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

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,

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