

Contents

Title page	i
Acknowledgments	iii
Abstract	v
Table of contents	xiii
List of tables	xv
List of figures	xvii
1 Introduction	3
2 Peculiarities and Challenges of the Arabic NLP	9
2.1 The Arabic Language: Scripture, Encodings and Morphology	12
2.2 Hardness of Information Retrieval and Question Answering	21
2.3 Hardness of Arabic Named Entity Recognition	34
2.4 Concluding Remarks	35
3 The Named Entity Recognition Task	37
3.1 How Important is the Named Entity Recognition Task?	40
3.2 The Named Entity Recognition Task: Definitions	44
3.3 The Named Entity Recognition Task State-of-Art	50
3.4 Concluding Remarks	58
4 Statistical Modeling Approaches for Named Entity Recognition	61
4.1 Maximum Entropy	64
4.2 Conditional Random Fields	68
4.3 Support Vector Machines	72
4.4 Concluding Remarks	83
5 A Maximum Entropy based Approach: 1 and 2 steps	85
5.1 The ANERcorp	87
5.2 1-step ME-based Approach	93
5.3 The 2-step Approach	98
5.4 Concluding Remarks	103

6	ME, CRFs and SVMs with Different Data-sets	107
6.1	Features	109
6.2	Data	114
6.3	Experiments and Results	116
6.4	Results Discussion and Error Analysis	119
6.5	Concluding Remarks	126
7	A Classifiers Combination Approach	129
7.1	Features	131
7.2	Classic and Fuzzy Borda Voting Scheme	133
7.3	Data	138
7.4	Experiments and Results	139
7.5	Results Discussion and Error Analysis	146
7.6	Concluding Remarks	150
8	Using Large External Resources	153
8.1	Introduction	154
8.2	Extracting Feature from Aligned Corpus	156
8.3	Experiments and Results	160
8.4	Results Discussion and Error Analysis	163
8.5	Concluding Remarks	165
9	Conclusions	167
9.1	Findings and Research Directions	169
9.2	Thesis Contributions	171
9.3	Further Challenges	172
	Bibliography	175
A	Bayesian Networks	187