文实例示例	108

图 3.24 客体子类中俄文部分实例（为程序自动排序，中俄实例已建立内部关联）	109
图 3.25 客体子类中文实例示例	110
图 3.26 客体子类俄文实例示例	111
图 3.27 行为子类中俄实例部分示例（为程序自动排序，中俄实例已建立内部关联）	112
图 3.28 行为子类中文实例示例	113
图 3.29 行为子类俄文实例示例	113
图 3.30 语料描述类结构设计图	115
图 3.31 МОПКУС 单语料描述方法	116
图 3.32 МОПКУС 可比语料描述方法	117
图 3.33 原始语料子类与领域知识类子类间关系	118
图 3.34 原始语料子类实例示例	119
图 3.35 原始语料子类中俄文部分实例（为程序自动排序，中俄实例已建立内部关联）	120
图 3.36 可比语料子类实例示例	123
图 3.37 双语本体命名规则示例	124
图 3.38 构建双语关联的本体对象属性示例	125
图 3.39 构建维基语料（corpora_of_wiki）双语关联的本体对象属性示例	127
图 3.40 构建新闻语料（corpora_of_news）双语关联的本体对象属性示例	127
图 4.1 第四章研究内容示意图	130
图 4.2 基于 МОПКУС 获取乌克兰事件领域俄汉新闻可比语料的基本思路	131
图 4.3 基于本体的种子词获取应用程序	135

图 4.4	2016 年 1 月 12 日维基百科中俄文词条数量统计	140
图 4.5	维基百科多语言链接示例	141
图 4.6	基于本体获取俄汉可比语料技术流程	142
图 4.7	维基百科中俄网页命名规则及中文种子词网址示例	143
图 4.8	维基百科中文词条的俄文链接网页源码示例	144
图 4.9	维基百科俄文词条的中文链接网页源码示例	144
图 4.10	俄汉维基百科单词条可比语料获取的程序核心代码片段	147
图 4.11	俄汉维基单词条可比语料获取的程序界面	147
图 4.12	俄汉维基百科批量词条可比语料获取的程序核心代码片段	148
图 4.13	俄汉维基百科批量词条可比语料获取的程序界面	148
图 4.14	新华网俄文搜索示例	151
图 4.15	新闻采集软件界面及选项设置	152
图 4.16	新闻采集软件网页下载过程	153
图 4.17	抽取后俄文新闻 XML 文本样式示例	153
图 4.18	俄文新闻语料数据文件示例	154
图 4.19	中文新闻语料数据文件示例	155
图 4.20	基于 МОПКУС 的俄汉可比语料构建设计	156
图 4.21	中文语料导入与处理程序示例	161
图 4.22	批量语料导入与处理程序	161
图 4.23	可比语料子类实例处理示例	165
图 4.24	维基可比语料实例与数据属性的建立	167
图 4.25	中俄新闻语料每日数量比较 (2013/11/20—2014/11/20)	168
图 4.26	中俄新闻语料可比对数据分布统计	169
图 4.27	俄汉新闻可比语料实例与数据属性的建立	170
图 4.28	基于词重叠的相似度算法	171

图 4.29	词重叠算法核心代码示例 (C#)	171
图 4.30	向量空间模型算法及核心代码示例 (C#)	173
图 4.31	基于机器翻译引擎的俄汉文本相似度计算流程	175
图 4.32	新闻语料各层面信息可比程度统计	176
图 4.33	维基语料各层面可比程度统计	176
图 5.1	第五章研究内容示意图	180
图 5.2	基于 МОПКУС 以关键词检索语料途径一	183
图 5.3	基于 МОПКУС 以关键词检索语料途径二	184
图 5.4	基于 МОПКУС 以关键词检索语料	184
图 5.5	基于 МОПКУС 使用句子检索语料	187
图 5.6	基于翻译的跨语言信息检索方法	188
图 5.7	基于本体的俄汉语料信息检索方法	189
图 5.8	基于 МОПКУС 的跨语言语料检索解决方案 (以俄文关键词 为例)	191
图 5.9	在 МОПКУС 中可比语料实例关系图	192
图 5.10	基于 МОПКУС 的维基可比语料检索路径 (以语料“欧盟” 为例)	193
图 5.11	基于 МОПКУС 的俄汉新闻可比语料检索路径	194
图 5.12	МОПКУС 语料的可比关系推理方法	196
图 5.13	基于 МОПКУС 的中文语料复杂检索程序展示	197
图 5.14	基于 МОПКУС 的俄文语料复杂检索程序展示	197
图 5.15	基于多维特征的语料可比度评估方法 (CEMDF) 的三个基 本指标	199
图 5.16	CEMDF 的基本方法	200
图 5.17	基于多维特征的语料可比度评估系统程序 (一)	205
图 5.18	基于多维特征的语料可比度评估系统程序 (二)	205