An investigation into compositional features and feature merging for maximum entropy-based parse selection
Mullen, Anthony James

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2002

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
An Investigation into Compositional Features and Feature Merging for Maximum Entropy-Based Parse Selection

Tony Mullen
The work in this thesis has been carried out under the auspices of the Behavioral and Cognitive Neurosciences (BCN) research school, Groningen, and the Dutch Graduate School in Logic OzsL (Onderzoeksschool Logica), Amsterdam.

Groningen Dissertations in Linguistics 37
ISSN 0928-0030
Typeset in \LaTeX
RIJKSUNIVERSITEIT GRONINGEN

An Investigation into Compositional Features and Feature Merging for Maximum Entropy-Based Parse Selection

Proefschrift


door

Anthony James Mullen

tegen op 15 mei 1971 te Fairbanks, Alaska (Verenigde Staten)
Promotor : Prof. dr. ir. John Nerbonne

Referenten : Dr. Robert Malouf
            Dr. Miles Osborne

Beoordelingscommissie : Prof. dr. Chris Brew
                         Prof. dr. Eric Brill
                         Prof. dr. Lambert Schomaker
Acknowledgements

This thesis would not have been possible without the help of many people. In particular I would like to thank John Nerbonne for his supervision as my promotor. His experience and guidance helped to ensure that the work was done to the best of my ability. Others whose direct input was crucial throughout the research and writing of this thesis were my “referents” Miles Osborne and Rob Malouf. I thank Miles for putting me onto the topic in the first place and giving me much valuable guidance in the field, as well as giving me the opportunity to spend a wonderful academic year in Edinburgh. Rob gave me no end of technical advice and I thank him for his patient explanations and helpful collaborations. To the list of colleagues deserving of gratitude I would like to add Gertjan van Noord, to whom I am grateful for his patience and assistance with my various questions and problems throughout my time in Groningen. Many other people in the department of Alfa-Informatica also gave me help in many ways, and I would like to acknowledge Gosse Bouma, Mark-Jan Nederhof and Peter Kleiweg for all their help. In addition, I am grateful to Shoji Yoshikawa, Rob Visser, and the fourth floor secretarial staff for their assistance with the day-to-day issues of computers, money, and miscellaneous, respectively.

Among those less directly related to this work itself, I am grateful to Christine Appel not only for her friendship and support throughout the past few years but for her professional collaboration, which allowed me to expand my own research interests into a field I would not have been qualified to enter without her. Rob Koeling, Stasinos Konstantopolous, and Ivelyn Stoianov deserve mention not only for their friendship but for their help and advice in a variety of technical areas, from maxent modeling to Linux installation to lamp repair. I would also like to thank Carl Vogel, who helped me to get the position in Groningen in the first place, and once I was here, helped me to minimize the number of embarrassing questions I had to ask right off the bat. Stasinos Konstantopolous and Begoña Villada Moirón deserve special mention for their help as my “paranimfs.”

There are many other people whose friendship and company helped get me through the sometimes tedious four years of grad school. I would like to single out Elissaveta Katchouleva to thank for all of her support and encouragement, and, as always, I am grateful to my family, whose love and support has helped me in all that I’ve done.
Contents

1 Introduction ............................................. 1
  1.1 The emergent field of NLP .......................... 1
    1.1.1 Linguists versus statisticians? ............... 3
    1.1.2 Interdisciplinary approaches to NLP .......... 3
  1.2 The problem of ambiguity ............................ 4
    1.2.1 Statistical parsing ............................ 7
    1.2.2 Maximum entropy and NLP ..................... 8
  1.3 Where does this work stand? ....................... 8
  1.4 The structure of this thesis ...................... 9

2 The task of parse selection .......................... 11
  2.1 Syntax and structure in language .................. 11
    2.1.1 Dependency structures ......................... 12
  2.2 The basics of parsing ................................ 13
    2.2.1 Earley parsing ................................ 13
  2.3 Ambiguity and large grammars ..................... 16
    2.3.1 Statistical parsing and “parse selection” .... 18
  2.4 Performance evaluation .............................. 18
  2.5 Approaches to statistical grammars ............... 19
    2.5.1 PCFGs ........................................ 20
    2.5.2 Lexicalized parsing ............................ 21
    2.5.3 Attribute-value grammars and maximum entropy-based parsing .... 25
    2.5.4 Other approaches ................................ 26
  2.6 Commonalities in statistical approaches .......... 26

3 The maximum entropy technique ...................... 29
  3.1 The maximum entropy principle .................... 29
  3.2 Basic definitions in probability and information theory .... 30
    3.2.1 Expected value ................................ 31
    3.2.2 Entropy ........................................ 31
    3.2.3 Cross entropy .................................. 33
    3.2.4 KL divergence .................................. 34
3.2.5 Log-likelihood ........................................... 34
3.3 Features and maxent ........................................ 35
3.4 Parametric form .............................................. 36
3.5 Maximum Entropy Modeling ................................. 41
3.6 Improved iterative scaling .................................. 41
  3.6.1 The IIS Algorithm ...................................... 42
3.7 Noise and overfitting ....................................... 43
  3.7.1 Maximum entropy and statistical dependencies ...... 44
3.8 Feature Selection ............................................ 44
3.9 Implementation ................................................ 45

4 Log-linear approaches ......................................... 47
  4.1 History-based parsing of WSJ text with maxent models . 47
    4.1.1 Maxent as a classifier ................................ 48
    4.1.2 A simple classification example ....................... 48
    4.1.3 Parsing as classification .............................. 49
  4.2 Stochastic modeling of attribute-value grammars ........ 50
    4.2.1 Why independence assumptions fail .................. 50
    4.2.2 The log-linear solution ............................... 51
    4.2.3 Abney .................................................. 51
    4.2.4 Johnson ................................................. 52
    4.2.5 Osborne ................................................ 53
  4.3 Salient points ............................................... 53

5 Maxent parsing with compositional features ................ 55
  5.1 Parsing with maximum entropy .............................. 55
  5.2 Representing parses as events ............................. 56
    5.2.1 A disambiguation example ............................ 59
    5.2.2 What it means ......................................... 61
  5.3 The training data .......................................... 61
  5.4 Compositional features .................................... 63
    5.4.1 The structure of the features ........................ 63
  5.5 Feature merging ............................................ 66

6 Improved models with feature merging ...................... 69
  6.1 The Alvey Tools experiments .............................. 69
    6.1.1 How this work differs from other work in the field .. 69
  6.2 The problem of overfitting ............................... 70
  6.3 Parse selection with compositional features ............. 71
  6.4 Feature merging and overfitting reduction ............... 72
  6.5 Experiments with Alvey Tools ............................. 74
    6.5.1 Evaluation ............................................ 75
  6.6 Results .................................................... 76
## CONTENTS

6.6.1 Performance of unmerged models .................................. 76  
6.6.2 Performance of merged models .................................. 76  
6.7 Frequency based feature cutoffs .................................. 79  
  6.7.1 Problems with feature cutoffs ................................. 79  
6.8 A combined approach .............................................. 81  
6.9 Increased training data and richer feature sets ................. 81  
  6.9.1 Backing off with compositional features ................. 83  
6.10 Conclusion ...................................................... 86  
  6.10.1 Why investigate further? ................................. 86  

7 Experiments with Alpino ........................................ 89  
7.1 Another context for feature merging .......................... 89  
7.2 Alpino: Wide-coverage Parsing of Dutch .................... 90  
  7.2.1 Grammar ................................................. 90  
  7.2.2 Robust Parsing ........................................... 91  
  7.2.3 Dependency Structures ................................ 91  
  7.2.4 Treebank ................................................ 92  
7.3 Maximum entropy modeling and Alpino ....................... 93  
7.4 The Features and Feature Merging ............................ 93  
  7.4.1 Noise reduction and feature merging .................... 94  
  7.4.2 Why feature merging? .................................. 95  
  7.4.3 Building merged models ................................ 95  
  7.4.4 Composition of lexicalized features .................... 96  
  7.4.5 Composition of multiple rule features .................. 96  
7.5 Experiments .................................................... 98  
  7.5.1 Experiments with lexicalized features ................. 98  
  7.5.2 Evaluation ............................................... 99  
7.6 Results with lexical features .................................. 99  
  7.6.1 Experiments with multiple-rule features .............. 100  
7.7 Conclusion .................................................... 101  

8 Conclusion ..................................................... 103  
8.1 Summation ...................................................... 103  
8.2 Discussion .................................................... 104  
  8.2.1 Selecting features to merge ............................ 106  
  8.2.2 Modeling with rich feature sets ....................... 107  
8.3 Future directions ............................................... 107  
  8.3.1 Modeling techniques .................................. 107  
  8.3.2 Smoothing operations .................................. 108  
  8.3.3 Parameter estimation .................................. 109  
  8.3.4 Search speed ........................................... 109  
  8.3.5 Parsing in practice ................................. 109
<table>
<thead>
<tr>
<th>CONTENTS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Samenvatting</td>
<td>110</td>
</tr>
<tr>
<td>References</td>
<td>116</td>
</tr>
</tbody>
</table>