Contents

Acknowledgements **v**

Abstract **vii**

Preface **xiii**

Contents **xvii**

1 Preliminaries **1**
- 1.1 Introduction . . . 1
- 1.2 Text classification . . . 2
  - 1.2.1 Support vector machines and boosting techniques . . . 4
  - 1.2.2 Machine-aided indexing . . . 5
- 1.3 Machine translation . . . 6
- 1.4 Statistical MT . . . 7
- 1.5 Statistical phrase-based translation systems . . . 10
  - 1.5.1 Generative phrase-based models . . . 10
  - 1.5.2 Heuristic phrase-based models . . . 12
- 1.6 Automatic MT evaluation metrics . . . 14
- 1.7 Computer-assisted translation . . . 15
- 1.8 The expectation-maximisation algorithm . . . 16
- 1.9 Finite mixture modelling . . . 18
- 1.10 Scientific contributions . . . 21

Bibliography **25**

2 Bilingual text classification **37**
- 2.1 Introduction . . . 37
- 2.2 Unigram mixture model . . . 38
  - 2.2.1 The model . . . 38
  - 2.2.2 Maximum likelihood estimation . . . 39
- 2.3 Bilingual unigram mixture model . . . 40
  - 2.3.1 The model . . . 40
  - 2.3.2 Maximum likelihood estimation . . . 40
  - 2.3.3 Smoothing . . . 41
- 2.4 Bilingual text classification using unigram models . . . 42
  - 2.4.1 Decision rules . . . 42
  - 2.4.2 Maximum likelihood estimation for supervised classification 43

2.5 Experimental results 44
2.5.1 Datasets 45
Traveller dataset 45
BAF corpus 46
JRC-Acquis corpus 46
2.5.2 Experimental results on Traveller and BAF 47
2.5.3 Experimental results on JRC-Acquis 51
2.5.4 Comparative results 52
2.6 Conclusions and future work 53

Bibliography **57**

3 Mixture of M1 models **61**
3.1 Introduction 61
3.2 The M1 Model 62
3.2.1 The model 62
The model using indicator vectors 63
3.2.2 Maximum likelihood estimation 64
3.3 Mixture of M1 models 65
3.3.1 The model 65
3.3.2 Maximum likelihood estimation 66
3.3.3 Smoothing 67
3.4 Bilingual text classification using the M1 model 68
3.4.1 Mixture of unigram-M1 models 69
The model 69
Maximum likelihood estimation 69
3.4.2 Bilingual text classification using the unigram-M1 model . 70
The decision rule 70
Maximum likelihood estimation for supervised classification 70
3.4.3 Experimental results 71
3.5 Mixture of M1 models applied to MT 73
3.5.1 Viterbi alignment 74
3.5.2 Evaluation of alignment quality 75
Corpora 75
Experimental results 76
3.5.3 Evaluation of translation quality 77
Corpora 77
Experimental setting and results 79
3.6 Conclusions and future work 81

Bibliography **83**

4 Mixture of M2 models 85
4.1 Introduction 85
4.2 The M2 model 85
4.2.1 The model 85
4.2.2 Maximum likelihood estimation 86
4.3 Mixture of M2 models 87
4.3.1 The model 87
4.3.2 Maximum likelihood estimation 87
4.3.3 Smoothing 88
4.4 Decoding algorithm 88
4.4.1 Decoding parameters 90
4.5 Experimental results 92
4.6 Further evaluation 94
4.6.1 Evaluation of alignment quality 95
4.6.2 Evaluation of translation quality 95
4.7 Conclusions and future work 95

Bibliography 99

5 Mixture of HMM alignment models 101
5.1 Introduction 101
5.2 The HMM alignment model 101
5.2.1 The model 101
5.2.2 Maximum likelihood estimation 102
5.3 Mixture of HMM alignment models 104
5.3.1 The model 104
5.3.2 Maximum likelihood estimation 104
5.3.3 Viterbi alignments 105
5.4 Experimental results 106
5.4.1 Evaluation of alignment quality 106
5.4.2 Evaluation of translation quality 107
5.5 Conclusions and future work 107

Bibliography 109

6 Computer-assisted translation based on stochastic finite-state transducers 111
6.1 Introduction 111
6.2 Machine translation with finite-state transducers 112
6.2.1 Learning finite-state transducers 113
6.3 Interactive and predictive search 114
6.3.1 Word graph derivation 116
6.3.2 Search for n-best translations given a prefix of the target sentence 117

Error-correcting parsing 118
N-best search 119
6.4 Experimental framework and results 120
6.4.1 XRCE and EU corpora 120
Corpora preprocessing 121
6.4.2 Translation quality evaluation 122
6.4.3 Experimental results 123
6.5 Conclusions and future work 126

Bibliography **129**

7 Conclusions **133**
7.1 Summary 133
7.2 Scientific publications 135
7.3 Future work 138

Bibliography **141**

A Additional experiments on bilingual text classification **143**
A.1 Experiments on Traveller and BAF datasets 143
A.1.1 Smoothed n-gram language models 143
A.1.2 Comparative results with SVM and boosting techniques . 145
A.2 Experiments on JRC-Acquis 145
A.2.1 Smoothed n-gram language models 145
A.2.2 Comparative results with boosting techniques 145

Bibliography **147**

B EM derivation **149**
B.1 EM algorithm for finite mixture models 149
B.1.1 E step for finite mixture models 149
B.1.2 M step for finite mixture models 150
B.2 EM algorithm for the unigram mixture model 151
B.2.1 M step for the unigram mixture model 151
B.3 EM algorithm for the M1 model 152
B.3.1 E step in the M1 model 152
B.3.2 M step in the M1 model 153
B.3.3 E step for the M1 mixture model 154
B.4 EM algorithm for the HMM model 155
B.4.1 E step in the HMM model 155
B.4.2 M step in the HMM model 157
B.4.3 E step for the HMM mixture model 158

C Symbols and acronyms **161**
C.1 Mathematical symbols 161
C.2 Acronyms 163

List of Figures **165**

List of Tables **167**