MACHINE LEARNING IN TRANSLATION

Peng Wang and David B. Sawyer

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

CONTENTS

List of Figures *xi*
List of Tables *xii*

Introduction 1

PART I
Human and machine approaches to translation 17

1 Convergence of two approaches to translation 19
Key concepts 19
1.1 Introduction 19
1.2 Language and translation – equivalence[1] in different systems of signs 21
1.3 A mathematical approach – machine learning in language and translation 22
1.4 Two approaches to meaning representation 25
1.5 A holistic view of translation 30

2 Levels of analysis 34
Key concepts 34
2.1 Introduction 34
2.2 Contextual levels in language comprehension and production 35
2.2.1 Levels of analysis in translation 35
2.2.2 The concept of granularity in NLP 36
2.2.3 Chunking: a cognitive aspect behind levels of analysis 37

2.2.4 Level of analysis requirements in machine learning 38
2.2.5 Translation memory: a case study for a balance of form and meaning 39
2.3 Drawing parallels between linguistic relations and levels of analysis 42
2.3.1 Syntagmatic vs. paradigmatic relations 42
2.3.2 Linguistic units for machine learning 44
2.3.3 Segmentation based on meaning 47
2.4 Application of linguistic levels in machine translation 48
2.5 Two approaches to levels of analysis 49
2.5.1 Sequences vs. meaning units 49
2.5.2 Other aspects of comparison 49
2.5.3 Implications for a translator 50

3 Predicative language models 52
Key concepts 52
3.1 Introduction 52
3.2 Language models 53
3.2.1 Language models for both humans and machines 53
3.2.2 n-*gram language models 55*
3.2.3 Conditions for prediction in n-*gram language models 55*
3.2.4 Neural probabilistic language models 57
3.2.5 Embeddings and meaning 58
3.3 Meaning representation in neural language models 59
3.3.1 Distributional semantics 59
3.3.2 Vectorization: from text to numbers 61
3.3.2.1 Words and vectors 61
3.3.2.2 Feature vectors 64
3.3.2.3 Word2vec 65
3.4 Predicative LMs and translation 66

PART II
Machine learning tasks in translation **69**

4 Machine translation 71
Key concepts 71
4.1 Introduction 71
4.2 The expansion of language technology in localization 72
4.2.1 MT and CAT: two lines of development 72
4.2.2 The role of MT and CAT in a localization process 75

4.3 MT paradigms 76
4.3.1 Interlingua: an ideal level of understanding 76
4.3.2 Rule-based MT 78
4.3.3 Statistical MT 79
4.3.4 Neural MT 82
4.3.4.1 Statistical approaches to MT 82
4.3.4.2 Recurrent neural network 83
4.3.4.3 Attention mechanism in MT 83
4.4 MT deployment in the localization process 86
4.4.1 MT deployment infrastructure 86
4.4.2 MT deployment considerations 87
4.5 MT as an instantiation of the process of human translation (HT) 89

5 Machine translation quality assessment and quality estimation 92
Key concepts 92
5.1 Introduction 92
5.2 Fundamental TQA considerations 94
5.2.1 Retrospective and prospective approaches 94
5.2.2 Linguistic relationships 94
5.2.3 Different automation scenarios in TQA 96
5.2.4 Factors in the TQA workflow 96
5.3 Two lines of thought for MT quality assessment 97
5.4 Automated metrics: using machines to assess machines 99
5.4.1 An overview of automated metrics 99
5.4.2 Automated metrics based on n*-gram precision 100*
5.4.2.1 The Bilingual Evaluation Understudy score 100
5.4.2.2 Factors that impact the performance of BLEU 102
5.4.2.3 Other n*-gram based automatic metrics 103*
5.4.3 Automated metrics based on edit distance 104
5.4.3.1 Translation Edit Rate (TER) 104
5.4.3.2 Human-targeted TER (HTER) 104
5.5 Quality estimation (QE) 105

6 Intentionality and NLP tasks in translation 108
Key concepts 108
6.1 Introduction 108
6.2 Classification of NLP tasks from the ML perspective 110
6.2.1 Two types of NLP tasks 110
6.2.2 Purposes of neural network models in translation 111
6.2.2.1 Sequence prediction 111

6.2.2.2 Sequence classification 112
6.3 Purposes of NLP tasks at each step of a continuous translation process 113
6.3.1 Continuous translation process 113
6.3.2 User purposes at each step of a continuous translation process 113
6.4 Intelligent NLP tasks in a continuous translation process 115
6.4.1 Extracting translatable strings from content in a larger context 115
6.4.1.1 Optical character recognition (OCR) 116
6.4.1.2 Automatic speech recognition (ASR) 116
6.4.1.3 Web crawling techniques 117
6.4.2 Preparing the translatable content 117
6.4.2.1 Domain detection 117
6.4.2.2 Automatic segmentation 118
6.4.3 Leveraging past experience/data through human or machine memory systems 119
6.4.3.1 Terminology extraction 119
6.4.3.2 Named entity recognition (NER) 119
6.4.4 Converting meaning across languages and cultures 120
6.4.5 Continuous learning 120
6.4.6 Adapting the target content format to the target locale and culture 121
6.5 Integration of NLP tasks 122
6.6 Intentionality: why machine learning matters in translation 123

PART III
Data in human and machine learning 127

7 Translator–computer interaction through language data 129
Key concepts 129
7.1 Introduction 129
7.2 Translation-driven language data 131
7.2.1 Corpus and language data 131
7.2.2 Translation-driven language data 132
7.2.2.1 Translation-driven corpora 132
7.2.2.2 Comparability of translation-driven corpora 134
7.2.3 Translation-driven language data 134
7.3 TCI in data design and acquisition 135
7.3.1 Human insights in data design and acquisition 135

7.3.2 Encoding translator's insights when creating language data 136
7.3.2.1 Degrees of translator intervention 136
7.3.2.2 The dynamic feature of data quality 138
7.4 TCI through data annotation 139
7.4.1 Raw data vs. annotated data 139
7.4.2 Framework for human annotations 140
7.4.3 The role of human annotation in machine learning 141
7.4.4 Automatic annotation 142
7.4.5 Annotation and markup language 143
7.5 TCI by using language data 144
7.5.1 Corpus-based analysis from the translator's perspective 145
7.5.1.1 Finding patterns in corpora 145
7.5.1.2 Corpus-based linguistic analysis 145
7.5.2 Data analytics for machine learning 146
7.5.2.1 Using data to train ML models 146
7.5.2.2 Using data to monitor and upgrade ML models 147

8 Balancing machine and human learning in translation 148
Key concepts 148
8.1 Introduction 148
8.2 Making inferences in learning 150
8.2.1 The mathematical analysis of learning processes 151
8.2.2 A mental multilingual language model 152
8.2.3 Intentions and connectivity of language models in translation 154
8.3 Personalized ML tools for human learning 155
8.3.1 The collective feature of technologies 155
8.3.2 From generic models to customized and personalized ML technologies 156
8.3.3 User accessibility to human-driven ML systems in translation 158
8.4 Using human-driven ML systems to facilitate human learning 159
8.4.1 Individual cognitive aspects of terminology 159
8.4.2 Vector space and terminological schematic context 160
8.4.3 Source-text oriented comparable corpora for ML and HL 163
8.4.4 Personalized intelligent NLP techniques for human learning 164

8.4.4.1 Machine learning and intelligent tutors 164
8.4.4.2 An intelligent tutor for translation 165

9 Impact of machine learning on translator education 167
Key concepts 167
9.1 Introduction 167
9.2 New ML-driven translation workflows 168
9.2.1 A machine-in-the-loop approach to translation 168
9.2.2 Translation as a language problem 169
9.2.3 Moving beyond routine translation tasks 170
9.2.4 The need for human–computer collaboration 171
9.2.5 The evolving role of translators driven by ML 171
9.3 Translator competences arising from machine learning 174
9.3.1 Primary competences in a continuous translation process 174
9.3.2 Two overarching ML-driven translator competences 175
9.3.2.1 ML-related competence 175
9.3.2.2 Continuous learning competence 176
9.3.3 NLP task-specific competences 176
9.3.3.1 Translation competence in the narrow sense 176
9.3.3.2 Linguistic competence 177
9.3.3.3 NLP competence 178
9.3.3.4 Cultural competence 178
9.4 Impact of machine learning on translator education 179
9.4.1 The educator's role in two approaches to learning environments 179
9.4.2 Incorporating ML into the teaching of translation 180
9.4.3 The impact of ML in a learner-centered environment 181
9.4.4 Building a learning platform for both educators and learners 183
9.5 The value of educators in the ML context 183

Epilogue – human-centered machine learning in translation 185
References 189
Index 203