

Lecture Notes in Artificial Intelligence

1171

Subseries of Lecture Notes in Computer Science

Edited by J. G. Carbonell and J. Siekmann

Lecture Notes in Computer Science

Edited by G. Goos, J. Hartmanis and J. van Leeuwen

Alexander Franz

Automatic Ambiguity Resolution in Natural Language Processing

An Empirical Approach



Springer

Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Author

Alexander Franz

Sony Corporation, D-21 Laboratory

6-7-35 Kitashinagawa, Shinigawa-Ku, Tokyo 141, Japan

E-mail: amf@pdp.crl.sony.co.jp

Cataloging-in-Publication Data applied for

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Franz, Alexander:

Automatic ambiguity resolution in natural language processing :
an empirical approach / Alexander Franz. - Berlin ; Heidelberg
; New York ; Barcelona ; Budapest ; Hong Kong ; London ;
Milan ; Paris ; Santa Clara ; Singapore ; Tokyo : Springer, 1996
(Lecture notes in computer science ; Vol. 1171 : Lecture notes in
artificial intelligence)

ISBN 3-540-62004-4

NE: GT

CR Subject Classification (1991): I.2.7, I.2, I.6.5, G.3, F.4.2-3

ISBN 3-540-62004-4 Springer-Verlag Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag. Violations are liable for prosecution under the German Copyright Law.

© Springer-Verlag Berlin Heidelberg 1996
Printed in Germany

Typesetting: Camera ready by author

SPIN 10549137 06/3142 - 5 4 3 2 1 0 Printed on acid-free paper

Foreword

Natural language processing is often called an “AI-complete” task, in the sense that in order to truly process language (i.e. to comprehend, to translate, to generate) full understanding is required, which is itself the ultimate goal of Artificial Intelligence. For those who seek solutions to practical problems, this is not a desirable property of NLP. However, it is possible to address reduced versions of the NLP problem without the prerequisite of having first solved all of the other arbitrarily-difficult AI problems. There are various ways to restrict the NLP problem: restrict the semantic domain, restrict the expressiveness of the syntax, focus on only one aspect of NLP at a time (e.g. phoneme recognition, Part-of-Speech tagging, morphological analysis), seek only approximate solutions (e.g. by replacing a complex cognitive model with a statistical component), and so on. The work described in this monograph pursues the latter two approaches with significant success.

The beauty of statistical techniques for NLP is that in principle they require only training data – not manual reprogramming – to solve new or extended versions of the same problem. For instance, a Part-of-Speech tagger should be as easily trainable for any subset of English (e.g. legal, medical, engineering texts) as for the original subset in which it was developed. Moreover, it should be applicable to other languages as well, after modifying the tagset and possibly the feature set. The drawbacks of statistical systems, however, are also significant. It is difficult to solve the more complex NLP problems statistically with acceptable accuracy. It is difficult to obtain enough training data for models with large feature sets. It is a significant challenge to create computationally-tractable models that cope with significant combinations of features. And, it is seldom clear a priori how to design the feature set or what statistical model to use. All these difficulties notwithstanding, significant progress has been made in statistical methods for speech recognition, Part-of-Speech tagging, lexical disambiguation, Prepositional Phrase (PP) attachment, and even end-to-end machine translation.

Dr. Franz’s contribution is to develop a statistical paradigm for NLP tasks that makes minimal restrictive a priori assumptions. Based on loglinear modeling with contingency tables, the key idea is to be able to explore models that consider features singly, in pairs, or in larger interacting subsets, rather than in a single pre-determined and often suboptimal manner. Of course,

VI Foreword

this approach requires careful selection of potentially meaningful features, as well as certain simplifying assumptions – such as feature partitioning – to achieve computational tractability. The results on Part-of-Speech tagging and multiple-PP attachment structural disambiguation show the advances of this modeling approach over the previous state of the art. Of course much more remains to be investigated with respect to statistical NLP and hybrid rule-based/statistical approaches, but the methodology of the research and clear initial advances have been established.

September 1996

Jaime Carbonell

Preface

This is an exciting time for Artificial Intelligence (AI), and for Natural Language Processing (NLP) in particular. Within the last five years or so, a newly revived spirit has gained prominence that promises to revitalize our field: the spirit of empiricism.

As described by Cohen (1995), the revival of empiricism can be felt throughout all of AI. For NLP, empiricism offers a new orientation and a new way of looking at problems involving natural language that focuses on naturally-occurring language data.

There are three main aspects of the empirical approach to NLP. The first aspect concerns the *exploration* of the natural language phenomenon under study. Initial, pre-theoretical observations are analyzed and structured with respect to “features” or statistical variables. The data is examined for trends, and initial ideas about causal influences and interactions are formed.

The growing availability of online text and speech corpora has made it possible to perform such exploratory data analysis on natural language data. This enterprise has just begun, and much remains to be learned. Nevertheless, I expect that this type of activity will in time come to be widely accepted as an essential component of NLP methodology.

The second aspect of the empirical method is related to *model construction*. Currently, many models in empirical NLP are statistical models of the simplest type, implicitly assuming one of the common statistical distributions and estimating parameters directly from the observed training data.

This is mostly a reflection of the youth of the empirical NLP enterprise. After collecting, exploring, and structuring data, fitting a standard statistical model is the most obvious next step. In the future, I expect that the models will become more complex, combining both symbolic and statistical elements. This is likely to develop into a major research focus.

The third aspect of the empirical approach is probably the most familiar. It relates to *formal experiments*, statistical hypothesis testing, and the rejection or confirmation of scientific hypotheses. In the so-called hard sciences, this has long been a part of the standard methodology.

Not so in AI. Within NLP, even though formal hypothesis testing remains quite rare, this aspect of empiricism has already lead to a widespread concern with quantitative evaluation. At the current state of the art, the main concern

usually lies with measuring the accuracy of a model at performing a specified task, such as recognizing a spoken word or determining the syntactic structure of a sentence.

If standardized data collections are used, then the accuracies obtained by different models can be compared directly, and conclusions about the fidelity of the different models can be drawn. This is currently not always the case, however; it is often difficult to interpret the reported accuracy measurements. As the field develops, I expect that there will be somewhat less of an emphasis on competition between different implemented systems, and a growing emphasis on drawing general conclusions about language processing.

In this book, we demonstrate the empirical approach to NLP by tackling one of the main problems in natural language analysis, the problem of automatic ambiguity resolution. Using data from the University of Pennsylvania Treebank, we investigate three particularly problematic types of syntactic ambiguity in English: unknown words, lexical Part-of-Speech ambiguity, and Prepositional Phrase attachment ambiguity.

It has often been suggested that effective ambiguity resolution requires the integration of multiple sources of knowledge. In this work, we will show how to construct procedures for automatic ambiguity resolution that achieve this aim in a precisely defined sense: By adopting the loglinear class of statistical models, we are able to take into account the interactions between different features, and thus obtain a Bayesian posterior probability distribution over the response variable that is properly conditioned on the combinations of the explanatory variables.

Our scientific result pertaining to the theory of natural language ambiguity can be summarized in one sentence: Ambiguity resolution procedures that take into account the interactions between analysis features obtain higher disambiguation accuracy than procedures that assume independence. This result is derived through a series of experiments that provide a rigorous evaluation of our models, and a thorough comparison with methods that have been described previously in the literature.

While this result does not *prove* that handling feature interactions is necessary, it certainly provides a strong indication. In doing so, this work suggests a number of avenues for further research on the theory of ambiguity resolution. At the same time, the techniques described here yield higher disambiguation accuracy than previously described methods, so they are directly useful for applied work on natural language analysis. More broadly, the methods for data analysis, modeling, and experimental evaluation that are described in this book are relevant to anyone working in NLP or AI.

This book is based on my PhD dissertation submitted to the Computational Linguistics Program at Carnegie Mellon University in 1995. I am deeply indebted to my advisor, Jaime Carbonell, for his continuous help, advice, and support. I am also grateful to the other members of my thesis committee, Ted Gibson, Michael “Fuzzy” Mauldin, and Teddy Seidenfeld, for

their guidance and encouragement. I would like to thank my fellow Computational Linguistics graduate students; the members of the Computational Linguistics community in Pittsburgh; my friends and colleagues at the Center for Machine Translation and at Carnegie Group Inc.; Gerald Gazdar, who fostered my first interests in natural language; and the Sony research members. Finally, I wish to thank Keiko Horiguchi for making life wonderful.

Tokyo, September 1996

Alexander Franz

Table of Contents

1. Introduction	1
1.1 Natural Language Ambiguity	2
1.2 Ambiguity and Robust Parsing	3
1.2.1 Grammatical Coverage	4
1.2.2 Ambiguity Resolution Schemes	5
1.3 Corpus-Based Approaches to NLP	6
1.3.1 Empirical Orientation	6
1.3.2 Naturally-Occurring Language	7
1.3.3 Emphasis on Evaluation	7
1.4 Statistical Modeling for Ambiguity Resolution	8
1.5 Overview of this Book	8
2. Previous Work on Syntactic Ambiguity Resolution	11
2.1 A Note on Reported Error Rates	11
2.2 The Problem of Unknown Words	11
2.2.1 AI Approaches to Unknown Words	12
2.2.2 Morphological Analysis of Unknown Words	12
2.2.3 Corpus-Based Approaches to Unknown Words	13
2.3 Lexical Syntactic Ambiguity	13
2.3.1 Rule-Based Lexical Syntactic Ambiguity Resolution	13
2.3.2 Frequency-Based POS Tagging	14
2.3.3 Hidden Markov Models for Lexical Disambiguation	16
2.3.4 N-Gram Based Stochastic POS Tagging	17
2.4 Structural Ambiguity	19
2.4.1 Syntactic Approaches	19
2.4.2 Semantic Approaches	20
2.4.3 Pragmatic Approaches	22
2.5 Prepositional Phrase Attachment Disambiguation	23
2.5.1 Using Lexical Associations for PP Attachment Disambiguation	27
2.5.2 Systematic Ambiguity in PP Attachment	28
2.5.3 PP Attachment and Class-Based Generalization	28
2.5.4 A Maximum Entropy Model of PP Attachment	29
2.5.5 Learning Symbolic PP Attachment Rules	30

XII Table of Contents

2.6	Critique of Previous Approaches.....	30
2.6.1	Syntactic Approaches	31
2.6.2	Semantic and Pragmatic Approaches.....	31
2.6.3	Corpus-Based Approaches	32
3.	Loglinear Models for Ambiguity Resolution.....	35
3.1	Requirements for Effective Ambiguity Resolution	35
3.1.1	Automatic Training.....	35
3.1.2	Handling Multiple Features	35
3.1.3	Modeling Feature Dependencies	36
3.1.4	Robustness	36
3.2	Ambiguity Resolution as a Classification Problem	36
3.2.1	Making Decisions under Uncertainty	36
3.2.2	Statistical Classification	37
3.2.3	Expected Loss and the Zero-One Loss Function	38
3.2.4	Minimum Error Rate Classification	38
3.2.5	Maximizing Utility	39
3.3	The Loglinear Model.....	40
3.3.1	Categorical Data Analysis	40
3.3.2	The Contingency Table	40
3.3.3	The Importance of Smoothing	41
3.3.4	Individual Cells	42
3.3.5	Marginal Totals and Expected Cell Counts	42
3.3.6	Interdependent Variables and Interaction Terms.....	43
3.3.7	The Iterative Estimation Procedure.....	44
3.3.8	Example of Iterative Estimation.....	45
3.3.9	Definition of a Loglinear Model	46
3.4	Statistical Inference.....	47
3.4.1	The Bayesian Approach	47
3.4.2	Bayesian Inference Using the Contingency Table.....	47
3.5	Exploratory Data Analysis.....	48
3.5.1	The Exploratory Nature of this Approach	48
3.5.2	Searching for Discriminators	49
4.	Modeling New Words.....	51
4.1	Experimental Data and Procedure.....	51
4.1.1	Problem Statement	51
4.1.2	Experimental Data	51
4.1.3	Modeling Procedure	53
4.2	Exploring the Data	53
4.2.1	The Initial Feature Set	53
4.2.2	Decreasing the Size of the Model	54
4.2.3	Eliminating Low-Information Features	55
4.2.4	Choosing the Interaction Terms	57
4.3	Evaluation and Experimental Results	59

4.3.1	Measuring Residual Ambiguity: An Example	60
4.3.2	Results from Previous Work	62
4.3.3	Constructing the Loglinear Model	63
4.3.4	Experimental Results	66
4.3.5	Effect of Number of Features on Performance	67
4.3.6	Number of Features and Residual Ambiguity	68
4.3.7	The Tradeoff between Accuracy and Ambiguity	69
5.	Part-of-Speech Ambiguity	71
5.1	Stochastic Part-of-Speech Tagging	71
5.1.1	Maximizing Tag Sequence Probability	71
5.1.2	Making Independence Assumptions	72
5.1.3	The Tagging Algorithm	73
5.2	Estimation of Probabilities	74
5.2.1	Tagged versus Untagged Training Corpora	74
5.2.2	Jeffreys' Estimate	74
5.2.3	Linear Interpolation	75
5.2.4	Deleted Interpolation	76
5.2.5	Other Smoothing Schemes	77
5.3	Stochastic Tagging with Unknown Words	77
5.3.1	Experimental Data	78
5.3.2	Results from Previous Work	78
5.3.3	Boxplots	78
5.3.4	Error Distribution	79
5.3.5	Tagging Error Density	81
5.3.6	Normal Probability Plot	82
5.3.7	The Role of Contextual Probabilities	83
5.3.8	Bigrams versus Trigrams	84
5.3.9	The Importance of Smoothing Trigrams	85
5.3.10	Lexical Probabilities and Unknown Words	85
5.3.11	POS Tagging and Unknown Words	87
5.3.12	Unknown Word Model Results	88
5.3.13	Effect of Statistical Model on New Text	88
5.4	Errors Analysis for the Trigram-Based Tagger	89
5.4.1	Qualitative Error Analysis	90
5.4.2	Quantitative Error Analysis	91
5.4.3	Overall Error Distribution of the Stochastic Tagger	91
5.4.4	Confusion Matrix of the Stochastic Tagger	91
5.4.5	Results of Error Analysis	92
5.5	Using a Loglinear Model for Lexical Disambiguation	93
5.5.1	Errors before Correction	93
5.5.2	Features for Tagging Correction	94
5.5.3	Results of Tagging Correction	95
5.5.4	Summary of Results	95

6. Propositional Phrase Attachment Disambiguation	97
6.1 Overview of PP Experiments	97
6.2 Features for PP Attachment	97
6.3 Experimental Data and Evaluation	99
6.4 Experimental Results: Two Attachments Sites	100
6.4.1 Baseline: Right Association	100
6.4.2 Results of Lexical Association	101
6.4.3 Results of the Loglinear Model	102
6.5 Experimental Results: Three Attachment Sites	102
6.5.1 Additional PP Patterns	102
6.5.2 Baseline: Right Association	103
6.5.3 Results of Lexical Association	103
6.5.4 Results of Enhanced Lexical Association	104
6.5.5 Results of the Loglinear Model	104
6.5.6 Analysis of Results	105
6.6 Human Performance on PP Attachment	107
6.7 Discussion of PP Experiments	107
7. Conclusions	109
7.1 Summary of this Work	109
7.1.1 Modeling Unknown Words	109
7.1.2 Part-of-Speech Disambiguation	109
7.1.3 Propositional Phrase Attachment Disambiguation	110
7.2 Contributions of this Work	110
7.2.1 Automatic Natural Language Ambiguity Resolution	110
7.2.2 Statistical Language Modeling	111
7.2.3 Towards a Theory of Ambiguity	111
7.3 Future Work	112
7.3.1 Improving the Models	113
7.3.2 Application to Other Ambiguity Problems	113
7.3.3 Integration with Other Knowledge Sources	114
7.3.4 Costs and Benefits of Loglinear Ambiguity Resolution Models	115
7.4 Towards a Unified Model	115
A. Entropy	117
B. Penn Treebank Tags	119
C. Obtaining Random Samples	121
D. Confusion Matrices for POS Tagging	123
E. Converting Treebank Files	125
F. Input to and Output from the Estimation Routines	129