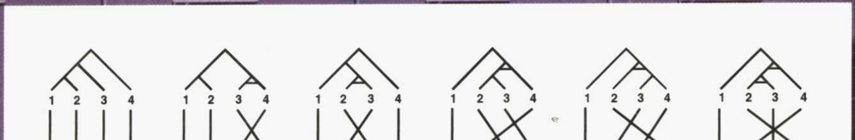
Handbook of Natural Language Processing



edited by

Robert Dale
Hermann Moisl
Harold Somers

Handbook of Natural Language Processing

edited by

Robert Dale
Macquarie University
Sydney, Australia

Hermann Moisl
University of Newcastle
Newcastle-upon-Tyne, England

Harold Somers
UMIST
Manchester, England

MARCEL
DEKKER

MARCEL DEKKER, INC.   NEW YORK • BASEL

ISBN: 0-8247-9000-6

This book is printed on acid-free paper.

Headquarters
Marcel Dekker, Inc.
270 Madison Avenue, New York, NY 10016
tel: 212-696-9000; fax: 212-685-4540

Eastern Hemisphere Distribution
Marcel Dekker AG
Hutgasse 4, Postfach 812, CH-4001 Basel, Switzerland
tel: 41-61-261-8482; fax: 41-61-261-8896

World Wide Web
http://www.dekker.com

Current printing (last digit)
10 9 8 7 6 5 4 3 2 1

PRINTED IN THE UNITED STATES OF AMERICA

Preface

The discipline of Natural Language Processing (NLP) concerns itself with the design and implementation of computational machinery that communicates with humans using natural language. Why is it important to pursue such an endeavor? Given the self-evident observation that humans communicate most easily and effectively with one another using natural language, it follows that, in principle, it is the easiest and most effective way for humans and machines to interact; as technology proliferates around us, that interaction will be increasingly important. At its most ambitious, NLP research aims to design the language input–output components of artificially intelligent systems that are capable of using language as fluently and flexibly as humans do. The robots of science fiction are archetypical: stronger and more intelligent than their creators and having access to vastly greater knowledge, but human in their mastery of language. Even the most ardent exponent of artificial intelligence research would have to admit that the likes of HAL in Kubrick's *2001: A Space Odyssey* remain firmly in the realms of science fiction. Some success, however, has been achieved in less ambitious domains, where the research problems are more precisely definable and therefore more tractable. Machine translation is such a domain, and one that finds ready application in the internationalism of contemporary economic, political, and cultural life. Other successful areas of application are message-understanding systems, which extract useful elements of the propositional content of textual data sources such as newswire reports or banking telexes, and front ends for information systems such as databases, in which queries can be framed and replies given in natural language. This handbook is about the design of these and other sorts of NLP systems. Throughout, the emphasis is on practical tools and techniques for implementable systems; speculative research is minimized, and polemic excluded.

The rest of this preface is in three sections. In Section 1 we provide a sketch of the development of NLP, and of its relationship with allied disciplines such as linguistics and cognitive science. Then, in Section 2, we go on to delineate the scope of the handbook in terms of the range of topics covered. Finally, in Section 3, we provide an overview of how the handbook's content is structured.

1. THE DEVELOPMENT OF NATURAL LANGUAGE PROCESSING

A. History

In the 1930s and 1940s, mathematical logicians formalized the intuitive notion of an effective procedure as a way of determining the class of functions that can be computed algorithmically. A variety of formalisms were proposed—recursive functions, lambda calculus, rewrite systems, automata, artificial neural networks—all equivalent in terms of the functions they can compute. The development of these formalisms was to have profound effects far beyond mathematics. Most importantly, it led to modern computer science and to the computer technology that has transformed our world in so many ways. But it also generated a range of new research disciplines that applied computational ideas to the study and simulation of human cognitive functions: cognitive science, generative linguistics, computational linguistics, artificial intelligence, and natural language processing. Since the 1950s, three main approaches to natural language processing have emerged. We consider them briefly in turn.

- Much of the NLP work in the first 30 years or so of the field's development can be characterized as "NLP based on generative linguistics." Ideas from linguistic theory, particularly relative to syntactic description along the lines proposed by Chomsky and others, were absorbed by researchers in the field, and refashioned in various ways to play a role in working computational systems; in many ways, research in linguistics and the philosophy of language set the agenda for explorations in NLP. Work in syntactic description has always been the most thoroughly detailed and worked-out aspect of linguistic inquiry, so that at this level a great deal has been borrowed by NLP researchers. During the late 1980s and the early 1990s, this led to a convergence of interest in the two communities to the extent that at least some linguistic theorizing is now undertaken with explicit regard to computational issues. Head-driven Phase Structure Grammar and Tree Adjoining Grammar are probably the most visible results of this interaction. Linguistic theoreticians, in general, have paid less attention to the formal treatment of phenomena beyond syntax, and approaches to semantics and pragmatics within NLP have consequently tended to be somewhat more ad hoc, although ideas from formal semantics and speech act theory have found their way into NLP.
- The generative linguistics-based approach to NLP is sometimes constrasted with "empirical" approaches based on statistical and other data-driven analyses of raw data in the form of text corpora, that is, collections of machine-readable text. The empirical approach has been around since NLP began in the early 1950s, when the availability of computer technology made analysis of reasonably large text corpora increasingly viable, but it was soon pushed into the background by the well-known and highly influential opinions of Chomsky and his followers, who were strongly opposed to empirical methods in linguistics. A few researchers resisted the trend, however, and work on corpus collections, such as the Brown and LOB corpora, continued, but it is only in the last 10 years or so that empirical NLP has reemerged as a major alternative to "rationalist" lin-

guistics-based NLP. This resurgence is mainly attributable to the huge data storage capacity and extremely fast access and processing afforded by modern computer technology, together with the ease of generating large text corpora using word processors and optical character readers.

Corpora are primarily used as a source of information about language, and a number of techniques have emerged to enable the analysis of corpus data. Using these techniques, new approaches to traditional problems have also been developed. For example, syntactic analysis can be achieved on the basis of statistical probabilities estimated from a training corpus (as opposed to rules written by a linguist), lexical ambiguities can be resolved by considering the likelihood of one or another interpretation on the basis of context both near and distant, and measures of style can be computed in a more rigorous manner. "Parallel" corpora—equivalent texts in two or more languages—offer a source of contrastive linguistic information about different languages, and can be used, once they have been aligned, to extract bilingual knowledge useful for translation either in the form of traditional lexicons and transfer rules or in a wholly empirical way in example-based approaches to machine translation.

- Artificial neural network (ANN)-based NLP is the most recent of the three approaches. ANNs were proposed as a computational formalism in the early 1940s, and were developed alongside the equivalent automata theory and rewrite systems throughout the 1950s and 1960s. Because of their analogy with the physical structure of biological brains, much of this was strongly oriented toward cognitive modeling and simulation. Development slowed drastically when, in the late 1960s, it was shown that the ANN architectures known at the time were not universal computers, and that there were some very practical and cognitively relevant problems that these architectures could not solve. Throughout the 1970s relatively few researchers persevered with ANNs, but interest in them began to revive in the early 1980s, and for the first time some language-oriented ANN papers appeared. Then, in the mid-1980s, discovery of a way to overcome the previously demonstrated limitations of ANNs proved to be the catalyst for an explosion of interest in ANNs both as an object of study in their own right and as a technology for a range of application areas. One of these application areas has been NLP: since 1986 the volume of NL-oriented—and specifically NLP—research has grown very rapidly.

B. NLP and Allied Disciplines

The development of NLP is intertwined with that of several other language-related disciplines. This subsection specifies how, for the purposes of the handbook, these relate to NLP. This is not done merely as a matter of interest. Each of the disciplines has its own agenda and methodology, and if one fails to keep them distinct, confusion in one's own work readily ensues. Moreover, the literature associated with these various disciplines ranges from very large to huge, and much time can be wasted engaging with issues that are simply irrelevant to NLP. The handbook takes NLP to be exclusively concerned with the design and implementation of effective natural language input and output components for computational systems. On the basis of this definition, we can compare NLP with each of the related disciplines in the following ways.

1. Cognitive Science

Cognitive science is concerned with the development of psychological theories; the human language faculty has always been central to cognitive theorizing and is, in fact, widely seen as paradigmatic for cognition generally. In other words, cognitive science aims in a scientific sense to explain human language. NLP as understood by this book does not attempt such an explanation. It is interested in designing devices that implement some linguistically relevant mapping. No claims about cognitive explanation or plausibility are made for these devices. What matters is whether they actually implement the desired mapping, and how efficiently they do so.

2. Generative Linguistics

Like cognitive science, generative linguistics aims to develop scientific theories, although about human language in particular. For Chomsky and his adherents, linguistics is, in fact, part of cognitive science; the implication of this for present purposes has already been stated. Other schools of generative linguistics make no cognitive claims, and regard their theories as formal systems, the adequacy of which is measured in terms of the completeness, internal consistency, and economy by which all scientific theories are judged. The rigor inherent in these formal systems means that, in many cases, they characterize facts and hypotheses about language in such a way that these characterizations can fairly straightforwardly be used in computational processing. Such transfers from the theoretical domain to the context of NLP applications are quite widespread. At the same time, however, there is always scope for tension between the theoretical linguist's desire for a maximally economical and expressive formal system and the NLP system designer's desire for broad coverage, robustness, and efficiency. It is this tension that has in part provoked interest in methods other than the strictly symbolic.

3. Artificial Intelligence

Research in artificial intelligence (AI) aims to design computational systems that simulate aspects of cognitive behavior. It differs from cognitive science in that no claim to cognitive explanation or plausibility is necessarily made with respect to the architectures and algorithms used. AI is thus close in orientation to NLP, but not identical: in the view we have chosen to adopt in this book, NLP aims not to simulate intelligent behavior per se, but simply to design NL input–output components for arbitrary computational applications; these might be components of AI systems, but need not be.

4. Computational Linguistics

Computational linguistics (CL) is a term that different researchers interpret in different ways; for many the term is synonymous with NLP. Historically founded as a discipline on the back of research into Machine Translation, and attracting researchers from a variety of neighboring fields, it can be seen as a branch of linguistics, computer science, or literary studies.

- As a branch of linguistics, CL is concerned with the computational implementation of linguistic theory: the computer is seen as a device for testing the completeness, internal consistency, and economy of generative linguistic theories.
- As a branch of computer science, CL is concerned with the relationship between natural and formal languages: interest here focuses on such issues as language recognition and parsing, data structures, and the relationship between facts about language and procedures that make use of those facts.
- As a branch of literary studies, CL involves the use of computers to process large corpora of literary data in, for example, author attribution of anonymous texts.

Within whatever interpretation of CL one adopts, however, there is a clear demarcation between domain-specific theory on the one hand and practical development of computational language processing systems on the other. It is this latter aspect—practical development of computational language processing systems—that, for the purposes of this book, we take to be central to NLP. In recent years, this approach has often been referred to as "language technology" or "language engineering."

Our view of NLP can, then, be seen as the least ambitious in a hierarchy of disciplines concerned with NL. It does not aim to explain or even to simulate the human language faculty. What it offers, however, is a variety of practical tools for the design of input–output modules in applications that can benefit from the use of natural language. And, because this handbook is aimed at language-engineering professionals, that is exactly what is required.

2. THE SCOPE OF THIS BOOK

Taken at face value, the expression "natural language processing" encompasses a great deal. Ultimately, applications and technologies such as word processing, desktop publishing, and hypertext authoring are all intimately involved with the handling of natural language, and so in a broad view could be seen as part of the remit of this book. In fact, it is our belief that, in the longer term, researchers in the field will increasingly look to integration of these technologies into mainstream NLP research. But there are limits to what can usefully be packed between the covers of a book, even of a large book such as this, and there is much else to be covered, so a degree of selection is inevitable. Our particular selection of topics was motivated by the following considerations.

Any computational system that uses natural language input and output can be seen in terms of a five-step processing sequence (Fig. 1):

1. The system receives a physical signal from the external world, where that signal encodes some linguistic behavior. Examples of such signals are speech waveforms, bitmaps produced by an optical character recognition or handwriting recognition system, and ASCII text streams received from an electronic source. This signal is converted by a transducer into a linguistic representation amenable to computational manipulation.
2. The natural language analysis component takes the linguistic representation from step 1 as input and transforms it into an internal representation appropriate to the application in question.

3. The application takes the output from step 2 as one of its inputs, carries out a computation, and outputs an internal representation of the result.
4. The natural language generation component takes part of the output from step 3 as input, transforms it into a representation of a linguistic expression, and outputs that representation.
5. A transducer takes the output representation from step 4 and transforms it into a physical signal that, in the external world, is interpretable by humans as language. This physical signal might be a text stream, the glyphs in a printed document, or synthesized speech.

Ultimately, the input to a language-processing system will be in graphemic or phonetic form, as will the outputs. Phonetic form corresponds to speech input. Traditionally, however, the only kind of graphemic input dealt with in language-processing systems is digitally encoded text. In the case of such graphemic input it is generally assumed either that this will be the native form of the input (an assumption that is indeed true for a great deal of the data we might wish to process) or that some independent process will map the native form into digitally encoded text. This assumption is also made in much—although certainly not all—work on speech: a great deal of this research proceeds on the basis that mapping speech signals into a textual representation consisting of sequences of words can be achieved independently of any process that subsequently derives some other representation from this text. It is for these reasons that we have chosen in the present work to focus on the processing techniques that can be used with textual representations, and thus exclude from direct consideration techniques that have been developed for speech processing.

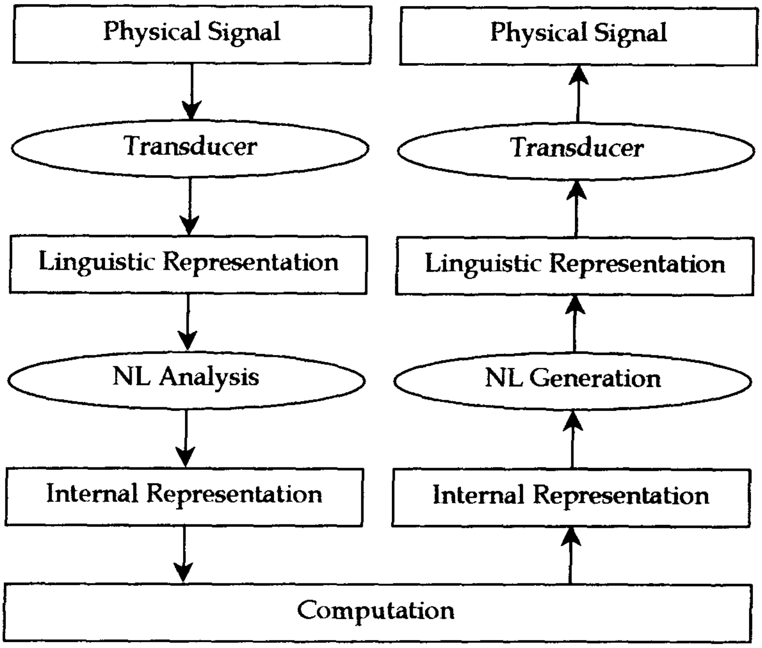


Fig. 1 The five steps in computational processing of natural language input and output.

Subject, then, to the foregoing restrictions of scope, coverage of NLP tools and applications in this handbook attempts to strike a balance between comprehensiveness and depth of coverage with the aim of providing a practical guide for practical applications.

3. THE STRUCTURE OF THIS BOOK

After some five decades of NLP research, and of NLP-related work in the aforementioned disciplines, a huge research literature has accumulated. Because the primary function of this handbook is to present that material as accessibly as possible, clarity of organization is paramount. After much discussion, we realized that no organization was going to be ideal, but two reasonable possibilities did emerge: a historically based organization and a topic-based one. The first of these would partition the material into three sections corresponding to the three main strands of development described in the foregoing historical sketch, and the second would attempt to identify a set of research topics that would comprehensively describe NLP and NLP-related research to date by presenting the relevant work in each topic area. There are drawbacks to both. The main problem with the historically based organization is that it might be seen to reinforce the distinctions between the three approaches to NLP: although these distinctions have historical validity, they are increasingly being broken down, and for the handbook to partition the material in this way would distort the current state of increasingly ecumenical NLP research. The main problem with topic-based organization was identification of a suitable set of research areas. Demarcations that seemed natural from one point of view seemed less so from another, and attempts to force all the material into canonical compromise categories simply distorted it, with consequent effects on clarity. A morphology–syntax–semantics categorization, for example, seems perfectly natural from the point of view of linguistics-based NLP, but far less so from an empirical or ANN-based one. There are, moreover, important topics unique to each of the main strands of historical development, such as representation in the ANN-based approach. In fact, the more closely one looks, the more problematical identification of an adequate set of topics becomes.

In the end, the historically based organization was adopted. The motivation for this was the need for clarity: each of the three approaches to NLP can be described in its own self-contained section without the distortion that a topic-based organization would often require. To mitigate the problem of insularity, each section indicates areas of overlap with the others. Inevitably, though, the problem has not thereby been eliminated.

The handbook is, then, organized as follows. There are three main parts corresponding to the three main strands of historical development:

1. Symbolic approaches to NLP, which focuses on NLP techniques and applications that have their origins in generative linguistics
2. NLP based on empirical corpus analysis
3. NLP based on artificial neural networks

Again for clarity, each section is subdivided into the following subsections:

1. An introduction that gives an overview of the approach in question
2. A set of chapters describing the fundamental concepts and tools appropriate to that approach
3. A set of chapters describing particular applications in which the approach has been successfully used

Beyond this basic format, no attempt has been made to impose uniformity on the individual parts of the handbook. Each is the responsibility of a different editor, and each editor has presented his material in a way that he considers most appropriate. In particular, given that work in ANN–based NLP will be relatively new to many readers of this book, the introduction to Part III includes a detailed overview of how these techniques have developed and found a place in NLP research.

Some acknowledgments are appropriate. First and foremost, we express our thanks to the many contributors for their excellent efforts in putting together what we hope will be a collection of long-standing usefulness. We would also like to thank the editorial staff of Marcel Dekker, Inc., for their considerable patience in response to the succession of delays that is inevitable in the marshaling of a book of this size and authorial complexity. Finally, Debbie Whittington of the Microsoft Research Institute at Macquarie University and Rowena Bryson of the Centre for Research in Linguistics at Newcastle University deserve special mention for their administrative help.

Robert Dale
Hermann Moisl
Harold Somers

Contributors

James Allen Department of Computer Science, University of Rochester, Rochester, New York

Elisabeth André German Research Center for Artificial Intelligence, Saarbrücken, Germany

Ion Androutsopoulos Institute of Informatics and Telecommunications, National Centre for Scientific Research "Demokritos," Athens, Greece

Eric Brill Microsoft Research, Redmond, Washington

John A. Carroll School of Cognitive and Computing Sciences, University of Sussex, Brighton, England

Chun-Hsien Chen Department of Information Management, Chang Gung University, Kwei-Shan, Tao-Yuan, Taiwan, Republic of China

Jim Cowie Computing Research Laboratory, New Mexico State University, Las Cruces, New Mexico

Ido Dagan Department of Mathematics and Computer Science, Bar-Ilan University, Ramat Gan, Israel

Robert Dale Department of Computing, Macquarie University, Sydney, Australia

Georg Dorffner Austrian Research Institute for Artificial Intelligence, and Department of Medical Cybernetics and Artificial Intelligence, University of Vienna, Vienna, Austria

George Ferguson Department of Computer Science, University of Rochester, Rochester, New York

Barbara J. Grosz Division of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts

Simon Haykin Communications Research Laboratory, McMaster University, Hamilton, Ontario, Canada

George E. Heidorn Microsoft Research, Redmond, Washington

Stefan Heil Technical University of Munich, Munich, Germany

Vasant Honavar Department of Computer Science, Iowa State University, Ames, Iowa

Richard I. Kittredge* Department of Linguistics and Translation, University of Montreal, Montreal, Quebec, Canada

Ludovic Lebart Economics and Social Sciences Department, Ecole Nationale Supérieure des Télécommunications, Paris, France

Karen E. Lochbaum U S WEST Advanced Technologies, Boulder, Colorado

Simon Lucas Department of Electronic Systems Engineering, University of Essex, Colchester, England

Yuji Matsumoto Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Nara, Japan

David D. McDonald Brandeis University, Arlington, Massachusetts

Tony McEnery Department of Linguistics, Bowland College, Lancaster University, Lancaster, England

Kathleen R. McKeown Department of Computer Science, Columbia University, New York, New York

Dieter Merkl Department of Software Technology, Vienna University of Technology, Vienna, Austria

Risto Miikkulainen Department of Computer Sciences, The University of Texas at Austin, Austin, Texas

Bradford W. Miller Cycorp, Austin, Texas

Hermann Moisl Centre for Research in Linguistics, University of Newcastle, Newcastle-upon-Tyne, England

**Also affiliated with*: CoGenTex, Inc., Ithaca, New York.

Michael Oakes Department of Linguistics, Lancaster University, Lancaster, England

David D. Palmer The MITRE Corporation, Bedford, Massachusetts

Rajesh G. Parekh Allstate Research and Planning Center, Menlo Park, California

Massimo Poesio Human Communication Research Centre and Division of Informatics, University of Edinburgh, Edinburgh, Scotland

Alain Polguère Department of Linguistics and Translation, University of Montreal, Montreal, Quebec, Canada

Dragomir R. Radev School of Information, University of Michigan, Ann Arbor, Michigan

Martin Rajman Artificial Intelligence Laboratory (LIA), EPFL, Swiss Federal Institute of Technology, Lausanne, Switzerland

Eric K. Ringger* Department of Computer Science, University of Rochester, Rochester, New York

Graeme Ritchie Division of Informatics, University of Edinburgh, Edinburgh, Scotland

Christer Samuelsson Xerox Research Centre Europe, Grenoble, France

Jürgen Schmidhuber IDSIA, Lugano, Switzerland

Candace L. Sidner Lotus Research, Lotus Development Corporation, Cambridge, Massachusetts

Teresa Sikorski Zollo Department of Computer Science, University of Rochester, Rochester, New York

Harold Somers Centre for Computational Linguistics, UMIST, Manchester, England

Richard Sproat Department of Human/Computer Interaction Research, AT&T Labs–Research, Florham Park, New Jersey

Henry S. Thompson Division of Informatics, University of Edinburgh, Edinburgh, Scotland

**Current affiliation*: Microsoft Research, Redmond, Washington

Takehito Utsuro Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Nara, Japan

Stefan Wermter Centre for Informatics, SCET, University of Sunderland, Sunderland, England

Yorick Wilks Department of Computer Science, University of Sheffield, Sheffield, England

Mats Wirén Telia Research, Stockholm, Sweden

Michael Witbrock Lycos Inc., Waltham, Massachusetts

Dekai Wu Department of Computer Science, The Hong Kong University of Science and Technology, Kowloon, Hong Kong

David Yarowsky Department of Computer Science, Johns Hopkins University, Baltimore, Maryland

Contents

Part II Empirical Approaches to NLP