

# CONTENTS

<b>Preface</b>	<b>9</b>
<b>1 Introduction</b>	<b>13</b>
1.1 Large Language Models: From MT and Back . . . . .	14
1.2 Overview of Unsupervised MT . . . . .	15
1.3 Structure of the Book . . . . .	16
<b>2 Background</b>	<b>19</b>
2.1 Language Data Resources . . . . .	20
2.1.1 Monolingual Corpora . . . . .	20
2.1.2 Parallel Corpora . . . . .	21
2.1.3 Comparable Corpora . . . . .	21
2.1.4 Pseudo-Parallel Corpora . . . . .	22
2.1.5 Synthetic Parallel Corpora . . . . .	22
2.1.6 Pre-Trained Models . . . . .	22
2.2 Cross-lingual Information in Monolingual Data . . . . .	23
2.3 Languages of the World . . . . .	24
2.4 Low-Resource Languages . . . . .	25
2.5 The Extent of This Study . . . . .	26
<b>3 NLP Fundamentals</b>	<b>29</b>
3.1 Word Embeddings . . . . .	30
3.1.1 Static Word Embeddings . . . . .	30
3.1.2 Contextual Word Embeddings . . . . .	32
3.1.3 Cross-lingual Word Embeddings . . . . .	32
3.2 Transformer Language Models . . . . .	34
3.2.1 Architecture . . . . .	35
3.2.2 Input Embeddings . . . . .	37
3.2.3 Self-Attention . . . . .	38

3.2.4	Unsupervised Pre-Training . . . . .	39
3.2.5	Multilingual Pre-Training . . . . .	42
3.2.6	Internal Representations . . . . .	42
3.3	Machine Translation . . . . .	43
3.3.1	Neural Machine Translation . . . . .	43
3.3.2	Phrase-Based Machine Translation . . . . .	46
3.3.3	Machine Translation Evaluation . . . . .	47
<b>4</b>	<b>Approaches to Unsupervised MT</b>	<b>51</b>
4.1	Model-Centric Approaches to UMT . . . . .	52
4.1.1	Model Architecture . . . . .	52
4.1.2	Model Initialization . . . . .	55
4.1.3	Training Strategies . . . . .	57
4.1.4	Decoding Strategies . . . . .	59
4.2	Data-Centric Approaches to UMT . . . . .	60
4.2.1	Pseudo-Parallel Data . . . . .	60
4.2.2	Synthetic Data . . . . .	61
4.2.3	Multilingual Data . . . . .	63
<b>5</b>	<b>Parallel Corpus Mining</b>	<b>65</b>
5.1	Related Work . . . . .	67
5.2	Methodology . . . . .	68
5.2.1	Pre-trained Multilingual Masked Language Models . . . . .	68
5.2.2	Fine-tuning MLMs with a Translation Objective . . . . .	68
5.2.3	Fine-tuning MLMs for Unsupported Languages . . . . .	69
5.2.4	Sentence Embeddings . . . . .	70
5.2.5	Searching in Multilingual Embedding Space . . . . .	70
5.3	Experiments . . . . .	71
5.3.1	Model . . . . .	71
5.3.2	Data . . . . .	72
5.3.3	Training . . . . .	72
5.3.4	Benchmarks . . . . .	72
5.4	Results . . . . .	73
5.4.1	Evaluation I: Parallel Corpus Mining . . . . .	73
5.4.2	Evaluation II: Corpus Deshuffling . . . . .	75

5.4.3	Analysis: Representations Across Layers . . . . .	77
5.4.4	Parallel Corpus Mining for Unsupported Languages . . . . .	77
5.5	Takeaways . . . . .	80
<b>6</b>	<b>Unsupervised Machine Translation Methodology</b>	<b>83</b>
6.1	Unsupervised Cross-Lingual Embeddings . . . . .	84
6.1.1	Seed Lexicon . . . . .	84
6.1.2	Self-Refinement . . . . .	85
6.1.3	Applications in Unsupervised MT . . . . .	85
6.2	Unsupervised Phrase-Based Machine Translation . . . . .	86
6.2.1	Cross-Lingual Phrase Embeddings . . . . .	86
6.2.2	Initial Phrase Table Induction . . . . .	87
6.2.3	Language Model . . . . .	87
6.2.4	Unsupervised Tuning . . . . .	87
6.2.5	Back-Translation . . . . .	87
6.3	Unsupervised Neural Machine Translation . . . . .	88
6.3.1	Vocabulary . . . . .	89
6.3.2	Architecture . . . . .	89
6.3.3	Pre-Training . . . . .	90
6.3.4	Fine-Tuning for Translation . . . . .	93
6.3.5	Baselines . . . . .	94
<b>7</b>	<b>Experiments &amp; Results</b>	<b>95</b>
7.1	Phrase-Based Unsupervised MT . . . . .	96
7.1.1	Data . . . . .	97
7.1.2	Model & Training . . . . .	97
7.1.3	Results & Discussion . . . . .	99
7.1.4	Takeaways . . . . .	101
7.2	Hybrid Unsupervised MT . . . . .	101
7.2.1	Data . . . . .	102
7.2.2	Model & Training . . . . .	102
7.2.3	Results & Discussion . . . . .	104
7.2.4	Takeaways . . . . .	106
7.3	Effect of Pre-Training Strategies . . . . .	108
7.3.1	Data . . . . .	108

7.3.2	Model & Training . . . . .	109
7.3.3	Results & Discussion . . . . .	111
7.3.4	Takeaways . . . . .	114
7.4	Boosting Unsupervised MT with Pseudo-Parallel Data . . . . .	115
7.4.1	Data . . . . .	115
7.4.2	Model & Training . . . . .	115
7.4.3	Results & Discussion . . . . .	118
7.4.4	Takeaways . . . . .	122
7.5	Limitations of Unsupervised MT . . . . .	124
7.5.1	Data . . . . .	124
7.5.2	Model & Training . . . . .	125
7.5.3	Results & Discussion . . . . .	125
7.5.4	Takeaways . . . . .	127
7.6	Pseudo-Parallel Data in Semi-Supervised MT . . . . .	128
7.6.1	Data . . . . .	128
7.6.2	Model & Training . . . . .	128
7.6.3	Results & Discussion . . . . .	130
7.6.4	Takeaways . . . . .	131
<b>8</b>	<b>Discussion</b>	<b>133</b>
	<b>Conclusion</b>	<b>142</b>
	Acknowledgements . . . . .	144
	<b>Bibliography</b>	<b>145</b>
	<b>Appendix</b>	<b>163</b>
	A.1 Additional Evaluation (COMET and chrF++) . . . . .	163
	A.2 Tools and Configuration . . . . .	166
	<b>List of Figures</b>	<b>167</b>
	<b>List of Tables</b>	<b>171</b>
	<b>List of Abbreviations</b>	<b>175</b>