

Contents

List of Figures	xiii
List of Tables	xv
1 Introduction	1
1.1 Language Modeling	1
1.2 N-gram Models	2
1.2.1 Estimating an N-gram model	2
1.2.2 Smoothing Techniques	3
1.3 Non-Local Dependencies	5
1.4 Maximum Entropy Language Modeling	7
1.4.1 Advantages of the Maximum Entropy Method	8
1.4.2 Disadvantages of the Maximum Entropy Method	9
1.5 Measuring Language Model Performance	10
2 Maximum Entropy Modeling	13
2.1 The Maximum Entropy Principle	13
2.1.1 Examples: Dice	13
2.1.2 The General Problem	16
2.1.3 The Maximum Entropy Solution	16
2.2 Maximum Entropy Language Modeling	19
2.3 An Example: Building a Maximum Entropy Trigram Model	21
2.4 Features of Maximum Entropy Language Models	23
2.4.1 Feature Categorization By Information Source	24
2.4.2 Feature Categorization By Order	26
2.4.3 Feature Patterns	27
2.5 Training Maximum Entropy Language Models	28
2.5.1 Computing the Normalization Factors	30
2.5.2 Computing Feature Expectations	31
2.5.3 Updating α by Newton's Method	32

3	Hierarchical Training Methods	34
3.1	Introduction	34
3.1.1	Lower Bound and Upper Bound on Complexity	36
3.2	Hierarchical Training Method for Maximum Entropy Models with only Nested Features	37
3.2.1	Computing Normalization Factors in a Trigram Model	38
3.2.2	Extension for N-gram Models	41
3.3	Exploiting the Feature Hierarchy for Computing Maximum Entropy Models	48
3.3.1	Relations between Features	49
3.3.2	Representing the Feature Hierarchy of ME Models in a Digraph	51
3.4	Expansion for Models with Non-Nested Features	51
3.4.1	Training Models with Non-Nested and Overlapping Features .	53
3.4.2	Model with M Kinds of Bigram Constraints	57
3.4.3	Analysis of Complexity	60
3.5	Generalized Hierarchical Training Methods	63
3.5.1	Training MN-gram Models Hierarchically	63
3.5.2	Training a General ME Model Hierarchically	67
3.6	Divide-and-Conquer	70
3.7	Feature Expectation	73
3.8	Examples: The Training of Several Practical ME Models with Non- nested Features	76
3.8.1	Training ME Models with Syntactic Constraints	76
3.8.2	Training Composite ME Model with Both Topic and Syntactic Dependencies	78
3.9	Experimental Results for Speed-up	79
3.9.1	Nominal Speed-up vs. Real Speed-up	80
3.9.2	Hierarchical Training for N-gram Models	80
3.9.3	Generalized Hierarchical Training for Syntactic Model	82
3.9.4	Divide-and-Conquer and Topic-Dependent Model	83
3.9.5	Divide-and-Conquer Combined with Generalized Hierarchical Training	84
3.10	Summary	85
4	Topic Dependent Language Models	87
4.1	Introduction	87
4.2	Exploiting Topic Dependence	89
4.2.1	Topic Classification	90
4.2.2	Selection of Topic Sensitive Words	91
4.2.3	Formulation of the ME Topic Model	92
4.2.4	Topic Assignment to the Test Utterance	93
4.3	Switchboard Results	94

4.3.1	Baseline Results	95
4.3.2	Topic Assignment During Testing	95
4.3.3	Analysis of Recognition Performance	101
4.4	Comparison with Related Methods	103
4.4.1	Maximum Entropy vs. Linear Interpolation	103
4.4.2	Topic Dependent Model vs. Cache-Based Model	104
4.5	Broadcast News Experiments	106
4.5.1	Performance of ME models	106
4.5.2	Comparison with Interpolated Models	108
4.6	Summary	108
5	Syntactic Language Models	110
5.1	Syntactic Parsing	110
5.2	Quantitative Measurement for Syntactic Dependencies	112
5.3	Formulation of the Maximum Entropy Model	114
5.4	Tokenization Issues (Switchboard vs. Penn Treebank)	116
5.5	Experimental Results in Switchboard	117
5.5.1	Head-Word Model	117
5.5.2	Non-Terminal Model	119
5.5.3	Full Syntactic Model	119
5.6	Analysis of Recognition Performance	120
5.7	Comparison with Related Methods	122
5.7.1	Syntactic Model vs. Class-Based Model	122
5.7.2	Maximum Entropy vs. Linear Interpolation	124
5.8	Broadcast News Results	125
5.8.1	Baselines	125
5.8.2	Syntactic Models	126
5.8.3	Limitations of Computer Speed and Memory Space	127
5.9	Summary	128
6	Composite Language Model with Both Topic Dependencies and Syntactic Dependencies	130
6.1	Introduction	130
6.2	Formulation	131
6.3	Overall Experimental Results for Switchboard	132
6.4	Analysis of Performance for Switchboard	134
6.4.1	Role of Topic Dependencies	134
6.4.2	Role of Syntactic Dependencies	136
6.4.3	Four-Way Analysis	137
6.5	Broadcast News Results	138
6.6	Summary	140

7	Improving the Computation when Using ME Models	141
7.1	Converting ME Models to ARPA Format	142
7.1.1	ARPA Format for Back-off Models	142
7.1.2	Mapping the ME N-gram Model Parameters to ARPA Back-off Model Parameters	144
7.1.3	Speed-up	146
7.1.4	Using the ME Trigram Model in the First Pass of Speech Recog- nition	147
7.2	Approximating Models with Topic Features	148
7.2.1	Experimental Results	150
7.3	Caching Recent Histories	152
7.3.1	Experimental Results	156
7.4	Summary	156
8	Conclusion	158
8.1	Contributions and Summary	158
8.1.1	Accuracy	158
8.1.2	Efficiency	159
8.1.3	Flexibility	161
8.1.4	Summary of Experimental Results	162
8.2	Future Work	162
A	Sign-Test	166
B	Documentation of the Maximum Entropy Toolkit	169
B.1	Implementation of Hierarchical Training Algorithms	169
B.1.1	Feature Expectation	170
B.1.2	Merging Partial Feature Expectations and Updating Model Pa- rameters	173
B.1.3	Memory Concerns	173
B.1.4	Computing Conditional Probabilities Using ME Models	174
B.2	Data Structures	174
B.2.1	Model parameters	174
B.2.2	Features	175
B.2.3	History	177
B.2.4	Future	179
C	User Manual	181
C.1	Overview	181
C.1.1	Availability	181
C.1.2	Functionality	182
C.1.3	System Requirements	182

C.1.4	Installation	182
C.1.5	Tree Structure of Directories	183
C.1.6	Executables	184
C.2	Training an ME Model	185
C.2.1	Data Preparation and Feature Selection	185
C.2.2	Computing Feature Expectations	185
C.2.3	Updating Parameters	186
C.3	Computing Probabilities Using ME Models (Evaluation)	186
C.4	File Formats	186
C.4.1	Model Parameter File	187
C.4.2	N-gram File	187
C.4.3	Discounting File	189
C.4.4	History Equivalence Class File (By Information Source)	189
C.4.5	History File	190
C.4.6	Training Tuple File	191
C.4.7	History Equivalence Class File (By Order)	191
C.4.8	Tuple File for History Equivalence Classes and Futures	192
C.4.9	Test Tuple File	192
C.4.10	Argument File	193
C.5	Advanced Topics	195
C.5.1	Reducing the Number of Iterations	195
C.5.2	Add or Drop features	195
C.5.3	Using Other Smoothing Methods	195
C.5.4	High Order Features	196
C.6	Exercises	196
C.6.1	Building a Trigram Model	196
C.6.2	Building a Flat Model	197
C.6.3	Building a Skipped N-gram Model	197
C.7	Troubleshooting	197
C.8	Update	198

Bibliography