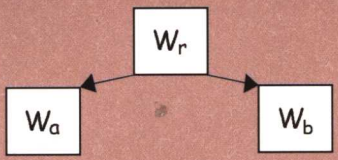
基于词联接的
自然语言处理技术及其应用研究

JIYU CI LIANJIE DE
ZIRAN YUYAN CHULI JISHU JIQI YINGYONG YANJIU

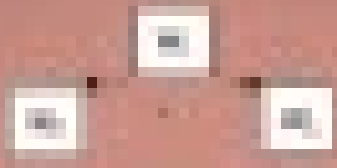


李良炎 著

學林出版社

基于词联接的

自然语言处理技术及其应用研究



本书出版得到国家自然科学基金项目（60173060）的支持

基于词联接的
自然语言处理技术及其应用研究

JIYU CI LIANJIE DE
ZIRAN YUYAN CHULI JISHU JIQI YINGYONG YANJIU

李良炎　著

学林出版社

图书在版编目(CIP)数据

基于词联接的自然语言处理技术及其应用研究 / 李良炎著. —上海：学林出版社，2007.9
ISBN 978-7-80730-456-2

Ⅰ.基… Ⅱ.李… Ⅲ.自然语言处理-研究 Ⅳ.TP391

中国版本图书馆 CIP 数据核字(2007)第 148724 号

基于词联接的自然语言处理技术及其应用研究

作　　者—— 李良炎
责任编辑—— 李晓梅
封面设计—— 周剑峰
出　　版—— 上海世纪出版股份有限公司
学林出版社(上海钦州南路 81 号 3 楼)
电话：64515005　传真：64515005
发　　行—— 新华书店上海发行所
学林图书发行部(上海钦州南路 81 号 1 楼)
电话：64515012　传真：64844088
照　　排—— 南京展望文化发展有限公司
印　　刷—— 上海肖华印务有限公司
开　　本—— 889×1194　1/32
印　　张—— 5.5
字　　数—— 14 万
版　　次—— 2007 年 9 月第 1 版
2007 年 9 月第 1 次印刷
印　　数—— 3000 册
书　　号—— ISBN 978-7-80730-456-2/G·129
定　　价—— 15.00 元

前　言

随着人类社会信息化程度和计算机软硬件水平的提高，自然语言处理（Natural Language Processing，简称 NLP）技术逐渐成为计算机应用和人工智能研究的热点，其基本技术目标是让计算机具有类似人的语言智能，例如能够像人一样听、说、读、写。围绕 NLP 技术逐渐形成了一个专门的学科——计算语言学。该学科属于交叉学科，涉及语言学、心理学、心理语言学、脑科学、计算机科学、哲学、逻辑学、人工智能、数学、信息论、文学、美学等诸多学科或领域。从 20 世纪中叶以来，虽然不同学科和领域的无数研究者投入了大量的研究资源，探索出了一些有效的技术，取得了一定的成果，但离实现基本技术目标还显得非常遥远。值此世纪之初，有必要从更高、更深的层面重新审视 NLP 技术的研究背景、研究目标和研究途径，在继承现有技术的基础上大胆创新，探索出切实可行、面向未来的新技术。本研究旨在朝这个方向作出努力，以抛砖引玉，促进 NLP 技术的发展。

在国家自然科学基金项目“计算机辅助文学艺术创作研究——诗词曲联”（60173060，2002—2004）的支持下，在深入分析 NLP 技术背景的基础上，本研究提出并初步构建了基于词联接的 NLP 技术（Term Connection Technique for NLP，简称 TCT），并应用到诗词语言处理系统（Poetry Processing System，简称 PPS）中。理论研究和应用研究的结果表明，TCT 是一种继承并发展已有技术、高度综合与包容、操作性强、有效的 NLP 技术。本书的主

要内容包括：

第一章介绍了 TCT 的技术背景，包括 NLP 的根本难点、现阶段 NLP 的根本目标、NLP 技术的发展阶段和趋势，提出了一系列较为独特的观点。NLP 的根本难点在于自然语言具有认识性和不确定性，现阶段目标应是受限语言智能仿知技术。这一目标是在分析了 NLP 根本难点、现阶段的计算机技术、已有人工智能和 NLP 研究成果的基础上提出来的，因此是切实可行的。NLP 技术分为技术探索、规则技术兴起、统计技术兴起三个阶段，主要发展趋势是实例技术可能成为主流技术，规则技术和统计技术可能成为辅助技术。

第二章介绍了 TCT 的基本原理、技术结构、哲学基础、理论基础、方法基础，从宏观上初步构建了 TCT。TCT 的基本原理是基于受限自然语言环境建立动态语料库，基于动态语料库建立词联接实例知识库和高级知识库，基于知识库进行受限自然语言处理。TCT 具有完整、简明、合理的技术结构，包括：TCT 知识表示技术(TCTR)、TCT 知识获取技术(TCTO)、TCT 语言分析技术(TCTA)、TCT 语言评价技术(TCTE)、TCT 语言生成技术(TCTC)、TCT 语言修改技术(TCTM)、TCT 语言输入输出技术(TCTIO)。TCT 的哲学基础是易学，这是本研究最具特色的地方。易学是中国古代哲学的精华，具有朴素的唯物辩证思想和系统观，对于认识和认识建模都具有很强的指导作用。TCT 的理论基础是神经认知语言学，该理论强调语言理论不悖于大脑神经事实，综合了联接主义和符号主义等理论的长处，是一种面向 NLP 的语言学理论。TCT 的方法基础是基于实例的知识加工，以实例为基础，可以更好地综合各种知识加工技术。

第三、四、五、六章分别介绍了 TCT 技术结构中的 TCTR、TCTO、TCTA、TCTE，从微观上初步构建了 TCT。由于研究资源有限，在斟酌各个技术模块关系的前提下，着重研究这四大模

块。在每个技术模块的研究中都注意继承与创新相结合，注意各个技术模块在功能上的独立性与系统性。与传统的 NLP 技术相比较，各个技术模块都具有各自的特色和创新点：在 TCTR 中提出了自然语言知识结构和要素、抽象概念树（Abstract Concept Tree，简称 ACT）、体验语义和价值语义等；在 TCTO 中提出了语言知识设计、语料标注规范、语言知识获取与管理等；在 TCTA 中提出了词联接最大语义符合度计算和最优句树搜索的初级语言分析算法；在 TCTM 中提出了豪放与婉约语言风格的计算模型。

第七章介绍了基于 TCT 的 PPS 开发和测试。由于构建 TCT 之初就认识到了自然语言认识性中的体验性这一难点，而体验性是自然语言尤其是文学语言的重要特征，因此 TCT 应当能够比传统的 NLP 技术更好地处理文学语言。诗词语言是一种典型的文学语言，开发 PPS 对传统的 NLP 技术和 TCT 都是一种挑战。本研究基于当前水平的 TCT 开发了 PPS，完成了语言知识类设计、语言知识库设计、模块设计、界面设计，并在大量数据准备工作的基础上进行了诗词语料标注测试、诗词语言初级分析测试、诗词语言豪放与婉约风格的评价测试。测试结果表明，PPS 取得了成功，验证了 TCT 的有效性。

本书是在笔者的博士学位论文基础上完成的。主要研究工作得到了重庆大学计算机学院陈廷槐教授、吴中福教授和何中市教授的指导。他们严谨的治学态度、勤勉的工作作风、不倦的教学风范和创新的思维方法给了我极大的影响和帮助。谨以此书献给我的恩师、学友和家人！

李良炎

2007 年 3 月 10 日

目　录

第一章 技术背景

自然语言处理(Natural Language Processing,简称 NLP)在我国又称"语言信息处理",根据我国 GB12200.1－90 的定义,指"用计算机对自然语言的音、形、义等信息进行处理,即对字、词、句、篇章的输入、输出、识别、分析、理解、生成等的操作与加工"[1]。NLP 是计算机应用和人工智能研究的热点和难点,是计算机技术和信息社会高速发展到一定阶段的必然课题,具有极大的挑战性和广阔的应用前景。

1.1 NLP 的根本难点——认识性和不确定性

NLP 技术的终极目标是实现一种能够像个体人一样进行语言加工的智能机器。从信息论角度来看,这种加工包括语言输入、语言存储、语言运算、语言输出等基本过程。现阶段由于计算机在这些方面的突出表现,理所当然成为 NLP 的基本工具。然而 NLP 研究并不顺利。从 20 世纪 50 年代最初的小小成功而盲目乐观,到 60 年代遭遇挫折而陷入低潮,再到 80 年代至今出现研究热潮而艰难探索,人们逐渐得出这一结论:NLP 是非常困难的。

要列出 NLP 的困难之处,显然不胜枚举,但找到根本难点是关键。NLP 的根本难点是什么?在于**自然语言具有认识性和不确定性**。人类具有很强的认识能力和处理不确定性的能力,而现阶段的符号处理式计算机无法具备克服这一根本难点的能力。当然计算机技术是不断发展的,如量子计算机、DNA 计算机现在已崭

露头角[2]。如果将来这些计算机能够克服 NLP 的根本难点，那么实现 NLP 的终极目标是可能的。

为了准确地理解自然语言的不确定性和认识性，可以通过与人工语言的对比来分析。本节主要参考文献为[3～24]。

1.1.1　自然语言的认识性

认识是哲学和心理学的重要概念，是存在物具有主体性的重要标志，是存在物与其存在环境之间信息交换的重要手段。从人类的角度来看，认识是世界在人类意识中的反映。世界即人类的存在环境，包括客观世界和主观世界。自然语言作为人类基本的交流工具，在认识过程中发挥着重要作用，因此具有认识性。显然，人类认识的复杂性将会带来自然语言的复杂性，进一步给 NLP 带来意想不到的困难。人工语言是人类设计的、适用于计算机的语言，其认识性是很有限的。计算机必须依赖人工语言，人类的认识主要依赖自然语言而不是依赖人工语言。

自然语言的认识性表现为认知性、体验性，决定于人类认识的复杂性。人类具有很强的认识能力，这是目前的计算机不具备的[3]。

体验与认知是心理学的一对重要范畴，但在这里考察两者的起源与准确意义并不重要。重要的是，用这两个词来描述人类认识世界的两种基本方式[5, 6]可能更容易让人理解。从多个层面来看，可以确信这两种基本方式的存在。从物质层面来看，人类的大脑分为左脑和右脑；从意识层面来看，人类的意识分为理性意识与感性意识；从心理层面来看，人类的思维分为逻辑思维和形象思维。在哲学、生理学、心理学等多学科的研究中，人类逐渐发现：左脑、理性意识、逻辑思维之间具有密切联系，是人类理性认识世界的基础；右脑、感性意识、形象思维之间具有密切联系，是人类感性认识世界的基础。在这里，我们称人类对世界的理性认识方式为**认知**，人类对世界的感性认识方式为**体验**。由此可见，理性、感

性是两个更基本的概念。然而要弄清这两个概念实非易事，很可能面临哲学上的争执而无法做具体工作。因此本书在这里只给出基于人类普遍价值——真、善、美[7~9]对理性和感性的理解。

理性是人类的**求真意识**，**感性**是人类的**求美意识**。人的认识被意识所主导，认知是在求真意识主导下的认识，体验是在求美意识主导下的认识。显然，在认知过程中一切都是围绕世界的本来面目所谓"**真**"来展开的，在体验过程中一切都是围绕世界的和谐关系所谓"**美**"来展开的。人类的生存必须依赖于认知和体验。通过对世界的认知人类发展了科学，通过对世界的体验人类发展了艺术，增强了人类改造和适应世界的能力。由此可见，有必要像重视左脑与右脑一样重视认知与体验。由于科学与艺术的重要区别，可以将艺术中的自然语言称为文学语言，将科学领域中的自然语言称为实用语言。人们日常生活领域中的自然语言往往是文学语言与实用语言的混合体。

其一，**实用语言具有认知性**，人工语言具有有限的认知性。实用语言是为人类的认知服务的，必须准确地记录人类的认知过程和成果，并准确地交流和传承。从这种要求来看，实用语言强调客观和共性。但作为自然语言的实用语言很难达到这种理想化要求。从客观要求来看，人类对世界的认知力求客观，但由于世界的巨大复杂性和人类认知能力的局限性，很难真正弄清世界的本来面目。科学求真的过程实际上是一个不断观察、假设、验证的过程。从哲学意义上看，人类的认知过程永远离不开人类这一主体，其认知结果不是世界的本真结论。从共性要求来看，人类对实用语言的理解力求共性，但由于人类知识结构和表达能力的个体差异，常常很难对同一实用语言有完全相同的理解。语言交流的过程实际上是一个编码、传输、解码的过程。虽然与目前的计算机相比，人类在语言交流中的容错能力是很强的，但从信息论角度来看，人类的语言交流总是有信息误差、信息耗损、信息补缺现象的。

人工语言的客观与共性优于实用语言，通过人类预先设计和计算机稳定运行已经得到有力保证。但人工语言往往只是实用语言中某些问题求解的形式化描述，是非常局部和有限的。例如，很多问题是无法用人工语言来描述、求解并在目前的计算机上运行的[2]。

其二，**文学语言具有体验性**，人工语言不具有体验性。文学语言是为人类的体验服务的，必须准确地记录人类的体验过程和成果，并准确地交流和传承。从这种要求来看，文学语言强调主观与个性。与实用语言相比，文学语言的重心在人类自身的和谐状态而不在世界的真实面目，因此完全用一方的标准来要求另一方是行不通的。正是由于强调主观与个性，使得驾驭文学语言是一件远比驾驭实用语言困难的事情。对于人类如此，对于目前计算机来说更是如此。从主观要求来看，人类对世界的体验力求主观，在认知世界的基础上必须基于自身立场产生感受、价值、态度、兴趣、情感、动机、志向等体验性反应。文学语言的任务就是技巧性地描述一些基本的认知事实，重点却在表现这些体验性反应。对于人类特别是文学家来说，难点在于如何基于普通的认知事实产生高水平的体验，如何运用技巧有效地表达这种体验。从个性要求来看，人类对文学语言的理解力求个性，语言接受者必须从自身立场产生各种与语言发送者同质（原创性）和异质（再创性）的体验性反应。对人类特别是文学读者来说，难点在于如何基于文学语言中传递的非常有限的认知事实产生高水平的同质和异质体验。高水平的文学欣赏更强调异质体验。人工语言由于将客观与共性绝对化，因此是不考虑主观与个性的，不会携带体验信息。真正像人一样产生体验性反应对目前的计算机是绝不可能的，即使模拟也相当困难。

总之，人类认识的复杂性是自然语言认识性的根源，使得NLP出现重重困难。由于人工语言与实用语言在认知性上的天然联系，因此用目前的计算机处理实用语言比处理文学语言应该

容易。人们确实在文学语言处理上遇到了超乎想象的困难，往往倾向于回避这一问题并寄希望于未来。人们甚至没有从 NLP 角度对文学语言进行定义。鉴于此，本书从四个角度给出对文学语言的理解。

其一，词典对“文学”的定义：“以语言文字为工具形象化地反映社会生活斗争的艺术，包括戏剧、诗歌、小说、散文等”[13]。根据该定义可见，文学语言的突出特点是具有形象性。这种定义是不严格的，没有抓住文学语言的本质特征。

其二，本书从心理学角度的定义：文学语言是体验型语言，实用语言是认知型语言。文学语言主要为人类的体验活动服务，最大限度满足人类体验需求是其基本目的。不能丰富人类体验的自然语言不可能是有水平的文学语言。实用语言主要为人类的认知活动服务，最大限度满足人类认知需求是其基本目的。不能丰富人类认知的自然语言不可能是有水平的实用语言。

其三，本书从计算语言学和篇层面的定义：文学语言是符合文学体裁的语言，实用语言是符合实用体裁的语言。**体裁**就是作品的语义类型。从计算语言学角度来看，体裁就是一系列规则和约束，因此具有操作性、鉴别性。体裁有很多种类型[14]。文学体裁如诗歌、小说、散文、戏剧等，实用体裁如科技论文、新闻、信函、公文、合同、说明书等。每种体裁都从篇层面对语言的内容和形式作了具体规定。从篇层面来看，文学语言与实用语言界限分明。例如，一篇作品不可能同时是小说和新闻。

其四，本书从计算语言学和句层面的定义：文学语言是运用积极修辞技巧的语言，实用语言是运用消极修辞技巧的语言。**修辞技巧**就是句子的语义类型。积极修辞力求生动、新颖，消极修辞力求真实、准确。从计算语言学角度来看，修辞技巧就是一系列规则和约束，因此具有操作性、鉴别性。积极修辞如比喻、拟人、夸张、用典等，消极修辞如说明、证明、注释等。从句层面来看，文学

语言与实用语言界限分明。例如一个句子不可能同时是比喻句和证明句。

从不同层面来看，文学语言与实用语言常常是有机融合的。篇层面的文学语言常常包含句层面的实用语言，篇层面的实用语言常常包含句层面的文学语言。文学作品一般不可能全是比喻句。非文学作品也需要文学语言点缀以提高表达效果，如外交辞令中用典就会增色不少。正因为如此，面向日常生活领域的 NLP 不可能回避文学语言处理问题。

1.1.2 自然语言的不确定性

确定性与不确定性之争是科学的一个重大命题。从日常经验和科学传统来看，人们更倾向于认为对象具有确定性。在这种前提下，正向和反向的长期预测是有效的。然而普利高津解构了这种世界观[25]，不确定性才是基本的自然法则，确定性只是表象。在不确定的前提下，正向和反向的短期预测可能是有效的，长期预测很可能是无效的。人类按照确定性法则构造了人工语言，使得计算机从理论上能够处理任何用人工语言表述的操作。但计算机依然会受到来自软件描述和硬件运行方面的不确定性干扰。相比之下，自然语言的不确定性就更突出了。

自然语言的不确定性表现为无限性、动态性、模糊性、或然性，决定于世界的巨大复杂性。人类具有处理不确定性问题的强大能力，这是目前的计算机所不具备的。

其一，**自然语言是无限的**，人工语言是有限的。正是因为自然语言无限性的存在，我们无法认为自然语言是确定的。由于自然语言是人类认识世界的符号集，世界时空的无限性决定了自然语言的无限性，如自然语言应当包含无限词集、无限语义集。由于人工语言是关于人类认识的主观世界的符号集，主观世界的有限性决定了人工语言的有限性，如人工语言只包含有限词集、有限语义集。以计算机作为工具处理自然语言，首先必须用人工语言对自然语