智能 科/学/技/术/著/作/丛/书

固定短语的自动提取研究

刘 荣 著

科学出版社

智能科学技术著作丛书

固定短语的自动提取研究

刘　荣　著

科学出版社

北京

内 容 简 介

本书以固定短语自动提取为研究目标，围绕结合紧密、使用稳定原则，采取统计与规则相结合的算法，并通过历时考察最终取得固定短语。本书主要内容包括：领域高频种子词提取、通过统计量对短语的考察、通过句法规则对短语的考察、通过语义对短语的考察、历时分析对短语的考察。

本书适合高校语言学与应用语言学专业、计算语言学专业读者阅读，也可作为计算机专业学生的参考用书。

图书在版编目(CIP)数据

固定短语的自动提取研究/刘荣著. —北京：科学出版社，2016

(智能科学技术著作丛书)

ISBN 978-7-03-047383-7

Ⅰ. 固… Ⅱ. 刘… Ⅲ. 自然语言处理-研究 Ⅳ. TP391

中国版本图书馆 CIP 数据核字(2016)第 033338 号

责任编辑：魏英杰 / 责任校对：何艳萍

责任印制：张 伟 / 封面设计：陈 敬

科学出版社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

北京教图印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2016 年 2 月第 一 版 开本：B5(720×1000)

2016 年 2 月第一次印刷 印张：11 1/2

字数：228 000

定价：90.00 元

(如有印装质量问题，我社负责调换)

《智能科学技术著作丛书》编委会

《智能科学技术著作丛书》序

“智能”是“信息”的精彩结晶，“智能科学技术”是“信息科学技术”的辉煌篇章，“智能化”是“信息化”发展的新动向、新阶段。

“智能科学技术”(intelligence science&technology，IST)是关于“广义智能”的理论方法和应用技术的综合性科学技术领域，其研究对象包括：

- “自然智能”(natural intelligence，NI)，包括“人的智能”(human intelligence，HI)及其他“生物智能”(biological intelligence，BI)。
- “人工智能”(artificial intelligence，AI)，包括“机器智能”(machine intelligence，MI)与“智能机器”(intelligent machine，IM)。
- “集成智能”(integrated intelligence，II)，即“人的智能”与“机器智能”人机互补的集成智能。
- “协同智能”(cooperative intelligence，CI)，指“个体智能”相互协调共生的群体协同智能。
- “分布智能”(distributed intelligence，DI)，如广域信息网、分散大系统的分布式智能。

“人工智能”学科自1956年诞生的，五十余年来，在起伏、曲折的科学征途上不断前进、发展，从狭义人工智能走向广义人工智能，从个体人工智能到群体人工智能，从集中式人工智能到分布式人工智能，在理论方法研究和应用技术开发方面都取得了重大进展。如果说当年“人工智能”学科的诞生是生物科学技术与信息科学技术、系统科学技术的一次成功的结合，那么可以认为，现在“智能科学技术”领域的兴起是在信息化、网络化时代又一次新的多学科交融。

1981年，“中国人工智能学会”(Chinese Association for Artificial Intelligence，CAAI)正式成立，25年来，从艰苦创业到成长壮大，从学习跟踪到自主研发，团结我国广大学者，在“人工智能”的研究开发及应用方面取得了显著的进展，促进了“智能科学技术”的发展。在华夏文化

与东方哲学影响下，我国智能科学技术的研究、开发及应用，在学术思想与科学方法上，具有综合性、整体性、协调性的特色，在理论方法研究与应用技术开发方面，取得了具有创新性、开拓性的成果。“智能化”已成为当前新技术、新产品的发展方向和显著标志。

为了适时总结、交流、宣传我国学者在“智能科学技术”领域的研究开发及应用成果，中国人工智能学会与科学出版社合作编辑出版《智能科学技术著作丛书》。需要强调的是，这套丛书将优先出版那些有助于将科学技术转化为生产力以及对社会和国民经济建设有重大作用和应用前景的著作。

我们相信，有广大智能科学技术工作者的积极参与和大力支持，以及编委们的共同努力，《智能科学技术著作丛书》将为繁荣我国智能科学技术事业、增强自主创新能力、建设创新型国家做出应有的贡献。

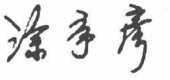
祝《智能科学技术著作丛书》出版，特赋贺诗一首：

智能科技领域广

人机集成智能强

群体智能协同好

智能创新更辉煌



中国人工智能学会荣誉理事长

2005 年 12 月 18 日

序　一

作为他的导师，以我对他的了解，我想可以用“持之以恒，铁杵成针”来概括他的学术研究。

语言信息处理是当代语言学的热点，是一门交叉学科，需要将语言学、计算机科学、数学等多学科知识交叉结合。其任务艰难，工作量大，需掌握的新知识、新方法、新技术多，要解决的新问题艰巨，都必须持之以恒，下苦功夫才有可能完成。

刘荣博士一直坚持学术研究。他属于那种“为伊消得人憔悴，衣带渐宽终不悔”的人。在三年求学期间，他不去想如何职场升迁，也不是为了读个博士学位“镀镀金”，一直都坚持在 DCC[①] 学习，真正做到了脱产学习，为此还放弃了很多赚钱的机会，这对一个三十多岁的人来说，是难能可贵的。他选择的研究对象“固定短语”，是一个非常困难的题目，不是随便研究研究就能出成果的选题。

《现代汉语词典》既收录“词”也收录“语”，“语”是指成语、谚语、歇后语、惯用语等这些非常固定的短语。但是在实际语言生活中，有大量“结合紧密、使用稳定”的短语，例如“小升初”“改革开放”“全国人民代表大会常务委员会”，这些短语收录与否就很难说。这些“语”在频度、使用度、流通度等方面常常高于一般的熟语，甚至远高于已经收入词典的偏僻成语，如何发现并分析它们是一个难题，也是语言应用中迫切需要解决的实际问题。在语言教学（尤其是第二语言教学）、词典编纂、语言翻译中，包括机器翻译、自然语言理解等语言信息处理方面，都需要解决“语”的收录与说解。“结合紧密、使用稳定”完全是人的一种感觉，我们称为语感。但如果没有量化的操作标准，依然无法判断哪些“语”是固定短语，应该收入“语表”。收了也说不清、道不明，难以服人，这恰恰是对“语”的语感。

我们有一些博士在动态流通语料库的基础上对某一种“语“及”语

---

① 2001 年，北京语言大学建立“DCC 博士研究室”，后归入“应用语言学研究所”。DCC 就是英文 Dynamic Circulating Corpus（动态流通语料库）的缩写。研究室每周有一次 DCC 相关研究的讨论，同学、老师、外请的专家学者都可能是主报告人，这一讨论课持续多年至今，客观上大大推进了动态语言知识更新的研究。

表“(比如:流行语、术语、字母词语、新词语、熟语单位等)进行定量和定性分析研究,刘荣博士知难而上,在我们实验室前期研究的基础上,选择对整个“结合紧密、使用稳定”的“语”进行定量和定性分析研究,这是需要决心和勇气的。

对于他来说,这是一条十分难走的研究之路,仅仅就外文资料而言,他就坚持通读了 ACL(Association of Computational Linguistics,计算语言学协会)1994 年至今所有相关专题的论文,共 236 篇。他的研究选用 DCC 2006～2008 年 15 份主流媒体报纸中的教育领域文本作为考察对象,对教育领域“结合紧密、使用稳定”的固定短语进行提取研究。教育领域 3 年语料总计文本数量 142,069 个,计字节数量 216,154,807 字节,提取的候选串总计 24,116,507 个。他从统计量、句法、语义、历时使用考察等多个角度对语的“结合紧密、使用稳定”做了定量与定性分析。基于大规模真实语料库对固定短语进行了计算机提取研究,经综合运用“统计量+规则+人工评测”的方法,获得了一个固定短语表,并对表中的固定短语进行了一定的分析。

他对“结合紧密,使用稳定”原则的具体判断提供了基本方法和手段,为固定短语的深入研究提供了一种量化考察途径。

该书紧密围绕国际语言学界热点问题,基于动态流通语料库,采用多种方法多种角度探讨汉语中的固定短语问题,在理论与实际应用方面都有较大意义。这是对刘荣研究的肯定,绝非溢美之词。

刘荣最重要的创新就是对“结合紧密、使用稳定”原则确定了一种机器可操作的具体办法,这可以说开拓了传统语言学对于固定短语研究的新方法。

他目前所有本体研究课题都以动态语言知识更新理论为基石,又将研究对象由中文扩展到英文。这说明他对 DCC 研究方向的坚持与领悟,也是对动态语言知识更新理论发展的贡献。

我为他对学术研究的执著、对动态语言知识更新的赤诚而感动,他告诉我:持之以恒,铁杵成针。我为有这样的学生而骄傲。

希望刘荣能够始终如一,是为序。

张 普

2014 年 11 月 5 日

序　二

甲午岁末，刘荣副教授申报的《语言信息处理学科建设和研究》项目获得太原理工大学学科建设专项资助。承蒙抬爱，邀我作序。自知“隔行如隔山”、“术业有专攻”，实不敢妄自品头论足。然几番相让不得允，又恐“却之”有“不恭”之嫌，故捧卷在手，细细拜读。

刘荣副教授博学善思，涉猎广泛，文理兼修，学识渊博。本科期间就读于科技英语专业，硕士阶段转而研究信息工程，考取博士研究生时又选择了语言学与应用语言学方向。数载寒暑，几番耕耘，造就了他包容开放的学术态度和兼容并蓄的知识结构。近年来，刘荣副教授主要从事语料库建设与加工、文本分类、固定短语提取、计算机辅助对外汉语教学和词典编纂等方面的研究工作，取得了不少令人瞩目的成绩，本书就是其中之一。

当前，全球业已步入信息技术迅猛发展的大数据时代，这使得语言的信息处理和大数据应用成为可能和常态。作为涵盖语言学、数学、计算机科学等几大领域的交叉学科，计算语言学正以“基于大规模真实数据提供高质量研究成果”的突出优势，在语言学研究领域中占据重要地位。

时下，“学说中国话”正在成为世界的潮流；同时，遍布全球的孔子学院更在积极创造推广汉语教学、传播中国文化的品牌和平台；加之，现代汉语的不断创新和发展，新词新语的日益成型和完善，对与时俱进地推进汉语本体研究和应用研究提出了新的命题。特别是，通过信息处理手段对新词新语进行及时收录和提取研究，既有本体研究的意义，更有实际应用的价值，对于弘扬民族文化、推动国家繁荣具有重大促进作用。

该书有诸多亮点。从整体来看，对固定短语的定义有所突破，在短语计算方面有较大创新。作者基于国家语言监测中心平面媒体分中心语料库，选取 2006～2008 年语料，采用统计与规则相结合的计算方法，通过历时考察方法确定固定短语。其所用语料规模之大，包括统计、句

法、语义、历时考察等在内的方法之全面，在文科研究中均属罕见。特别是，作者充分发挥文理交叉的知识优势，结合具体的研究课题，采取步步推进的方案，利用数据挖掘的方法自动发现固定短语，业已广泛应用于多个实际领域。

学问人，常思进；唯笔耕不辍，方著作等身。诚挚祝贺喜获殊荣，热切期待再谱新篇。

是为序。

梁丽萍

2014 年 12 月

前　言

20世纪80年代，固定短语作为一个术语被学界认可。此前，学界习惯将此种语言现象称为固定词组。此后，学界对固定短语的研究日益深入，对于固定短语的认识有了长足的发展。固定短语不再局限于词典所收录的历经长时间发展的成语、谚语、歇后语、惯用语。为了和上述固定短语区别，学者们提出了“准固定短语”或“类固定短语”。虽然名称不一样，但是固定短语所指的范围却是宽泛了许多。

随着信息化浪潮、计算机与互联网越来越深入大众的生活，汉字输入、信息检索等自然语言处理领域迫切需要固定短语从而提高准确率和效率。用Sag的话说，固定短语已经成为自然语言处理的瓶颈。那些已被词典或语典收录的固定短语可作为词库或底表，但自然语言处理更需要那些还未被词典或语典收录的“结合紧密、使用稳定”的固定短语。

语言随着社会发展而不断变化，尤其是改革开放以来，新词新语层出不穷。其中一些短语的频度、使用度、通用度、流通度常常是高于一般的熟语，甚至远高于已经收入词典的偏僻短语。这些新词新语是否也应该与传统的成语、谚语、歇后语、惯用语一样被收录？更重要的是如何去发现它们？

在上述背景下，为了给词典编纂提供客观、科学的新材料，促进汉语固定短语的研究，为中文信息处理提供颗粒度比词更大的资源，笔者认为，针对短语的结合紧密、使用稳定性质开展及时、全面、客观的研究是非常必要的。考虑到学界前辈已经对成语、谚语、歇后语、惯用语进行了大量的研究，本书提出较为宽泛的固定短语定义，并主要探讨如何将“结合紧密、使用稳定”原则具体化和可计算化，从而提出固定短语的自动提取方法。

本书以动态语言知识更新理论为指导，以动态流通语料库为实验

平台，从多个维度对短语的结合紧密与使用稳定进行研究。为了方便评价提取效果，本书以一个领域为研究对象，但是提取方法可扩展至不同领域。

本书首次对结合紧密、使用稳定原则确定了机器可操作的具体办法。在多特征融合的框架内从统计、句法、语义的角度，对“结合紧密”程度进行量度；从历时考察的角度对“使用稳定”进行量度。本书所做研究为考察短语固定程度提供了一种从定量到定性分析的方法。

感谢太原理工大学学科建设专项经费“语言信息学科建设与研究”、教育部人文社会科学研究青年基金项目“基于动态语料库的新词语实时研究”、山西省归国留学人员科研资助项目(2012-041)经费资助。

由于笔者水平有限，错误和疏漏之处在所难免。恳请读者谅解和批评指正。

刘　荣

2015 年 5 月

目　录