

Contents

Acknowledgements	iii
Agradecimientos	iv
Abstract	vii
Resumen	ix
Resum	xi
Contents	xiii
List of Figures	xxi
List of Tables	xxv
1 Introduction	1
1.1 Motivation and Objectives	3
1.1.1 General Objectives	3
1.1.2 Specific Objectives	3
1.2 Research Questions	4
1.3 Main contributions	6
1.4 Outline of the Thesis	6
2 Plagiarism and Text Re-Use	11
2.1 Text Re-Use	12
2.1.1 Methods for Text Re-Use Commitment	13

2.2	Plagiarism	13
2.2.1	A History of Plagiarism	15
2.2.2	Plagiarism Commitment and Prevention	17
2.3	Computational Linguistics Meets Forensic Linguistics	20
2.3.1	Forensic Linguistics	20
2.3.2	(Dis)similarities between Computational and Forensic Linguistics	23
2.4	Plagiarism: An Explosion of Cases	24
2.4.1	Overview of Recent Cases of Plagiarism	25
2.4.1.1	Cases of Plagiarism in Academia	25
2.4.1.2	Cases of Plagiarism in Research	26
2.4.1.3	Cases of Plagiarism in Journalism	27
2.4.1.4	Cases of Plagiarism in Literature	28
2.4.1.5	Cases of Plagiarism in Politics	29
2.4.1.6	Cases of Plagiarism in Show Business	31
2.4.1.7	Discussion on the Explosion of Plagiarism Cases	33
2.5	Surveying Plagiarism in Academia	34
2.5.1	Overview of Surveys on Plagiarism	34
2.5.2	A New Survey on Plagiarism Attitudes	38
2.5.2.1	General information	38
2.5.2.2	Scholar Practices	39
2.5.2.3	Attitudes Respect to Plagiarism	41
2.5.2.4	Final Opinions	44
2.6	Automatic Text Re-Use and Plagiarism Detection	46
2.7	Commercial Plagiarism Detection Tools	48
2.8	Chapter Summary	52
3	Text Analysis for Re-Use and Plagiarism Detection	53
3.1	Text Representation	53
3.1.1	Pre-Processing	54
3.1.1.1	Character Normalisation	54
3.1.1.2	Tokenisation	54
3.1.1.3	Stemming and Lemmatisation	55
3.1.1.4	Sentence Identification	55

3.1.1.5	Punctuation Removal	55
3.1.1.6	Words Filtering	56
3.1.2	Bag of Words Representation	56
3.1.3	<i>n</i> -Grams	56
3.1.4	Cognates	58
3.1.5	Hash Model	59
3.2	Weighting	60
3.2.1	Boolean Weighting	60
3.2.2	Real Valued Weighting	61
3.2.2.1	Term Frequency	61
3.2.2.2	Document Frequency	62
3.3	Text Similarity	62
3.3.1	Vector Space Models	63
3.3.1.1	Boolean Models	64
3.3.1.2	Real-Valued Models	66
3.3.2	Probabilistic Models	67
3.3.2.1	Kullback-Leibler Distance	68
3.3.2.2	Machine Translation	69
3.4	Stylometric Measures	72
3.4.1	Text Statistics	72
3.4.2	Syntactic Features	74
3.4.3	Part of Speech Features	74
3.4.4	Closed-Class and Complex Words Features	74
3.5	Chapter Summary	76
4	Corpora and Evaluation Measures	77
4.1	Overview of Corpora and Evaluation Measures Exploitation	78
4.2	Corpora for Plagiarism and Text Re-Use Detection	79
4.2.1	METER Corpus	81
4.2.2	Co-derivatives Corpus	84
4.2.3	PAN-PC Corpora	87
4.2.3.1	PAN-PC Conception	87
4.2.3.2	Cases Generation	91

4.2.3.3	PAN-PC-09	92
4.2.3.4	PAN-PC-10	93
4.2.3.5	PAN-PC-11	94
4.2.3.6	Potential Future Improvements to the PAN-PC Corpora	95
4.2.4	Short Plagiarised Answers Corpus	97
4.2.5	CL!TR 2011 Corpus	99
4.3	Evaluation Metrics	101
4.3.1	Recall, Precision, and F -measure	101
4.3.2	Highest False Match and Separation	102
4.3.3	Especially Fitted Measures for Plagiarism Detection	104
4.3.3.1	Especially Fitted Recall and Precision	105
4.3.3.2	Granularity	106
4.3.3.3	Plagdet	106
4.4	Chapter Summary	107
5	Monolingual Detection of Text Re-Use and Plagiarism	109
5.1	Past Work	115
5.1.1	Approaches for Intrinsic Plagiarism Detection	115
5.1.1.1	Averaged Word Frequency Class	115
5.1.1.2	Character n -Gram Profiles	116
5.1.1.3	Kolmogorov Complexity Measures	116
5.1.1.4	Assumptions and Drawbacks	117
5.1.2	Approaches for External Plagiarism Detection	117
5.1.2.1	Detailed Analysis	117
5.1.2.2	Heuristic Retrieval	126
5.1.2.3	Knowledge-Based Post-Processing	128
5.1.2.4	Pre-processing	128
5.1.2.5	Detecting the Direction of Re-Use	129
5.2	Word n -Grams Retrieval	129
5.2.1	Experimental Setup	130
5.2.2	Results and Discussion	130
5.3	Containment-based Re-Use Detection	134
5.3.1	Experimental Setup	134

5.3.2	Results and Discussion	135
5.4	The Impact of Heuristic Retrieval	135
5.4.1	Proposed Heuristic Retrieval Model	136
5.4.1.1	Features Selection	136
5.4.1.2	Term Weighting	137
5.4.2	Experimental Setup	138
5.4.3	Results and Discussion	139
5.5	Chapter Summary	141
6	Cross-Language Detection of Text Re-Use and Plagiarism	143
6.1	Cross-Language Plagiarism Detection Process	144
6.1.1	Cross-Language Heuristic Retrieval	145
6.1.2	Cross-Language Detailed Analysis	145
6.2	Past Work	145
6.2.1	Intrinsic Cross-Language Plagiarism Detection	145
6.2.2	External Cross-Language Plagiarism Detection	147
6.2.2.1	Models based on Syntax	147
6.2.2.2	Models based on Thesauri	148
6.2.2.3	Models based on Comparable Corpora	149
6.2.2.4	Models based on Parallel Corpora	150
6.2.2.5	Models based on Machine Translation	150
6.3	Cross-Language Alignment-based Similarity Analysis	152
6.4	Document Level Cross-Language Similarity Experiments	153
6.4.1	Corpora for Model Training and Evaluation	154
6.4.2	Experimental Setup	154
6.4.3	Results and Discussion	155
6.5	Sentence Level Detection across Distant Languages	158
6.5.1	Experimental Setup	160
6.5.2	Results and Discussion	161
6.6	Chapter Summary	162
7	PAN International Competition on Plagiarism Detection	165
7.1	PAN @ SEPLN 2009	166
7.1.1	Tasks Overview	167

7.1.1.1	External Detection	167
7.1.1.2	Intrinsic Detection	170
7.1.2	Results and Discussion	171
7.2	PAN @ CLEF 2010	173
7.2.1	Tasks Overview	173
7.2.1.1	External Detection	173
7.2.1.2	Intrinsic Detection	175
7.2.2	Results and Discussion	176
7.2.2.1	External Detection	177
7.2.2.2	Intrinsic Detection	183
7.3	PAN @ CLEF 2011	183
7.3.1	Tasks Overview	184
7.3.1.1	External Detection	184
7.3.1.2	Intrinsic Detection	185
7.3.2	Results and Discussion	187
7.3.2.1	External Detection	187
7.3.2.2	Intrinsic Detection	189
7.3.2.3	Temporal Insights	189
7.4	Detection of Monolingual Plagiarism @ PAN	191
7.4.1	Results and Discussion	194
7.5	Detection of Cross-Language Plagiarism @ PAN	196
7.5.1	Cross-Language Detection Strategy	197
7.5.2	Experimental Setup	199
7.5.3	Results and Discussion	199
7.6	Chapter Summary	205
8	Plagiarism meets Paraphrasing	207
8.1	Paraphrase Typology	209
8.1.1	Morphology-based Changes	211
8.1.2	Lexicon-based Changes	211
8.1.3	Syntax-based Changes	213
8.1.4	Discourse-based Changes	213
8.1.5	Miscellaneous Changes	214

8.1.6	Semantics-based Changes	215
8.2	Building the P4P Corpus	215
8.3	Analysis of Paraphrase Plagiarism Detection	218
8.3.1	Clustering Similar Cases of Plagiarism in the P4P Corpus	219
8.3.2	Results and Discussion	222
8.4	Chapter Summary	226
9	Detection of Text Re-Use in Wikipedia	227
9.1	Related Work over Wikipedia	228
9.1.1	Monolingual Analysis	228
9.1.2	Analysis across Languages	229
9.2	Monolingual Co-Derivation in Wikipedia	230
9.2.1	Experimental Settings	231
9.2.2	Results and Discussion	233
9.3	Similarity of Wikipedia Articles across Languages	237
9.3.1	Experimental Settings	238
9.3.2	Results and Discussion	239
9.4	Extracting Parallel Fragments from Wikipedia	240
9.4.1	Model Description	241
9.4.2	Parallel and Comparable Corpora	241
9.4.3	Experimental Settings	242
9.4.4	Results and Discussion	243
9.5	PAN@FIRE: Cross-Language Indian Text Re-use	245
9.5.1	Proposed Task	245
9.5.2	Submissions Overview	246
9.5.3	Results and Discussion	248
9.6	Chapter Summary	250
10	Conclusions	253
10.1	Contributions	254
10.2	Research Answers	254
10.3	Future Trends of Research	257
	References	261

A	Generation of Dictionaries for CL-ASA	291
A.1	Dictionary Built from Parallel Corpora	291
A.2	Dictionary Built from Lexicographic Data	293
B	Related Publications	299
B.1	Journals	299
B.2	Conferences	300
B.3	Book Chapters	301
B.4	Workshops	302
C	Media Coverage	305
C.1	News	305
C.2	On Air and TV	306
	Index	309