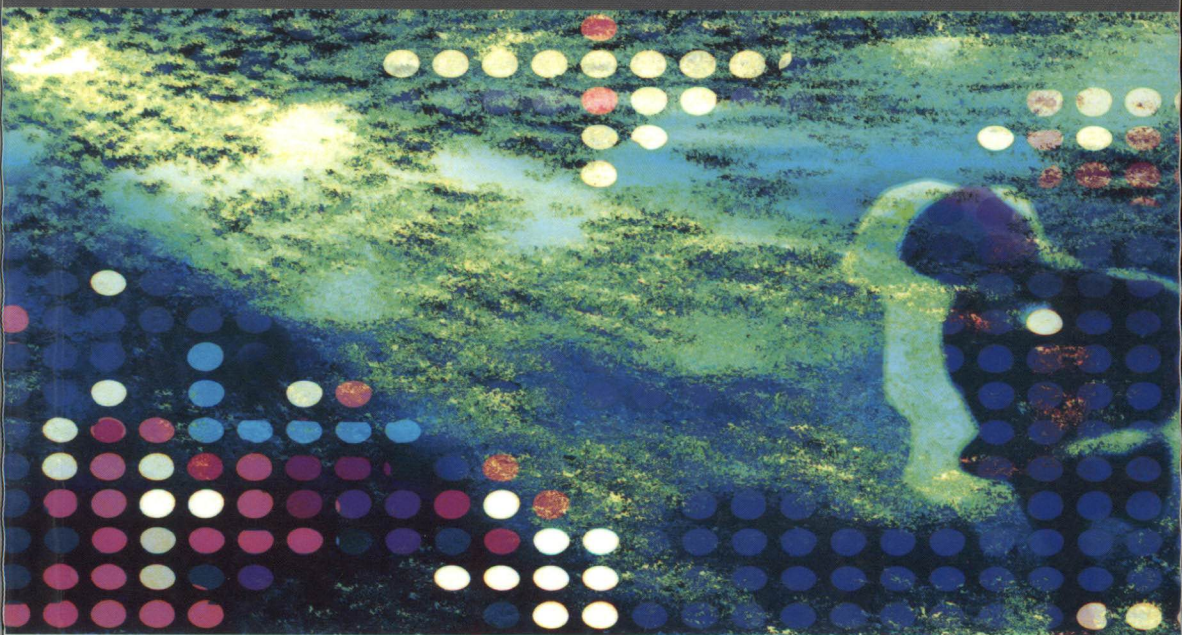




Expérimentations et évaluations en fouille de textes

un panorama des campagnes DEFT

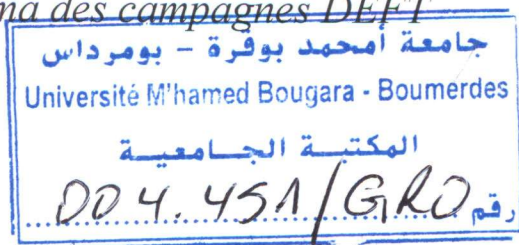
sous la direction de

Cyril Grouin
Dominic Forest



Expérimentations et évaluations en fouille de textes

un panorama des campagnes DEFT



Alaw.

sous la direction de
Cyril Grouin
Dominic Forest



hermes
Science
— publications —

Lavoisier

Table des matières

PREMIÈRE PARTIE. INTRODUCTION	21
Chapitre 1. Expérimentations et évaluations en fouille de textes : un panorama des campagnes DEFT	23
Cyril GROUIN et Dominic FOREST	
1.1. Introduction	23
1.2. La fouille de textes	25
1.3. Les campagnes d'évaluation	26
1.3.1. Présentation générale	26
1.3.2. Les campagnes en fouille de textes	28
1.3.3. Le Défi fouille de textes	28
1.4. Mesures d'évaluation	30
1.4.1. Rappel, précision, F-mesure	30
1.4.2. Macro et micro-moyennes	32
1.5. Les campagnes en genre et thèmes	33
1.5.1. L'édition 2006 : les ruptures thématiques	34
1.5.2. L'édition 2008 : l'identification en genres et en thèmes	34
1.6. Les campagnes en fouille d'opinions	36
1.6.1. L'édition 2007 : fouille d'opinions sur des avis argumentés	36
1.6.2. L'édition 2009, distinction objectivité/subjectivité multilingue	37
1.7. Les campagnes diachroniques	39
1.7.1. L'édition 2010 : variations diachroniques et géographiques	40
1.7.2. L'édition 2011 : variation diachronique	44

1.7.3. Méthodes d'évaluation	47
1.8. Conclusion	49
1.9. Remerciements	51
1.10. Bibliographie	52
 DEUXIÈME PARTIE. LES CAMPAGNES EN GENRES	
ET THÈMES	55
 Chapitre 2. Détecter les ruptures thématiques dans les discours : synergie entre supervision et non-supervision	
57	
Alain LELU et Martine CADOT	
2.1. Introduction : la tâche et sa difficulté	57
2.2. Principes mis en œuvre	58
2.3. Outils et environnements de travail utilisés	60
2.4. Réalisation des trois angles d'attaque	60
2.4.1. Hypothèses et vérifications	61
2.4.2. Les débuts de phrases : aspects formels et « attaques »	64
2.4.3. Aspects rhétoriques : classification non supervisée des phrases à partir de leur indexation « plein texte »	65
2.4.4. Aspects sémantiques	66
2.5. Apprentissage final sur les 5 indicateurs synthétisant nos points de vues	70
2.6. Conclusion	72
2.7. Bibliographie	73
 Chapitre 3. L'équipe du GRDS au DEFT2006 : Indexo-II	
75	
Lyne DA SYLVA, Graham RUSSELL, Yves MARCOUX et Frédéric DOLL	
3.1. Introduction	75
3.2. Travaux antérieurs	76
3.2.1. Segmentation par cohésion lexicale	76
3.2.2. Apprentissage automatique	77
3.3. Approche utilisée	78
3.3.1. Introduction	78
3.3.2. Prétraitement	78
3.3.3. Termes et pondération des termes	78
3.3.4. Cohésion lexicale	79

3.3.5. Chaînes lexicales	80
3.3.6. Taux d'introduction du vocabulaire	81
3.3.7. Entraînement du système	82
3.4. Contrastes entre diverses approches à la segmentation	83
3.4.1. Segmentation pour l'indexation de monographies	83
3.4.2. Segmentation pour la tâche de DEFT2006	89
3.5. Evaluation	90
3.5.1. Résultats obtenus	90
3.5.2. Caractéristiques du corpus et performance du système	90
3.5.3. Difficulté de la tâche d'évaluation	92
3.6. Conclusion	93
3.7. Bibliographie	94

**Chapitre 4. Pré-traitements classiques ou par analyse
distributionnelle : application aux méthodes de classification
automatique déployées pour DEFT2008**

97

Eric CHARTON, Nathalie CAMELIN, Rodrigo ACUNA-AGOST,
Rémi LAVALLEY, Rémy KESSLER et Silvia FERNANDEZ

4.1. Introduction	97
4.2. Caractéristiques de la tâche et des corpus	98
4.2.1. Commentaires sur le corpus de la tâche 1	98
4.2.2. Commentaires sur le corpus de la tâche 2	100
4.3. Approches envisagées	101
4.3.1. Méthode de validation des classifieurs	101
4.3.2. Méthodes de classification retenues	101
4.4. Implémentation des différents systèmes de classification	103
4.4.1. Méthodes de pré-traitement et normalisation des textes	103
4.4.2. Présentation des systèmes de classification implémentés pour DEFT2008	106
4.4.3. Stratégie de fusion	109
4.5. Expériences et résultats	110
4.5.1. Evaluations pendant l'apprentissage	111
4.5.2. Soumission 1. Fusion par vote ternaire majoritaire	111
4.5.3. Soumission 3. Utilisation des meilleurs systèmes	112
4.6. Conclusion et perspectives	112
4.7. Bibliographie	113

TROISIÈME PARTIE. LES CAMPAGNES EN FOUILLE

D'OPINIONS	115
----------------------	-----

Chapitre 5. Classification d'opinions et convergence des techniques symboliques, statistiques et distributionnelles

Luca DINI, Sigrid MAUREL, Paolo CURTONI et Beata DOBRZYŃSKA	117
---	-----

5.1. Introduction	117
5.2. Les corpus	118
5.3. Méthode symbolique	119
5.3.1. Grammaire	120
5.3.2. Lexique de sentiments	122
5.3.3. Grammaires pour DEFT2007	123
5.3.4. Listes de termes	125
5.4. Méthode statistique	126
5.5. Méthode hybride	127
5.6. Conclusion à propos de DEFT2007	128
5.7. Travaux depuis DEFT2007	130
5.7.1. LOR	130
5.7.2. RI	131
5.7.3. Comparaison de LOR et RI sur corpus	133
5.8. Bibliographie	134

Chapitre 6. DEFT2009 : essais d'optimisation d'une procédure de base pour la tâche 1

Yves BESTGEN	135
6.1. Introduction	135
6.2. Pré-traitement et analyse exploratoire	136
6.3. Essais d'optimisation du classifieur pour le corpus français	140
6.3.1. Paramètres classiques	140
6.3.2. Prise en compte de la longueur des documents	142
6.4. Analyse des effets spécifiques des paramètres	144
6.5. Résultats de la phase de test et analyses complémentaires	147
6.6. Conclusion	149
6.7. Bibliographie	150

Chapitre 7. Détection de la subjectivité et catégorisation de textes subjectifs par une approche mixte symbolique et statistique 153

Matthieu VERNIER, Laura MONCEAUX et Béatrice DAILLE

7.1. Introduction	153
7.2. Contexte motivant la participation à DEFT2009	155
7.2.1. Travaux reliés et tâches réalisées pour DEFT2009	155
7.2.2. Qu'est-ce que la subjectivité ?	156
7.3. Tâche 1 : catégorisation de textes Objectif/Subjectif	158
7.3.1. Choix des descripteurs	158
7.3.2. Mise en œuvre informatique	163
7.4. Tâche 2 : détection des passages subjectifs	168
7.4.1. Notre outil de détection des évaluations dans les blogs	168
7.4.2. Adaptation de notre outil et résultats	171
7.5. Conclusion	173
7.6. Bibliographie	174

QUATRIÈME PARTIE. LES CAMPAGNES DIACHRONIQUES 175

Chapitre 8. Datation d'un article de journal par analyse lexicale et statistique 177

Pierre ALBERT, Flora BADIN, Maxime DELORME, Nadège DEVOS,
Sophie PAPAZOGLOU et Jean SIMARD

8.1. Introduction	177
8.2. Corpus	178
8.2.1. Le corpus d'entraînement	178
8.2.2. Caractéristiques du corpus	179
8.3. Les méthodes	180
8.3.1. Les réformes de l'orthographe	180
8.3.2. Les entités nommées	182
8.3.3. Apprentissage par l'utilisation de <i>Conditional Random Fields</i>	183
8.3.4. Récurrence des termes	186
8.3.5. Fusion des résultats	187
8.4. Soumission	188
8.5. Conclusion	191
8.6. Bibliographie	192

Chapitre 9. Système du LIA pour la campagne DEFT2010	195
Stanislas OGER, Mickael ROUVIER, Nathalie CAMELIN, Rémy KESSLER, Fabrice LEFÈVRE et Juan-Manuel TORRES-MORENO	
9.1. Introduction	195
9.2. Présentation des tâches et des corpus	196
9.2.1. Variations diachroniques	196
9.2.2. Origine géographique	197
9.3. Présentation des systèmes initiaux	198
9.3.1. Protocole expérimental	198
9.3.2. Systèmes extra-linguistiques : <i>Mick_MLP</i> et <i>Mick_SVM</i>	199
9.3.3. Modèle de langage à base de n -grammes de caractères : <i>Jmt</i> et <i>Jmt_basic</i>	200
9.3.4. Systèmes à base de boosting : <i>Boost_basic</i> et <i>Rk_icsiboost</i>	202
9.3.5. Système hybride pour résoudre la tâche <i>Décennies</i>	202
9.4. Systèmes de fusion	204
9.4.1. Fusion majoritaire	204
9.4.2. Fusion par classification	204
9.5. Evaluation	206
9.5.1. Evaluation stricte	206
9.5.2. Indice de confiance pondéré	206
9.5.3. Discussion	207
9.6. Résultats	207
9.7. Conclusion	210
9.8. Bibliographie	210

Chapitre 10. Apprentissage supervisé et paresseux pour la fouille de textes	213
Christian RAYMOND et Vincent CLAVEAU	

10.1. Introduction	213
10.2. Contribution à la tâche 1 : approches avec modèle	214
10.2.1. Pré-traitement des données	214
10.2.2. Premier aperçu avec un arbre de décision	216
10.2.3. Un peu plus de souplesse	220
10.2.4. Conclusion	222
10.3. Contribution à la tâche 1 : k -plus proches voisins	224
10.3.1. Vision de la tâche : <i>lazy-learning</i>	224

10.3.2. Mesures de similarité	226
10.3.3. Procédure de vote	227
10.3.4. Résultats	228
10.4. Contribution à la tâche 2 : appariement résumé/article	228
10.4.1. Vision de la tâche	229
10.4.2. Résultats	229
10.5. Conclusion	229
10.6. Bibliographie	230
 Chapitre 11. Méthodes pour l'archéologie linguistique :	
datation par combinaison d'indices temporels	231
Anne GARCÍA-FERNANDEZ, Anne-Laure LIGOZAT, Delphine BERNHARD et Marco DINARELLI	
11.1. Introduction	231
11.2. Corpus et prétraitements	232
11.2.1. Découpage du corpus d'entraînement	233
11.2.2. Lemmatisation	233
11.2.3. Correction orthographique	233
11.3. Indices chronologiques	234
11.3.1. Dates de naissance de personnes	234
11.3.2. Néologismes et archaïsmes	235
11.3.3. Réformes orthographiques	237
11.4. Système de catégorisation temporelle	238
11.4.1. Système SVM	239
11.4.2. Paramétrage	239
11.4.3. Attributs de classification	240
11.4.4. Intégration des indices chronologiques	240
11.5. Expériences et résultats	241
11.5.1. Métrique d'évaluation	241
11.5.2. Résultats	242
11.6. Conclusion	244
11.7. Remerciements	244
11.8. Bibliographie	245
 Index	 247

La fouille de textes est une activité combinant traitements informatiques et données linguistiques avec comme objectif principal l'extraction et l'organisation automatique des informations présentes dans les textes. Deux familles de méthodes permettent d'atteindre ce but : celles à base de connaissances d'experts et celles reposant sur un apprentissage automatique supervisé.

Une campagne d'évaluation consiste à confronter les systèmes développés par plusieurs équipes sur un même jeu de données et en un temps limité. Créé en 2005 à l'image des campagnes anglo-saxonnes, le défi fouille de textes (DEFT) est aujourd'hui la seule campagne d'évaluation francophone en fouille de textes.

Cet ouvrage rassemble les méthodes utilisées lors des différentes éditions du défi. Les thématiques relèvent de la classification de documents en genres et thèmes, de la fouille d'opinions et de l'identification de la période de parution d'un document.

Les coordonnateurs

Ingénieur d'Etudes CNRS au LIMSI, Cyril Grouin travaille sur l'anonymisation de documents cliniques et sur les entités nommées étendues. Il coorganise les campagnes d'évaluation DEFT sur la fouille de textes.

Dominic Forest est professeur agrégé à l'Ecole de bibliothéconomie et des sciences de l'information de l'Université de Montréal. Ses recherches portent sur la fouille de textes et les humanités numériques.