王东波 著

面向非结构化文本的知识发现

基于英汉双语平行语料库的句法级知识挖掘和抽取研究

中国社会科学出版社

王东波 著

面向非结构化文本的知识发现

基于英汉双语平行语料库的句法级知识挖掘和抽取研究

中国社会科学出版社

图书在版编目(CIP)数据

面向非结构化文本的知识发现：基于英汉双语平行语料库的句法级知识挖掘和抽取研究/王东波著．—北京：中国社会科学出版社，2013.5

ISBN 978-7-5161-2603-5

Ⅰ.①面… Ⅱ.①王… Ⅲ.①知识工程—数据收集—研究 Ⅳ.①TP182

中国版本图书馆 CIP 数据核字(2013)第 100209 号

出 版 人 赵剑英
责任编辑 王 琪
责任校对 鲍凤英
责任印制 王 超

出 版 中国社会科学出版社
社 址 北京鼓楼西大街甲 158 号（邮编 100720）
网 址 http://www.csspw.cn
中文域名：中国社科网 010-64070619
发 行 部 010-84083685
门 市 部 010-84029450
经 销 新华书店及其他书店

印刷装订 三河市君旺印装厂
版 次 2013 年 5 月第 1 版
印 次 2013 年 5 月第 1 次印刷

开 本 710×1000 1/16
印 张 14.75
插 页 2
字 数 227 千字
定 价 39.00 元

凡购买中国社会科学出版社图书，如有质量问题请与本社联系调换
电话：010-64009791

序

“除了上帝，任何人都必须用数据来说话”这句美国谚语将会在今天这个真正的大数据时代使世人更深刻、更真实、更全面地感受数据的力量和“无微不至”。大数据的迅猛发展和其所具有的战略意义，引起了世界的广泛关注。2012 年 1 月，达沃斯世界经济论坛发布《大数据，大影响》报告，宣称数据已经成为一种新的经济资产类别，就像货币和黄金一样，是 21 世纪的石油。2012 年 3 月奥巴马政府发布了《大数据研究和发展倡议》书，投资 2 亿美元启动“大数据研究和发展计划”；2012 年 11 月，国家发展和改革委员会在其发布的高技术服务业研发及产业化专项中，设立了“大数据分析软件开发和服务创新”专项；2013 年 2 月，科技部发布 2014 年度 973 计划指南，设立了“大数据计算的基础研究”项目。在以更大规模的数据、更多样化的数据、更实时的数据、数据的价值稀疏性或者不精确性为特征的大数据中，80% 的数据是非结构化的。如何从非结构化数据，尤其是非结构化文本数据中挖掘出有价值的知识就成为备受关注和研究的热点。王东波博士从英汉双语平行语料这一非结构化文本入手，通过把情报学的知识、自然语言处理和文本挖掘的技术结合起来的方法，围绕词汇、简单短语结构和复杂短语结构这三个句法层级进行的知识抽取和挖掘正是在这一研究趋势下的积极而有益的尝试。

针对已有面向非结构化文本知识挖掘研究过于依赖模型和统计的现状，王东波博士以其本科阶段所习得的语言学知识为切入点，从英、汉这两种印欧语系和汉藏语系为代表的语言特征入手，面向网络构建了大规模的英汉双语平行语料库，并基于英汉平行语料这一非结构化文本，

挖掘出了词汇句法动态组合的分布规律，以介宾短语结构为例探讨了短语结构的知识抽取，在由复杂短语结构组成的短句基础上构建了类别知识挖掘的模型。本书的主要研究内容如下：

面向网络获取了英汉双语通用和专门平行语料，构建了相应的英汉双语平行语料库。在该部分主要围绕确定抓取网站、制定抓取底表、通过抓取工具获取网页、抽取英汉双语平行语料对、清洗英汉双语平行语料对和对英汉双语平行语料对进行去重处理等问题展开了探讨。

在词汇这一级，结合情报学中的相应方法和知识，挖掘出了英汉双语词汇在句法功能分布复杂度上呈现洛特卡现象的规律。在该部分，基于英汉双语平行语料库、宾州大学英汉树库和清华汉语树库，统计了英语和汉语词汇句法功能的分布，分析了英语和汉语词汇的句法功能分布复杂度情况，计算得出了英语和汉语词汇的平均句法功能分布复杂度值，揭示了词汇句法功能分布复杂度所呈现的洛特卡现象。

在简单短语这一级，通过介宾短语结构，基于条件随机场这一机器学习模型，构建了英汉双语介宾短语结构知识抽取模型，并给出了英汉双语介宾短语结构知识抽取的流程。本书在该部分统计了介宾短语结构的内部和外部句法特征、给出了训练语料的预处理格式、详细说明了自身特征模板和添加特征模板的具体构成内容并与最大熵的性能进行了对比。

在复杂短语这一级，基于已有聚类算法，通过构建词汇和词性特征知识下的类别知识挖掘模型，完成了面向英汉双语专门平行语料的类别知识挖掘的探究。本书通过具体的实验证实了英汉双语词汇特征在类别知识挖掘中的性能，并给出了造成类别知识挖掘性能有差异的原因，同时使用词汇和词性的特征知识，在英汉名词，英汉名词、动词，英汉名词、动词和形容词这三种词汇和词性特征组合的基础上，探究了词性知识在类别知识挖掘上的具体表现。

王东波博士面向英汉双语平行语料这一非结构化数据抽取和挖掘的句法三个层面的知识不仅有利于知识库构建、知识服务、信息检索、信息计量等情报学中的相关研究开展，而且有助于自然语言处理中的歧义消解、知识抽取和机器与辅助机器翻译问题的解决。同时，本书所采用

的把情报学的方法和知识通过自然语言处理的技术和模型融入非结构化数据的知识挖掘和抽取中的策略具有方法论上的意义，在一定程度上具有推广价值。

王东波博士本科主修汉语言文学，硕士为中文信息处理专业，博士主攻情报学，不仅学习了语言学概论、现代汉语、古代汉语和现代汉语语法研究等语言学课程的知识，而且掌握了中文信息处理、计算语言学、语料库语言学和语言统计方法等课程的技术，还系统学习了情报学理论与方法、信息资源管理技术、信息处理与检索技术和信息智能处理与检索等情报学课程的方法，具有跨学科研究的良好基础。同时，读博期间，王东波博士在处理数据、发表论文、承担项目的过程中充分体现出了勤奋刻苦、锐意创新、百折不挠的个性。《周易》有言：君子藏器于身，待时而动。我相信，王东波博士在这个大数据时代，依据其“咬合度”极好的跨学科知识、技术和方法，凭借其勤奋、创新和坚持的个性，定能做出更多、更好的科研成果。

是为序。

苏新宁

2013 年 5 月 13 日

南京大学信息技术开发研究所

目　录

图 目 录

表 目 录

第一章

引　言

一　课题提出

课题的研究目标是基于面向网络构建的英汉双语平行语料库，结合自然语言处理和文本挖掘的相关技术和语料资源，使用情报学、机器学习和统计的相关方法，通过比较和使用英汉双语的不同语言特征，从由英汉双语语料组成的非结构化文本中挖掘或抽取出词汇、简单短语结构和复杂短语结构等句法层级上的语言知识。该课题的提出，主要出于以下三个方面的考虑。

（1）随着自然语言处理和文本挖掘技术的发展，从非结构化文本中挖掘和抽取相应的专门或通用知识以便更好地服务于基础和应用研究日益成为一种趋势。课题从更微观的句法级这一序列层级[①]入手，针对英汉双语语料这一非结构化文本对相应的知识进行挖掘或抽取研究，不仅是在这一大的研究趋势下的一种探究，而且是对整合英汉双语两种语言下的数据预处理、算法和评价方案以及特征在具体知识挖掘和抽取上的一个尝试。

（2）从英汉双语平行语料库这一非结构化数据集合中挖掘[②]和抽取句法级知识在理论研究和具体应用上具有巨大的需求。构建的英汉双语平行语料库可以直接向相关的研究提供素材，从英汉双语平行语料中挖

---

① 朱德熙：《语法讲义》，商务印书馆 1983 年版，第 15—28 页。

② Feldman, R., & Sanger, J., *The Text Mining Handbook*, London: Cambridge University Press, 2009, pp. 5—7, p. 225.

掘的词汇级句法知识不仅可以为歧义问题的解决提供相关的理论指导，而且可以对比英汉双语在词汇句法知识上的差异，从而满足解决英汉双语机器翻译、辅助机器翻译的需求。同样地，从英汉双语平行语料库获取的简单短语级和复杂短语级知识更可以直接满足树库构建、信息检索、知识推送服务等的需求。

（3）已有的英汉双语自然语言处理中的语料和技术与文本挖掘中的聚类算法确保了本研究的可行性。正是有目前标注比较成熟的英语和汉语树库，在进行词汇级知识挖掘的过程中才能完成多角度、各层次的验证，而本课题的研究之所以能得以开展，主要在于自然语言处理技术的发展和逐步成熟，同时各种聚类模型的开发和应用也为研究提供了相应的条件。

综上所述，在文本挖掘这一大的研究趋势下，结合理论和应用的具体需求，在自然语言处理和文本挖掘的相应资源、技术和方法的基础上，本书选取了从英汉双语平行语料中挖掘词汇级、简单短语级和复杂短语级这三个层面的句法知识的研究。

二 研究意义

基于英汉双语平行语料库的句法级知识挖掘和抽取对情报学、自然语言处理和相应的研究方法探究具有重要的意义。

（一）情报学方面

从具体研究内容出发，结合情报学的相关研究领域，本研究关于英汉双语平行语料库构建、词汇级、简单短语级、复杂短语级的知识挖掘和抽取对知识库构建、知识服务、信息检索、信息计量[①]等研究在某种程度上具有重要的意义。

1. 知识库构建和服务方面

面向网络构建的英汉双语平行语料库，不仅可以为英汉双语知识库的构

① 邱均平：《信息计量学》，武汉大学出版社 2007 年版，第 1—35 页。