

Contents

Zusammenfassung	
Abstract	
Acknowledgements	3
1 Introduction	1
1.1 Problem Statement	3
1.2 Research Questions	5
1.3 Research Methodology	6
1.4 Contribution	7
1.5 Publication	10
1.6 Outline	11
2 Fundamentals	12
2.1 Machine Learning Models	12
2.1.1 Artificial Neural Networks	12
2.1.2 Recurrent Neural Networks	14
2.1.3 Long Short-Term Memory (LSTM)	15
2.1.4 Bidirectional LSTM	16
2.1.5 Hidden Markov Model	17
2.1.6 Support Vector Machines	18
2.1.7 The Kernel-based Approaches	19
2.1.8 Decision Tree	20
2.2 Natural Language Processing	21
2.3 Application of Natural Language Processing	22
2.3.1 Part-of-Speech (POS) Tagging	22

2.3.2	Text Chunking	22
2.3.3	Parsing	22
2.3.4	Machine Translation	22
2.3.5	Named Entity Recognition	23
2.3.6	Coreference Resolution	23
2.3.7	Relation Extraction	23
2.3.8	Question Answering	23
2.3.9	Spam Filtering	23
2.4	History of Information Extraction	24
2.5	Dari Language	26
3	Data Collection and Data Preprocessing	29
3.1	Introduction	29
3.2	Data Collection	30
3.3	Data Preprocessing	31
3.3.1	Data Preprocessing Challenges in Dari Language	32
3.3.2	Document Triage	33
3.3.3	Text Segmentation	34
3.4	Dari Part-of-Speech	
	Tagging	38
3.4.1	DariPOS Corpus Building	39
3.4.2	DariPOS Corpus Annotation	40
3.4.3	Hidden Markov Model Based Part-of-Speech Tagging	42
3.4.4	Experimental Results	44
3.5	Conclusion	46
4	Dari Named Entity Recognition	48
4.1	Task Definition	50
4.2	Evaluation Metrics	52
4.2.1	MUC Evaluation	52
4.2.2	CoNLL Evaluation	54
4.3	Related Work	54

4.4	Research Observations	56
4.4.1	Language Factor	57
4.4.2	Domain Factor	58
4.4.3	Entity Type Factor	59
4.5	Datasets for Named Entity Recognition	59
4.6	Features	62
4.7	Experimental setup	63
4.7.1	Model Architecture	63
4.7.2	Linguistic Issues and Challenges	63
4.7.3	Developing NER Datasets for Dari Language	64
4.7.4	Dari Corpus Annotation for Named Entity Recognition	66
4.7.5	Data Format	68
4.8	Methodology	70
4.8.1	Hidden Markov Model versus Artificial Neural Network Models	70
4.8.2	Features Set Description	72
4.9	Results Analysis	72
4.9.1	Recurrent Neural Networks	73
4.9.2	Hidden Markov Model	75
4.9.3	Out of Vocabulary words	78
4.9.4	Conclusions	79
5	Coreference Resolution	81
5.1	Literature Review	84
5.1.1	Rule-Based Approaches to Coreference Resolution	84
5.1.2	Machine Learning Approaches to Coreference Resolution	85
5.2	Resources Used for Coreference Resolution	89
5.2.1	Message Understanding Conferences	91
5.2.2	OntoNotes	92
5.2.3	Automatic Content Extraction (ACE)	93
5.2.4	WikiCoref	94

5.3	Evaluation Scores	95
5.3.1	Message Understanding Conference (MUC)	96
5.3.2	B-Cubed (B^3)	97
5.3.3	Constrained Entity-Alignment F-measure (CEAF)	97
5.4	Experimental Setup	98
5.4.1	Coreference Resolution Model Architecture	98
5.4.2	Resources	99
5.5	Methodology	112
5.5.1	A Comparative Mention-Pair Model for Coreference Resolution in the Dari Language for Information Extraction	114
5.5.2	Decision Tree Classifier using Scikit-learn	125
5.6	Conclusion and Future Work	126
6	Relation Extraction	128
6.1	Predicting the Relationship Between a Given Entity Pair	132
6.2	Overview of Relation Extraction	133
6.3	Rule-based Systems	133
6.4	Supervised Systems	135
6.4.1	Feature-based Methods	136
6.4.2	Kernel-based Methods	137
6.4.3	Distantly Supervised	148
6.5	Experimental Setup	149
6.5.1	Kernel-based Relation Extraction on Dari language	149
6.5.2	Evaluation and Error Analysis	152
6.5.3	Conclusion	156
	Bibliography	170