

Contents

Title page	i
Acknowledgments	iii
Dedication	v
Abstract	vii
Table of contents	xv
List of tables	xix
List of figures	xxi
Notation	xxv
1 Introduction	1
1.1 Short texts	3
1.2 Narrow domain corpora	4
1.3 Narrow domain short-text corpora	5
1.4 Scientific abstracts	6
1.5 Challenges: corpora evaluation, clustering & validity	7
1.6 Thesis contributions	8
1.7 Thesis overview	10
2 Methods, techniques and datasets	13
2.1 Clustering methods	13
2.2 Term selection techniques	27
2.3 Datasets	30
3 Clustering narrow domain short-text corpora	47
3.1 Clustering vs. categorization	51
3.2 The clustering hypothesis	52
3.3 Related work	53
3.4 Experimental results	58
3.5 Concluding remarks	72
4 Evaluation of narrow domain short-text corpora	77
4.1 Domain broadness evaluation measures	80

4.2	Stylometry-based evaluation measure	90
4.3	Shortness-based evaluation measures	91
4.4	Class imbalance degree assessment measure	92
4.5	Structure-based evaluation measures	94
4.6	Experimental results	95
4.7	WaCOS: The Watermarking Corpus On-line System	109
4.8	Concluding remarks	114
5	The self-term expansion methodology	115
5.1	Term expansion using external knowledge	116
5.2	The self-term expansion technique	118
5.3	Term selection	120
5.4	Experimental results	121
5.5	Concluding remarks	130
6	Word sense induction	133
6.1	Peculiarities of the <i>WSI-SemEval</i> data collection	137
6.2	The proposed word sense induction system	138
6.3	Experimental results	140
6.4	Concluding remarks	152
7	Evaluation of clustering validity measures in short-text corpora	155
7.1	Correlation between internal and external clustering validity measures	156
7.2	The relative hardness of clustering corpora	164
7.3	Concluding remarks	168
8	Conclusions and further work	171
8.1	Findings and research directions	171
8.2	Major contributions	178
8.3	Further work	178
	Bibliography	181
A	Other external clustering validity measures	199
A.1	Pairwise Precision/Recall/Accuracy	199
A.2	MUC Precision/Recall	200
A.3	B-Cubed Precision/Recall	201
A.4	Purity/Inverse Purity	201
A.5	F-Purity/F-Inverse Purity	202

B	The specific behaviour of the evaluation measures	205
B.1	The <i>CICLing-2002</i> corpus	205
B.2	The <i>hep-ex</i> corpus	209
B.3	The <i>WebKB train</i> corpus	212
B.4	The <i>WebKB test</i> corpus	215
B.5	The <i>R8-Reuters train</i> corpus	218
B.6	The <i>R8-Reuters test</i> corpus	221
B.7	The <i>R52-Reuters train</i> corpus	224
B.8	The <i>R52-Reuters test</i> corpus	227
B.9	The <i>20 Newsgroups train</i> corpus	230
B.10	The <i>20 Newsgroups test</i> corpus	233
C	Word by word analysis in the WSI-SemEval data collection	237