

CONTENTS

1. Introduction	9
 Part I. THE RULES OF THE GAME	 15
2. Preliminaries	17
2.1. Frequencies and the Zipf's Law	17
2.2. Basics of Probability Calculus	18
2.2.1. <i>Probability Estimation</i>	19
2.2.2. <i>N-gram Models</i>	20
2.2.3. <i>Probabilistic Graphical Models</i> [◊]	21
2.3. Fundamental Paradigms: Rule-Based and Machine Learning Models . .	27
2.3.1. <i>Rule-Based Systems</i>	28
2.3.2. <i>Machine-Learning Systems</i>	28
 3. Deep Neural Networks	 35
3.1. The Basic Unit	35
3.1.1. <i>Biological Neuron Functionality</i>	35
3.1.2. <i>The Artificial Neuron</i>	36
3.2. Networks of Neurons: Basic Architectures	37
3.2.1. <i>Multilayer Perceptrons</i>	38
3.2.2. <i>Convolutional Networks</i>	42
3.2.3. <i>Recurrent Networks</i>	44
3.3. Networks of Neurons: Advanced Architectures	47
3.3.1. <i>Encoder/Decoder Networks</i>	47
3.3.2. <i>Transformers</i>	49
3.3.3. <i>Tokenisation in Transformer-Based Systems</i>	52
3.4. Artificial and Biological Neural Networks	54

4. Vector Representations (a.k.a. Embeddings)	59
4.1. Count-based methods	60
4.2. Don't count, predict!	62
4.3. Contextual Word-embeddings	65
4.4. Sentence/Text Embeddings	68
4.5. Transfer Learning	71
4.6. A Final Consideration on Vector Representations	72
5. Evaluation: the Will and the Way	75
5.1. System Evaluation	75
5.1.1. <i>Instance Classification - Discrete Outputs</i>	75
5.1.2. <i>Textual Output</i>	77
5.1.3. <i>Interactive Systems</i>	77
5.2. Steps in the evaluation process	77
5.3. International Evaluation Challenges	78
5.4. An Actual Problem: SOTA Chasing	79
 Part II. THE AUTOMATIC ANALYSIS OF ITALIAN	 83
6. Focus on Words	85
6.1. Italian “BERTology”	85
6.2. Morphological Analysis	86
6.2.1. <i>Italian Morphology</i>	87
6.2.2. <i>Computational tools to handle Italian Morphology</i>	88
6.2.3. <i>The AnIta Morphological Analyser</i>	88
6.2.4. <i>Evaluating AnIta Coverage</i>	94
6.3. Part-of-Speech Tagging	96
6.3.1. <i>Yes, but, which categories should we use?</i>	99
6.3.2. <i>PoS-Tagging Benchmarks for Italian</i>	101
6.3.3. <i>PoS-Tagging with Stochastic Models</i>	105
6.3.4. <i>Neural PoS-Taggers</i>	109
6.4. Lemmatisation	113
6.4.1. <i>The AnIta Lemmatiser</i>	114
6.4.2. <i>AnIta Lemmatiser Results at EVALITA 2011</i>	116
6.4.3. <i>A Neural Lemmatiser</i>	118
6.5. Named Entity Recognition	118
6.6. Word Sense Disambiguation	120
6.6.1. <i>Baseline Methods for WSD</i>	121
6.6.2. <i>A Quantum-Like Approach to Word Sense Disambiguation</i>	122
6.7. Frame Semantic Parsing	134
6.7.1. <i>Electra-AGE_FE System Architecture</i>	136
6.7.2. <i>The IFrameNet project</i>	144
6.7.3. <i>Electra-AGE_FE on Italian</i>	145

7. Sentence and Text Analysis	147
7.1. Syntactic Analysis	147
7.1.1. <i>General Parsing Strategies</i>	148
7.1.2. <i>Neural Dependency Parsers</i>	149
7.1.3. <i>Italian Treebanks</i>	155
7.1.4. <i>Neural Parsers Evaluation</i>	156
7.1.5. <i>An Ensemble of Neural Parsers</i>	159
7.1.6. <i>Transformer-based Parsing</i>	170
7.2. Two Words on Neural Machine Translation	172
7.3. Text Classification: Opinions, Polarity, Irony and Emotions	174
7.3.1. <i>Sentiment Analysis</i>	174
7.3.2. <i>Hate Speech Detection</i>	176
7.3.3. <i>Emotion Detection</i>	176
7.4. Text Generation/Summarisation	181
8. Let the Speech Speak!	189
8.1. Automatic Speech Recognition	189
8.1.1. <i>Italian Spoken Corpora for ASR</i>	190
8.1.2. <i>State of the Art for Italian ASR</i>	191
8.1.3. <i>NVIDIA NeMo ASR</i>	192
8.1.4. <i>Model Setup and Results</i>	194
8.2. Pitch Tracking	197
8.2.1. <i>A Neural PDA Smoother</i>	198
8.2.2. <i>Experiments and Results</i>	199
8.3. Prosodic Prominence Identification	202
8.3.1. <i>A Rule-based System</i>	208
8.3.2. <i>Solution based on Machine/Deep-Learning</i>	212
8.3.3. <i>A Reflection on Evaluation</i>	218
8.4. Pathological Language Analysis	220
8.4.1. <i>Linguistic Features as Digital Biomarkers</i>	221
8.4.2. <i>Mild Cognitive Impairment Detection</i>	229
8.4.3. <i>Anorexia Nervosa & Developmental Language Disorder detection</i>	239

Part III. SURFING ON “THE NEXT WAVE” 241

9. Large Language Models and Generative AI	243
9.1. Data for Pre-Training LLMs	247
9.2. Autoregressive Text Generation and Decoding	249
9.2.1. <i>From Model Predictions to Probabilities</i>	250
9.3. Major Milestones in Open-Weight LLM models	254
9.4. Evaluation of Generative AI	255
10. Elements of Prompt Engineering	265
10.1. The Basic Structure of a Prompt	265
10.2. Different techniques for building prompts	266
10.2.1. <i>Common Prompting Mistakes</i>	269

11. Retrieval Augmented Generation	271
11.1. A Common RAG Pipeline	272
11.2. Vector-based RAG	273
11.2.1. <i>From Documents to a Vector Index</i>	273
11.2.2. <i>Retrieval Beyond Nearest Neighbours</i>	274
11.2.3. <i>Context Construction</i>	274
11.2.4. <i>Strengths and Failure Modes</i>	274
11.3. GraphRAG	275
11.3.1. <i>Graph Construction from Text</i>	275
11.3.2. <i>Local and Global Retrieval</i>	275
11.3.3. <i>Advantages and Costs</i>	276
11.4. Hybrid Architectures	276
12. Agents: from Conversation to Autonomy	279
12.1. The Core Components	279
12.2. What Agents Add Over a Single LLM Interaction	280
12.3. Explaining Agents Through LangGraph Code	282
13. Summary of Machine Learning Paradigms for NLP	293
Bibliography	296
Acknowledgements	331
Index	332