

Contents

1	Introduction	1
1.1	Neural Networks Inspired by Humans	1
1.2	Multi-Lingual Machine Translation	2
1.3	Aims of This Book	3
1.4	Intended Audience for This Book	4
1.5	Book Structure	5
I	Background	7
2	Reverting the Babel Curse	9
2.1	The Benefits of Language Diversity	9
2.2	Why More than Two Languages in MT?	10
2.2.1	Efficiency	10
2.2.2	Flexibility	11
2.2.3	Quality	12
3	The Versatility of Neural Networks	13
3.1	Characteristics of Machine Translation Task	13
3.1.1	Sentence-Level Translation	14
3.1.2	Space of Possible Outputs for Sequence-to-Sequence Task	14
3.2	Processing Words	16
3.2.1	Word Embeddings Aware of Word Structure	17
3.2.2	Subword Representation	19
3.3	Processing Sentences	22
3.3.1	Sequence-to-Sequence Models	23
3.3.2	Transformer Model	23
3.4	Input and Output Versatility	27

3.5	Multi-Tasking	28
3.5.1	Multi-Tasking Benchmarks	29
3.5.2	Task Related Topics	30
3.5.3	Multi-Tasking Architectures	31
3.6	Back-Translation	32
4	Learning Skills and Pitfalls of Neural Networks	37
4.1	Devil in the Detail	37
4.2	Language Resources	38
4.2.1	What is Domain in Machine Translation?	39
4.2.2	Definition of Low-Resource Languages	40
4.2.3	Resource Quality	42
4.2.4	Corpus Cleaning	42
4.3	Measuring Training Progress	43
4.3.1	Convergence and Stopping Criterion	44
4.4	Training Instability	46
4.5	Lack of Generalization	49
4.6	Catastrophic Forgetting	50
4.7	The Cost of Multi-Tasking	52
5	The Battle Against Wishful Thinking	55
5.1	Machine Translation Evaluation	55
5.1.1	Manual Evaluation	56
5.1.2	Automatic Metrics	56
5.2	Statistical Significance	58
5.3	It Works in Low-Resource	59
5.4	Simple Contrastive Task	60
5.4.1	Dummy Diagonal Parse	60
5.4.2	Dummy Supertags	61
II	Multilingual Models	63
6	Overview	65
6.1	Classification of Multilingual Models	65

7	Transfer Learning	69
7.1	Terminology	70
7.2	Cold-Start	72
7.2.1	Cold-Start Direct Transfer	73
7.2.2	Direct Transfer Results	73
7.2.3	Parent Vocabulary Effect	75
7.2.4	Odia Subword Irregularity	77
7.2.5	Vocabulary Overlap	78
7.3	Vocabulary Transformation	79
7.3.1	Results with Transformed Vocabulary	80
7.3.2	Various Vocabulary Transformations	81
7.3.3	Training Time	82
7.4	Warm-Start	83
7.4.1	Vocabulary Shared between Parent and Child	84
7.4.2	Comparison of Shared Vocabularies	85
7.4.3	Warm-Starts with Ten Languages	87
7.4.4	Mixing Parent and Child	90
7.4.5	Analysis of Balanced Vocabulary	92
7.5	Warm vs Cold-Start	94
7.6	When Transfer is Negative?	96
7.6.1	Traces of Parent	97
7.6.2	Extremely Low-Resourced	100
7.6.3	Low-Resource Parent	101
7.6.4	No Shared Language	102
7.7	Position of Shared Language Matters?	104
7.7.1	Shared Language Position Affects Convergence Speed	104
7.7.2	Shared Language Position Affects Slope of Learning Curve	105
7.7.3	Parent Performance Drop	106
7.8	Rather Related Language, or More Data?	108
7.8.1	Artificially Related Language Pair	109
7.9	Linguistic Features, or Better Initialization?	112
7.9.1	Freezing Parameters	113
7.9.2	Swapping Direction in Parent and Child	115

7.9.3	Broken Word Order in Parent Model	116
7.9.4	Output Analysis	118
7.9.5	Various Lengths of Parent Sentences	120
7.9.6	Parent’s Performance Influence	121
7.9.7	Same Language Pair in Reverse Direction	124
7.10	Back-Translation with Transfer Learning	126
8	Observations and Advances in Multilingual MT	129
8.1	Single Source, Multilingual Target	130
8.2	Single Target, Multilingual Source	131
8.3	Multilingual Many-to-Many	133
8.3.1	Limited Sharing in Multilingual Many-to-Many Models	134
8.3.2	A Universal Model for Multilingual Many-to-Many	136
8.3.3	Sentence Representations in Multilingual Models	137
8.4	Massively Multilingual Models	139
8.5	Massive Massively Multilingual Models	142
9	Practical Aspects	147
9.1	KIT Lecture Translator	147
9.2	ELITR: European Live Translator	148
9.2.1	Many Languages, Many Problems	148
9.2.2	Distributed Development and Deployment	152
9.3	Computer Clusters	154
9.3.1	Computer Clusters from the Users’ Perspective	155
9.3.2	Conditions and Features of a Good Cluster	157
10	Conclusion: The Prospects	159
10.1	Mind the Gap in Understanding	159
10.1.1	Deep Models Do Not Understand Us	160
10.1.2	We Do Not Understand Deep Models	161
10.2	Massive Models in NLP	162
10.3	Ecological Trace of NMT Research	163
10.3.1	General Environmental Concerns	164
10.3.2	Carbon Footprint of Our Transfer Learning Experiments . .	165

CONTENTS

10.4 Final Words 167

Bibliography 168

List of Observations 185

Index 189