短语识别与信息抽取的隐马尔可夫模型的方法研究

METHOD STUDY ON HIDDEN MARKOV MODEL FOR PHRASE IDENTIFICATION AND INFORMATION EXTRACTION

李荣◎著

兵器工业出版社

山西省高校科技开发项目资助（项目编号：20101120、2013147）
忻州师范学院重点学科建设项目资助（项目编号：ZDXK201204）

短语识别与信息抽取的隐马尔可夫模型的方法研究

李 荣 著

兵器工业出版社

内容简介

短语识别与信息抽取是智能信息处理领域的重要组成部分。本书面向智能信息处理的实际需要，介绍了短语识别与信息抽取的 HMM 方法，并提出了改进混合 HMM 方法和其他模型算法，分别对这两个问题的几种实现方法进行了比较研究。最后，对短语识别与信息抽取的关系和其 HMM 方法进行了比较研究。

本书可作为高等院校计算机专业高年级学生的教学参考书，也可供从事智能信息处理与人工智能研究的相关人员参考。

图书在版编目（CTP）数据

短语识别与信息抽取的隐马尔可夫模型的方法研究 / 李荣著. -北京：兵器工业出版社，2013. 12

ISBN 978-7-80248-993-6

Ⅰ. ①短… Ⅱ. ①李… Ⅲ. ①人工智能-信息处理-研究 Ⅳ. ①TP18

中国版本图书馆 CIP 数据核字（2013）第 285105 号

出版发行：兵器工业出版社

发行电话：010-68962596，68962591

邮　　编：100089

社　　址：北京市海淀区车道沟 10 号

经　　销：各地新华书店

印　　刷：北京业和印务有限公司

版　　次：2013 年 8 月第 1 版第 1 次印刷

责任编辑：朱婧

封面设计：视觉揽胜

责任校对：郭芳

责任印刷：王京华

开　　本：140×203mm 1/32

印　　张：11

字　　数：220 千字

定　　价：22.00 元

前 言

中国正在频频地叩击信息化的大门，在我们面前伸展着一条有着无穷希望的路，但这却是一条充满荆棘坎坷的路。这荆棘与坎坷，是尚待完善的发展 Web 信息事业的体制和机制，是智能信息技术科学研究的蹒跚步履。在智能信息技术中，信息处理与抽取是其关键之一。

本书面向智能信息处理的实际需要，分别介绍了短语识别与信息抽取的隐马尔可夫模型（HMM）方法，并提出了另 3 种短语识别方法与 HMM 识别方法进行比较，共提出了 7 种信息抽取 HMM 方法进行比较分析。全书共分 6 章。

第 1 章是概述部分。首先引入论著主题，然后分别讨论了短语识别与信息抽取这两个问题的研究背景、研究意义、国内外研究现状与发展趋势。

第 2 章阐述了 HMM 的理论基础。简述了 HMM 模型，阐述了 HMM 的三个基本问题及对应解决算法，分析讨论了 HMM 算法在应用中应注意的问题，针对一阶 HMM 的缺陷，阐述了二阶 HMM 模型及对应问题的算法。

第 3 章详细阐述了短语识别的 HMM 方法及与其他识别法的比较研究。首先介绍了短语的基本知识，确立了短语的标注体系和分类标准；然后分别介绍了基于规则、基于 SVM、基于 HMM、基于遗传算法 HMM 的短语识别方法，这 4 种方法分别从方法原理、识别算法、实验系统和实验数据分析进行了详细的介绍，最后对这 4 种方法分别从理论与实验两方面进行了比较研究。

第 4 章详细阐述了信息抽取的 HMM 方法及比较研究。本章采用了基于 HMM、基于改进遗传算法 HMM、基于改进模拟退火 HMM、基于遗传退火 HMM、基于最大熵 HMM、基于混合条件

模型、基于自适应混合智能算法优化 HMM 共 7 种抽取方法，这 7 种方法分别从方法原理、改进之处、识别算法、实验系统和实验数据分析进行了详细的介绍，最后对这 7 种方法分别从理论与实验两方面进行了比较研究。

第 5 章分析了短语识别与信息抽取的内在关系，并对解决这两个问题的 HMM 方法进行了比较和分析。

第 6 章是结语部分。本章对研究工作进行了全面总结，简要概括了本研究取得的主要成绩，讨论了本研究工作对智能信息处理研究的意义，并提出了进一步研究的计划和目标。

本书的研究工作是跨现代汉语语法和智能信息处理两个领域进行的。一方面，研究的具体结果对推进智能信息处理技术的发展有直接的应用和参考价值；另一方面，从智能信息处理的角度来审视现代汉语语法研究，可以为研究工作提供一个清晰的实用背景，可以注意到以往面向人的研究不容易注意到的一些问题。希望对从事智能信息处理实际应用开发工作的科研人员、在计算机语言学这一交叉学科领域辛勤耕耘的研究人员，以及汉语语法研究工作者，都能起到一定的参考作用。

需要指出的是，本书得到了山西省高校科技开发项目(基于改进隐马尔可夫模型的 Web 信息抽取研究，No.20101120)，山西省高校科技开发项目(基于群智能优化算法的 Web 信息抽取研究，No.2013147)及忻州师范学院重点学科建设项目（ZDXK201204）的资助，同时也是忻州师范学院科研基金项目(基于 HMM 的汉语 NP 自动识别方法研究，No.200623)的研究成果的结晶。

本书是笔者在短语识别与信息抽取领域内多年研究成果的结晶，书中内容在得到许多专家学者的指导和宝贵意见后经过若干次调整修正，并经多次仔细校对，但错误疏漏之处，恐仍难免。在请读者包涵谅解的同时，也恳请专家同行多批评指正。

作者

2013 年 5 月

目 录

第1章 概 述

1.1 论著主题的提出

近年来，计算机的普及得以加速进行，互联网已进入千家万户，网上资源已经呈爆炸趋势。面对浩如烟海的网上信息，人们越来越需要搜索引擎、机器翻译、信息抽取等技术的帮助，各大网络公司都在为改进和开发这方面的产品而努力。另外，以智能信息处理为主要对象的语言工程已成为国际上关注的热点，计算机必须从传统的对文本形式的加工，发展到对文本内容的加工，才能从互联网上大量未经预处理的非结构化生语料中，抽取有用的知识。

句法分析在自然语言处理领域中具有十分重要的地位，同时也是公认的研究难题。通过句法分析得到输入的某种结构表示，如完整的分析树或分析树片段集合，是计算机理解自然语言的基础。而信息抽取通常只是对某一领域中数量有限的事件或关系进行抽取，并不需要得到句子的完整结构表示，加之完全分析技术的鲁棒性和时空开销都难以满足信息抽取系统的需要，所以越来越多的浅层句法分析技术已经成为信息抽取领域的一个趋势。

根据汉语的特点，中文信息的抽取具有一定的特殊性，需要自然语言处理技术——句法及语义分析的支撑。句法及语义

分析包括句法成分的识别和标引，关键词抽取，检索特征集的抽取、索引等。信息抽取的分析过程通常可称作“浅层的”或“部分的”句法及语义分析(只分析所需要的部分)，即找出代表指定信息的词汇、短语等块状语言结构，而不是去弄清楚每一语句的句法结构树。在语法分析阶段，一个主要问题是解决信息所包含的事件、消息或事实的有关名词性短语和动词性短语的识别问题。

由上述可知，短语识别作为一种主要的浅层句法分析技术，在信息抽取中具有重要的作用。其目标在于通过牺牲分析的完整性和深度为代价，换来分析信息的健壮性和效率，克服传统句法分析所遇到的困难，以便在大规模真实信息抽取中得到有效的应用。

鉴于上述思想的引导，基于多年来致力于短语识别与信息抽取的隐马尔可夫模型（Hidden Markov Model，HMM）方法的研究，提出了本论著主题。本研究工作是在解析短语识别与信息抽取二者关系的基础上，分别对 HMM 在这两个领域的应用进行了深入研究，且由此引出 HMM 改进方法与 HMM 混合方法，并把这些方法的识别与抽取效果进行比较分析和研究，以进一步提高其识别与抽取精度和效率，为后期深一层次的智能信息处理研究打下了坚实的理论与应用基础。

1.2 短语识别概述

1.2.1 研究背景及意义

互联网是一个以英语为主导语言的网络，极大地限制了我国在网上进行信息交换与实现资源共享的能力。这就迫切需要中文信息处理技术的发展适应当前形势的需求。而如何实现语言的计算机自动理解也正是当前计算语言学面临的一项难题。其实这项技术的关键是计算机的语言分析技术。汉语的计算机自动理解过程涉及分词、词性标注、短语标注、短语分析、语义理解等多级层次。每一层次的加工完成都需要形式化了的语言知识的介入。尽管国外计算语言学的发展为我们提供了多种语言知识形式的方法，但是，汉语知识本身的欠缺使在每个层次上汉语的自动理解都依然面临着重重困难。在分词和词性标注基础上，短语自动标注和短语分析研究，在国内也有了近十年的历史。近几年来，对汉语短语分析方法、依存关系标注、基本句型分析等方面的探索，为进行比较系统全面的短语分析积累了丰富的经验。自然语言处理经历字处理、词处理阶段，在理论上和实践上都取得了令人满意的成果，现在已经发展到用计算机进行短语处理和句法分析的时期。

短语分析常称作浅层句法分析。它是针对处理开放环境下自然语言处理的特点而提出的一种句法分析策略，其基本思想是：真实文本中句子变化极其复杂，性能好的完全句法分析器

在可预期的将来是不可能实现的，于是现阶段干脆“退而求其次”，先谋求对句子中的一些相对简单的成分（组块）的识别（这些成分对语言应用系统是十分有价值的），从而使得任务在某种程度上得到简化。对汉语来说，短语分析研究本身有非常重要的意义。因为短语在汉语中具有特别重要的地位，汉语短语是句子的重要组成部分，也是信息传递不可缺少的基本单位，它的内部结构比较稳定，往往作为一个整体和句子和其他成分发生作用，并且它的构造原则和句子的构造原则也基本一致。从语法系统说，汉语语言结构可以划分出词、短语、句子等不同的单位。短语是由两个或两个以上的词，按照一定的语法规则组成，表达一定意义的语言单位。短语作为汉语语法研究中的一个单位，正好处于联系词和句子的桥梁位置上。它的构造原则和句子的构造原则基本一致，这和英语的短语有很大的差别。朱德熙先生认为，“如果我们把各类词组结构和功能都足够详细地描写清楚，那句子的结构实际上也就描写清楚了，因为句子不过是独立的词组而已。”短语分析是对构成句子的短语内部结构成分、结构层次和结构关系进行分析，不涉及句子的语气、语调，不必考虑句子的语用因素，也暂不考虑句首修饰语，故短语分析也称为句法分析。句法分析有浅层和完全之分，浅层句法分析只要求识别句中某些结构相对简单的成分，如非递归的名词短语、“名词”+“动词”词语串、实体名等，不要求得到句子的完整句法分析树；而完全句法分析则要求通过一系列分析，最终得到句子的完整的句法树。完全句法分析是自然语

言处理的核心部分，是自然语言处理技术中从“词”的理解上升到“句”的理解的基石。近几年来，中文信息处理技术发展很快，进行现代汉语语料库句法自动分析研究的条件已基本成熟了。这是因为：① 经过十几年的研究，汉语自动切分和词性标注的处理技术已达到成熟，为进一步进行句法分析研究打下了很好的基础。② 近几年来，对汉语句法分析方法、依存关系标注、基本句型分析等方面的探索，为进行比较系统全面的句法分析积累了丰富的经验。句法作为汉语语法研究中的重要组成部分，表明了汉语句子基本构造方法和功能，对汉语的正确理解起着重要作用。从这个意义上讲，现代汉语短语分析研究的重要性是不言而喻的。概括地讲，其重要意义主要表现在以下几个方面：

1.使中文信息处理上升到“语言处理”阶段

“语言处理”基本单位是“句”，中文信息处理只有面向“句”的理解才算真正进入“语言处理”阶段。短语分析是从句子的内部结构、层次关系上来分析其内涵，虽然它并不是自足的，即单靠层次和结构关系的分析还不能完全达到了解语义的目的，但它的处理基本单位是句，是句子分析的开始，离开了短语分析就无所谓句子分析。也就是说，短语分析是中文信息处理进入“语言处理”阶段的关键。在整句级句法分析中所遇到的问题，如句法组合层次歧义、语义组合关系歧义等在短语识别的研究中同样存在，所以，汉语短语研究中所采用的技术对于整个句法分析具有方法论上的指导意义和普遍意义。

2.为机器翻译、对外教学等应用的实现奠定基础

文化交流国际化和贸易的全球化要求我们尽快实现机器自动翻译。但由于翻译不仅要求要求知识面广，而且是面向句的理解，目前的中文信息处理水平还都不能满足要求，所以机器翻译还无重大突破。虽然目前市场上出现了很多机器翻译产品，但其基本思路和水平是与语音输入原理基本一致的，信息处理单位仍是“字词”，不能算是真正意义上的翻译。众所周知，翻译有三条准则：“信、达、雅”。为使机器翻译达到理想的境界，我们应朝这三个方向努力。语义层面和语用层面的集成化或一体化分析，是使人译和机器翻译符合“信、达、雅”，则要求译者在修辞、语用等方面下功夫，尽可能反映原文作者的交际意图和行文的风格。唯有如此，才能提高译文的质量，实现翻译准确率高、可读强的目标。当然，这个目标是分层次、按步骤实现的，不可能一蹴而就。要实现翻译的“信、达、雅”，必须实现句子的句法、语义、语用三个层面上的理解，其中，汉语的句法分析是翻译中很重要的一部分，在自然语言处理中起着极其重要的作用。它还是机器翻译、文摘生成和情报检索的基础性课题之一。但是，面对大量真实文本分析的时候，出于汉语句子的复杂性和灵活性等固有的特点，对汉语句子的完全分析无论在时间上还是空间上都受到了极大的挑战。因此，要实现机器自动翻译首先要攻克自动短语分析这一难关。

3．对建立大型的汉语树库具有深远意义

“要想真正提高汉语信息处理的水平，首先必须开发大规

模、高质量的标注多种信息的能够共享的汉语树库。”这是现代汉语句子分析的最终目标，也是自然语言理解的真正实现。一个完整的汉语树库需要三方面的知识：句法结构、语义结构和语用结构，三者缺一不可。而在三者之中，句法结构又是基础，是树库建立的关键。目前，世界各研究机构也都开始这方面研究，如英国的 Lancaster-Leeds 树库项目和美国的 Penn 树库项目、台湾省中研院词库小组（CKIP）所建构的中文结构树项目等，但是，规模与实际需要还有相当大的差距。短语自动分析的顺利完成，对于将汉语语料库的多级加工处理推进到一个新的层次，以及构造大规模的汉语语料等方面，都具有非常重要的理论指导意义。汉语短语自动划分和标注技术的研究为构建大规模的汉语树库提供了强有力的支持。从这些意义上看，现代汉语短语自动识别研究具有很高的理论和实用价值。所以，汉语短语识别是目前自然语言处理领域中的一个较热的课题。

为了降低句法分析的难度，许多语言学家和计算机专家提出了“分而治之”（Divide-Conquer）的思想。即将整体的句法分析分解成若干个层次，可以根据每个层次各自的特点采取相同或者不同的分析策略。通过组块分析（短语分析）的方式，对语料进行简单的处理和分析，减少句法分析的难度，从而减轻翻译的繁重工作。组块分析策略就是这种思想的一种实现方案。通过引入句法块的方法，将问题分为三个阶段：

（1） 块识别：利用基于有限状态分析机制的块识别器识别出句子所有的块。

（2） 块内结构分析：对每个块内的成分赋予合适的句法结构。

（3） 块间关系分析：利用块连接器将各个不同的块构造成完整的句法结构树。

组块分析的这种思想对汉语分析是十分值得借鉴的。它可以为汉语句法分析提供良好的基础，起到很好的中介作用。因为，如果句子的层次分析直接在汉语自动分词和词性标注的基础上进行，那么在后续的分析中可能还要处理各种各样的复杂问题，这使得句法分析难度大大地增加。因而，在汉语句法分析的研究中，我们需要进行汉语短语的划分和标注。

汉语短语的自动标注就是要对一个已经完成了正确切分和词性标注处理的句子，经过自动分析处理，确定不同短语的边界位置，将它们用括号正确地划分出来，并标以合适的短语标记。以下两个例子是分别完成了正确切分与词性标注处理和进行了短语结构标注与短语功能标注的对照句。

例 1：阿基米德/n 在/p 陶罐/n 里/f 放/v 了/u 满满/a 的/u 水/n ，/w 将/p 陶罐/n 放/v 在/p 大/a 盆/n 里/f 。/w

阿基米德/n PP[在/p 陶罐/n 里/f] VP[放/v 了/u] NP[AP[满满/a 的/u]水/n], /w VBA[将/p 陶罐/n VC[放/v PP[在/p NP[大/a 盆/n]里/f]]]。/w

例 2：白杨树/n 不仅/c 象征/v 了/u 北方/f 的/u 农民/n，/w 尤其/d 象征/v 了/u 今天/t 我们/r 民族/n 斗争/n 中/f 所/u 不/d 可/v 缺/v 的/u 质朴/a ，/w 坚强/a ，/w 力求/v 上进