

Contents

Abstract	i
Resumen	iii
Resum	vii
Acknowledgments	xi
Preface	xiii
Thesis Structure	xiv
Related Publications	xvi
Contents	xx
1 Overview of Statistical Machine Translation	1
1.1 Introduction	1
1.2 Machine Translation	2
1.3 Statistical Machine Translation	4
1.3.1 Phrase-based Statistical Machine Translation	6
1.3.2 Language models	12
1.4 Continuous vector-space representation	14
1.4.1 Word embeddings	16
1.4.2 Continuous Skip-Gram Model	16
1.4.3 Sentence embeddings methods	18
1.5 Neural Statistical Machine Translation	19
1.5.1 Encoder-decoder architecture	20
1.5.2 Training	24
1.5.3 Decoding with beam search	25

1.6	Summary	25
2	SMT experimental framework	27
2.1	Introduction	27
2.2	Evaluation criteria	28
2.2.1	Bilingual Evaluation Understudy	28
2.2.2	METEOR metric	29
2.2.3	Translation Edit Rate metric	29
2.3	Corpora	30
2.4	Toolkits	36
2.5	Summary	38
3	Data selection preliminaries	39
3.1	Introduction	39
3.2	Adaptation	40
3.3	Adaptation in SMT	41
3.3.1	Off-line adaptation	42
3.4	Data selection	43
3.4.1	Cross-Entropy based methods	45
3.4.2	Infrequent ngrams recovery	47
3.5	Domain adaptation in NMT	48
3.6	Summary	49
4	Corpus selection for SMT training	51
4.1	Introduction	51
4.2	CRSDS technique	52
4.2.1	Similarity corpus	53
4.2.2	Sentences embedding methods	53
4.2.3	CRSDS technique	53
4.2.4	Bilingual-CRSDS technique	56
4.3	NNCDS technique	56
4.3.1	Neural network architecture	57
4.3.2	Semi-supervised selection	59
4.4	Experiments	60
4.4.1	Experimental setup	60
4.4.2	CRSDS experimental results	62
4.4.3	NNCDS experimental results	65

4.4.4	Comparative DS method using the in-domain corpus	65
4.4.5	DS method comparison using the source test corpus	77
4.5	Summary	83
5	Model combination	85
5.1	Introduction	85
5.2	Related work	87
5.3	Data selection method	88
5.4	Combination methods	89
5.4.1	Linear interpolation	89
5.4.2	Fill-up method	89
5.5	Experiments	90
5.5.1	Experimental setup	90
5.5.2	Interpolated language model results	91
5.5.3	Translation model combination results	94
5.5.4	Comparison with a concatenation approach	98
5.6	Summary	100
6	Looking for the right development corpus	103
6.1	Introduction	103
6.2	Related work	105
6.3	Development DS techniques	105
6.3.1	Levenshtein Distance DDS	106
6.3.2	DDS with vector-space representations	107
6.4	Experiments	110
6.4.1	Experimental setup	111
6.4.2	Controlled scenario results	111
6.4.3	Real scenario results	117
6.5	Summary	119
7	Data selection in NMT	121
7.1	Introduction	121
7.2	DS for training PBSMT and NMT approaches	123
7.3	Data selection to create synthetic data	123
7.3.1	Synthetic data creation method	124
7.4	Experiments	125

7.4.1	Experimental setup	125
7.4.2	Training a NMT system	125
7.4.3	Fine tuning with synthetic data	126
7.5	Summary	133
8	Conclusions	137
8.1	Summary	137
8.2	Future works	139
	Appendix A Log-linear weight adaptation	141
A.1	Introduction	141
A.2	Discriminative ridge regression for SMT	142
A.2.1	Sentence-by-sentence DRR	143
A.2.2	Batch DRR	145
A.3	Experiments	147
A.3.1	Corpora	148
A.3.2	Experimental setup	148
A.3.3	DRR experiments	149
A.3.4	Comparison between DRR, MERT and MIRA	151
A.4	Summary	160
	List of Symbols and Abbreviations	161
	List of Figures	164
	List of Tables	166
	List of Algorithms	169
	Bibliography	171