

Tommi Jauhiainen · Marcos Zampieri ·
Timothy Baldwin · Krister Lindén

Automatic Language Identification in Texts

Contents

1	Introduction to Language Identification	1
1.1	A Brief History of Language Identification (LI)	4
1.2	What is LI Used For?	7
1.3	What are the Main Challenges that Make LI Difficult?	9
	References	11
2	Features and Methods	19
2.1	On Notation	19
2.2	What Textual Features Are Used for LI and How Are They Collected and Calculated?	20
2.2.1	Feature Smoothing	28
2.3	What Classification Methods Are Used for LI and How Do They Work?	30
2.3.1	Decision Rules, Trees and Random Forests	31
2.3.2	Simple Scoring	32
2.3.3	Sum or Average of Values	33
2.3.4	Product of Values	36
2.3.5	Similarity Measures	39
2.3.6	Logistic Regression	42
2.3.7	Support Vector Machines	42
2.3.8	Neural Networks	43
2.3.9	Ensemble Methods	45
2.4	Machine Learning Toolkits and Libraries	47
	References	49
3	Evaluation and Measurement	65
3.1	How is LI Performance Evaluated? What Are the Measures and How Are They Calculated?	65
3.2	What Material Can Be Used in Training and Evaluating Language Identifiers?	70

3.3	LI Shared Tasks	73
	References	86
4	Specific Challenges of Variation and Text Types	99
4.1	Language Similarity	99
4.1.1	LI for Similar Languages, Varieties, and Dialects	100
4.2	Low-Resource Languages	106
4.3	Orthography and Its Variations	107
4.4	Short Texts	108
	References	109
5	Large Scale, Multi-domain Language Identification	117
5.1	Number of Languages	117
5.2	Unseen Languages	119
5.3	Multilingual Texts	121
5.4	Domain Compatibility	125
	References	126
6	Applications and Related Tasks	137
6.1	Applications	137
6.1.1	Monolingual NLP Components	137
6.1.2	Machine Translation	138
6.1.3	Multilingual Document Storage and Retrieval	138
6.2	Related Tasks	139
6.2.1	Native Language Identification	139
6.2.2	Author Profiling and Identification	140
	References	142
7	Conclusion and Future Directions	147