Table des matières

Introduction 9

Chapitre 1. Les entités nommées pour l'accès à l'information 11

1.1. Historique des programmes de recherche 12
1.1.1. La compréhension de documents : une tâche ambitieuse 12
1.1.2. Détecter des éléments de base : les entités nommées 13
1.1.3. Tendance : un retour au remplissage de formulaire 16
1.2. Quelles tâches pour utiliser les entités nommées comme « pivots » . . . 18
1.3. Conclusion 19

Chapitre 2. Les entités nommées, des unités référentielles 21

2.1. La problématique des entités nommées 22
2.1.1. Un ensemble hétérogène 22
2.1.1.1. Multiplicité des catégories 23
2.1.1.2. Diversité des mentions 24
2.1.2. Les formules définitoires existantes 26
2.1.3. Un objet TAL 30
2.2. Les notions de sens et de référence 30
2.2.1. Qu'est-ce que la référence ? 31
2.2.2. Qu'est-ce que le sens ? 32
2.3. Noms propres 35
2.3.1. Les critères traditionnels de définition du nom propre 36
2.3.2. Sens et fonctionnement référentiel du nom propre. 38
2.3.3. La « charge référentielle » du nom propre 41
2.4. Descriptions définies 42
2.4.1. Qu'est-ce qu'une *description définie* ? 42
2.4.2. Le sens des descriptions définies 45

2.4.3. Descriptions définies complètes et incomplètes 45
2.5. Sens et fonctionnement référentiel des entités nommées 47
2.5.1. Référence à un particulier 48
2.5.1.1. Le principe d'individuation 48
2.5.1.2. L'unicité référentielle 49
2.5.2. Autonomie référentielle 50
2.5.3. Une hétérogénéité « naturelle » 51
2.6. Conclusion 51

Chapitre 3. Ressources autour des entités nommées 53

3.1. Typologies du domaine général et en domaines de spécialité 54
3.1.1. La notion de catégorie 54
3.1.2. Évolution des typologies 55
3.1.3. Vers une structuration des entités 58
3.1.4. Et en dehors des campagnes d'évaluation ? 60
3.1.5. Petite comparaison illustrée 62
3.1.6. Différentes questions autour des entités 62
3.2. Corpus 64
3.2.1. Introduction 64
3.2.2. Corpus et entités nommées 65
3.2.2.1. Corpus issus des campagnes MUC et ACE 65
3.2.2.2. Corpus issus des campagnes françaises 66
3.2.2.3. Corpus issu de la campagne GermEval 68
3.2.2.4. Corpus issu de la campagne Evalita 68
3.2.2.5. Corpus issu de la campagne Harem 69
3.2.3. Conclusion 69
3.3. Lexiques et bases de données de connaissances 70
3.3.1. Bases lexicales 70
3.3.1.1. ANNIE 71
3.3.1.2. WordNet 71
3.3.1.3. Prolex 72
3.3.1.4. Geonames 72
3.3.1.5. JRC-Names 73
3.3.1.6. Dans le domaine biomédical 74
3.3.1.7. Conclusion 75
3.3.2. Les bases de connaissances 76
3.4. Conclusion 78

Chapitre 4. Reconnaître les entités nommées 81

4.1. Détection et classification des entités nommmées 82
4.2. Indices pour reconnaître les entités nommées 83
4.2.1. Décrire la morphologie des mots 83

4.2.2. Exploiter les bases lexicales 85
4.2.3. Indices contextuels 87
4.2.4. Conclusion 88
4.3. Techniques à base d'automates 88
4.4. Modèles guidés par les données et apprentissage 91
4.4.1. Modèles par classes majoritaires 93
4.4.2. Modèles à décisions contextuelles (HMM) 94
4.4.3. Modèles utilisant des indices multiples (softmax, MaxEnt) 95
4.4.4. Champs markoviens conditionnels (CRF) 97
4.5. Enrichissement non supervisé de méthodes supervisées 97
4.6. Conclusion 98

Chapitre 5. Lier les entités nommées aux référentiels 101

5.1. Les bases de références 102
5.2. Formalisation de la polysémie des mentions d'entités nommées 103
5.3. Étapes du processus de liaison des entités nommées 104
5.3.1. Recherche des mentions d'entités nommées 104
5.3.2. Sélection des candidats pour chaque mention 105
5.3.3. Désambiguïsation du référent 106
5.3.4. Liaison des entités 107
5.4. Performances des systèmes 107
5.4.1. Cas d'application, DBpedia Spotlight 108
5.4.2. Perspectives 109

Chapitre 6. Évaluation de la reconnaissance des entités nommées 111

6.1. Les mesures classiques : précision, rappel et F-mesure 112
6.2. Les mesures fondées sur un décompte des types d'erreurs 114
6.3. Évaluation des tâches connexes 119
6.3.1. Détection d'entités et mentions 120
6.3.2. Détection d'entités et liaison d'entités 121
6.4. Évaluation des technologies appliquées en amont 124
6.5. Conclusion 125

Conclusion 129

Annexe 1. Glossaire 133

Annexe 2. Récapitulatif des programmes de recherche sur les entités nommées 135

Annexe 3. Récapitulatif des corpus disponibles 141

Annexe 4. Formats d'annotation 147

Annexe 5. Entités nommées : les formules définitoires existantes 149

Bibliographie 153

Index 165