English Corpora and Automated Grammatical Analysis

英语语料库与自动语法分析

ALEX CHENGYU FANG

THE COMMERCIAL PRESS

English Corpora and Automated Grammatical Analysis

Alex Chengyu Fang

THE COMMERCIAL PRESS
2007 · Beijing

图书在版编目(CIP)数据

英语语料库与自动语法分析/方称宇著. —北京：商务印书馆，2007
ISBN 978-7-100-05659-5

Ⅰ.英… Ⅱ.方… Ⅲ.英语—语料库—语法 Ⅳ.H314

中国版本图书馆 CIP 数据核字(2007)第 162671 号

YINGYǓ YǓLIÀOKÙ YǓ ZÌDÒNG YǓFǍ FĒNXĪ

英语语料库与自动语法分析

方称宇 著

商 务 印 书 馆 出 版
(北京王府井大街36号 邮政编码100710)
商 务 印 书 馆 发 行
商务印书馆上海印刷股份有限公司印刷
ISBN 978-7-100-05659-5

2007年11月第1版 开本 787×960 1/16
2007年11月上海第1次印刷 印张 15¼
印数 1-3000册

定价：38.00元

For Lan

Preface

With this book, I would like to commemorate my father, an experimental scientist, who taught me so much about research, practice, and application, not by words but through deeds and demonstrated enthusiasm about novelties. This book could not have been completed without the unreserved support and care from Lan. I am grateful to my two children, Adrien and Elizabeth, for giving me so much joy and pleasure during the course of my research described in this book. I am grateful to all the others in the family, whose enduring love accompanied me throughout the years in London.

Much of the research described in this book was conducted during my association with the Survey of English Usage, University College London, when the International Corpus of English (ICE) was under way to record authentic uses of English worldwide. Pressure to process large quantities of such data for detailed linguistic description prompted me into this research, for which I am deeply indebted to the late Professor Sidney Greenbaum, director of the Survey and coordinator of the ICE project. This is an appropriate place to record my gratitude to him.

I would like to thank Dr Mark Huckvale of the Department of Phonetics and Linguistics, University College London, for his great patience, academic insight, and valuable comments on an earlier draft. His emphasis on practical applications of parsing technologies, especially in the area of linguistically informed speech processing, had an important influence on my research. Different parts of the book have been read by Mr Eric Atwell at Leeds University and Dr Jonathan Ginzburg at King's College London, to whom I would like to express sincere thanks. Of course, responsibility for all the shortcomings remains mine.

Credits must also go to Professor John Campbell of the Computer Science Department, University College London, and Professor Jan Svartvik of the Survey of Spoken English, University of Lund, for many an enlightening after-dinner discussion on various topics in Delancey

Street, Highfield, and Alyth Gardens. Mr Nigel Birch of the Engineering and Physical Sciences Research Council (UK) provided me with timely valuable support of the research.

Finally, I would like to acknowledge the support I have received from the Halliday Centre for Intelligent Applications of Language Studies as well as the Department of Chinese, Translation, and Linguistics of City University of Hong Kong. Not least, thanks go to Professor Housheng Qian of the Commercial Press for his great patience and kind assistance in the publication of this book.

October 2007 *A.C.F.*

前言

从 1993 年到 2005 年，我在伦敦大学学院 (University College London，简称 UCL) 从事科研和教学工作。本书记载了我多年来在语料库语言学和计算语言学这两个领域的主要研究心得和成果。

上世纪 90 年代，是英国语料库语言学发展的黄金时期。伦敦的 Randolph Quirk 教授和 Sidney Greenbaum 教授、兰开斯特的 Geoffrey Leech 教授、伯明翰的 John Sinclair 教授都在进行语料库的开发工作。

当时，Sidney Greenbaum 教授任 UCL 的英语用法调查中心 (Survey of English Usage) 主任，正在从事国际英语语料库 (The International Corpus of English) 的创建工作。100 万字的英国英语语料已经采集完毕，语法标码也已完成，但句法分析遇到不少困难。一是所用的句法分析系统不适用，每输入一个语句，常生成几十、上百、甚至上千棵句法树，然后再人工选取，十分耗时耗力。二是所用的形式语法不适用。当时的语法为英语书面语所写，而 100 万字的英国英语语料包含 60 万字的口语，所以几乎每天都要开会讨论一些语句的具体处理，语法的某些部分干脆需要重写，尤其是不同层次上的并列结构。尽管如此，最后还是有大约 30%的语句，自动句法分析系统根本无法应付。

于是，Sidney Greenbaum 教授和我在 1994 年一同撰写了一份项目申请书，然后约见了英国工程及物理科学研究委员会 (Engineering and Physical Sciences Research Council) 的有关人员，其中包括 Nigel Birch 先生和 Mark Tatham 教授，提出了我们的研究设想。这份申请最后通过了委员会的评审，获得了一笔约 50 万英镑的资助，专门用于研制一个新的自动句法分析系统并重写一部新的、可用于英语口语分析的形式语法。

研究项目的主要思路就是将已经分析过的语料库变成一个句法知识库，从中提取短语结构语法规则，并通过基于实例的手段，在

知识库中为待分析语句提取一棵最佳句法树。这样的句法分析机制涉及几个重要课题：首先需要一个高质量的自动词类标码系统，不仅能对大类进行判别，而且能对小类的细分进行快速、有效的精确分析，比如说动词的配价问题。然后，我们需要一个短语分析系统，将待分析语句处理成一个短语结构集，然后据此计算句法相似度，最终生成相应的句法树。这样一种句法分析途径，具有强劲、高效、精确和自动学习等特性，在对国际英语语料库及其他海量语料库的处理中得到广泛检测和验证。

本书对上述各个部分的研究进行了详细的描述，对系统的实际表现进行了深入的量化评测，并有专门章节来探讨句法分析的评测问题。除此之外，还探讨了介词短语的自动分析，特别是这类短语的句法功能的自动判定，因为这一研究和句法相似度分析有着密切的关系。同时， 本书还就自动语法分析在语音合成及语音识别中的应用做了相应的介绍和说明，希望对读者能有所帮助。

我的不少朋友及同事都看过本书的初稿或部分章节，并提出过许多建议，在此表示感谢，特别是伦敦大学学院的 John Campbell 教授和 Mark Huckvale 博士、伦敦国王学院的 Jonathan Ginzburg 博士、利兹大学的 Eric Atwell 先生、瑞典隆德大学的 Jan Svartvik 教授及商务印书馆上海信息中心主任钱厚生教授。当然，我对书中的所有错误负全责，并恳请读者提出宝贵批评和建议。

最后，我以此书来缅怀先父对我的言传身教和恩师 Sidney Greenbaum 教授对我的栽培，并感谢家人对我的关心和支持。

方称宇

2007 年 10 月

Abstract

This book describes novel research in the robust practical processing of English for detailed word class and syntactic annotations of natural texts. In particular, this book describes an automatic part-of-speech tagger and a robust parser, both of which demonstrate state-of-the-art performance in efficiency, accuracy, and robustness through their application for the annotation of a large corpus of transcribed speech and writing.

The first part of the book is a description of AUTASYS (An Automatic Text Analysis System of English), which combines both a statistical backbone and a set of heuristics. Its novelties include the mapping of the Lancaster-Oslo-Bergen (LOB) tag set, achieved through a bi-gram language model, to a syntactically rich set designed for the International Corpus of English (ICE), successfully achieved even without global syntactic information. This research was extensively tested and evaluated with corpus data. The accuracy was estimated at about 93%.

The book then describes the Survey Parser, which uses AUTASYS as a pre-processor for tokenised input tagged with the ICE tag set. It then produces either a complete parse that indicates both the syntactic category and the syntactic function for each unit of analysis or a partial analysis with syntactic functions or attachment undetermined. A novelty of this work is the use of a large context-free grammar automatically generated from the syntactically annotated ICE corpus. A second novelty is treating parsing as a search problem for a string that has been previously analysed. A major contribution by this work is the use of tag-feature combinations as the anchor of sub-trees for permissive derivations, a more generalised approach than other similarity-based parsing methodologies such as data-oriented parsing that typically make use of lexical anchors for sub-trees. Extensive tests and evaluations of this module were carefully conducted to determine the coverage, the labelling accuracy, and the attachment accuracy of the Survey Parser. Experiments demonstrated a combined accuracy of about 80%, com-

parable to the performance of some of the major systems reported in literature.

A special chapter was devoted to the study of the automatic determination of the syntactic functions of the prepositional phrase. A lexicalist approach was proposed that successfully resolved the attachment of 85% of the prepositional phrases without global parsing. As the final chapter concludes, the successful handling of prepositional phrases is expected to enhance significantly the performance of the Survey Parser. This parser will also benefit from a future study that aims at the automatic detection of clause boundaries.

Contents

List of Figures