Contents

List of Figures xi

List of Tables xiv

List of Symbols xv

List of Abbreviations xvii

Abstract xix

1 Introduction 1

1.1 Key issues in low-resource machine translation 2

1.1.1 Lack of annotated and labelled datasets 3

1.1.2 Supporting multiple dialects of a language 3

1.1.3 Morphological richness 3

1.1.4 Word order 3

1.1.5 Globalization pressure 4

1.2 Motivation 4

1.2.1 Language preservation 4

1.2.2 Educational applications 5

1.2.3 Knowledge expansion 5

1.2.4 Monitoring demographic and political processes 5

1.2.5 Emergency response 5

1.3 Research objectives 6

1.4 Contributions 7

2 Theoretical Background and Literature Review 11

2.1 Theoretical Background 11

2.1.1 Languages 11

2.1.2 Related or Similar Languages 12
2.1.3 Categories of machine translation 13
2.1.4 NMT 15
2.1.5 Reinforcement learning 17
2.2 Literature Review 18
2.2.1 Zero-shot Problem 18
2.2.2 Morphological Richness 20
2.2.3 Domain Shift Problem 21
2.2.4 Missing Context and Rare-word Problem 24
2.3 Research gap 25
2.3.1 Zero-shot problem 27
2.3.2 Morphological richness 27
2.3.3 Domain shift 27
2.3.4 Missing context and rare-word problem 28
2.4 Evaluation metrics 28
2.4.1 BLEU 28
2.4.2 chrF2 29
2.4.3 Translation Edit Rate (TER) 29

3 Common Phonetic-Orthographic Representation 31
3.1 Introduction 31
3.2 Proposed Approach 35
3.2.1 Text Encoder 35
3.2.2 Model Training 37
3.2.3 Text Decoder 37
3.3 Corpus and Experimental Settings 37
3.3.1 Corpus description 38
3.3.2 Training details 39
3.4 Results and Analysis 40
3.4.1 Similarity between languages 42
3.4.2 Analysis of language complexity 44
3.5 Ablation Study 49
3.5.1 Analysis of the proposed approach on distant language pairs . . 49
3.5.2 Unsupervised settings 49
3.5.3 Back-translation 50
3.6 Summary 50

4 Transfer Learning-based Semi-supervised Pseudo-corpus Generation 53
4.1 Introduction . . . 53
4.2 Transfer Learning-based Semi-supervised Pseudo-corpus Generation . . 56
4.2.1 TSL . . . 56
4.2.2 MSL . . . 58
4.2.3 TL-based pseudo-corpus generation . . . 60
4.2.4 Model Training . . . 61
4.3 Data and Experimental Setup . . . 62
4.3.1 Data Preparation . . . 62
4.3.2 Experimental Setup . . . 63
4.4 Results and Analysis . . . 65
4.4.1 Impact of TSL and MSL . . . 65
4.4.2 Impact of TLSPG . . . 66
4.4.3 Relatedness between languages . . . 66
4.4.4 Impact of out-of-vocabulary . . . 67
4.4.5 Impact of language complexity . . . 68
4.4.6 Using LSTM in-place of Transformer in TLSPG . . . 69
4.5 Summary . . . 71

5 Incorporating Reinforcement Learning in Domain Adaptation 73
5.1 Introduction . . . 73
5.2 Proposed RSSW Method . . . 76
5.2.1 Language model . . . 76
5.2.2 Policy network . . . 78
5.2.3 Translation model training . . . 81
5.2.4 End-to-end training . . . 82
5.3 Experiments . . . 83
5.3.1 Corpus description . . . 83
5.3.2 Experimental setup . . . 84
5.3.3 Results and ablation study . . . 85
5.4 Summary . . . 89

6 Incorporating Reinforcement Learning in Adversarial Network 91
6.1 Introduction . . . 91
6.2 System Model and Problem Discussion . . . 93
6.2.1 System model . . . 93
6.2.2 Problem statement . . . 96

6.3 Proposed Approach 97
6.3.1 Joint embedding 97
6.3.2 Generator 98
6.3.3 Generator training procedure 100
6.3.4 Discriminator 101
6.3.5 Model training 102
6.4 Performance Study 102
6.4.1 Datasets 103
6.4.2 Settings 104
6.4.3 Results and analysis 105
6.5 Summary 109

7 Conclusions and Future Work **111**

References **116**

List of Publications **135**