

# Extraction automatique d'information

*du texte brut au web sémantique*

Thierry Poibeau

BIBLIOTHEQUE DU CERIST

 **Hermès**

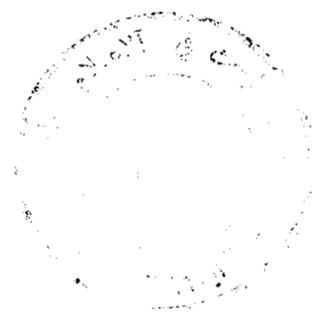
*Lavoisier*

---

IST 2775

BIBLIOTHEQUE DU CERIST

Extraction automatique d'information



## REMERCIEMENTS

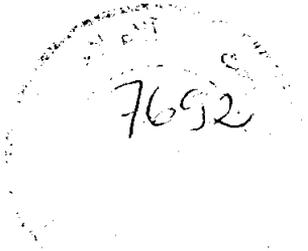
A l'instar de Panurge, je fais partie des « débiteurs et emprunteurs ». Je dois beaucoup à ceux qui ont contribué, d'une façon ou d'une autre, à ce travail. Que tous en soient remerciés : ceux dont les noms figurent ici, bien sûr, mais aussi tous les autres, ceux qui ont enrichi ma réflexion, parfois à leur insu, lors de séminaires ou de conférences.

Daniel Kayser et Adeline Nazarenko ont guidé et encadré ces travaux. Je leur sais gré de leur attentive bienveillance, de leurs conseils avisés et de leur constante amitié. Christian Fluhr, Christian Jacquemin, Patrick Saint-Dizier et Yorick Wilks m'ont apporté des conseils et des remarques qui me permettent de prolonger la réflexion sur des voies nouvelles, auxquelles je n'aurais pas songé.

Je remercie Célestin Sedogbo pour m'avoir permis de mener à bien ces travaux à Thales Recherche et Technologie. Mes collègues ont constamment nourri ma réflexion : il s'agit notamment d'Antonio Balvet, Nicolas Farcet, Bénédicte Goujon, Claire Laudy, Frédéric Meunier et Nathalie Richardet.

Cette étude a bénéficié de nombreuses collaborations, formelles ou informelles. J'ai notamment bénéficié de l'aide et des conseils de Florence Amardeilh, Sophie Bizouard, Nathalie Colineau, Anne Dister, Dominique Dutoit, David Faure, Natalia Grabar, Tristelle Kervel, Leïla Kosseim, Claire Laudy, Claire Nédellec, David Roussel, Max Silberstein, Monique Slodzian et Axelle Vinckx.

J'ai enfin une pensée pour les membres du LIPN et du CRIM, pour mes étudiants (notamment ceux de l'INaLCO), ma famille et mes amis qui ont contribué à rendre ces années de recherche agréables.



© LAVOISIER, 2003

LAVOISIER

11, rue Lavoisier

75008 Paris

Serveur web : [www.hermes-science.com](http://www.hermes-science.com)

ISBN 2-7462-0610-2

---

Le Code de la propriété intellectuelle n'autorisant, aux termes de l'article L. 122-5, d'une part, que les « copies ou reproductions strictement réservées à l'usage privé du copiste et non destinées à une utilisation collective » et, d'autre part, que les analyses et les courtes citations dans un but d'exemple et d'illustration, « toute représentation ou reproduction intégrale, ou partielle, faite sans le consentement de l'auteur ou de ses ayants droit ou ayants cause, est illicite » (article L. 122-4). Cette représentation ou reproduction, par quelque procédé que ce soit, constituerait donc une contrefaçon sanctionnée par les articles L. 335-2 et suivants du Code de la propriété intellectuelle.

BIBLIOTHEQUE DU CERIST

# Extraction automatique d'information

*du texte brut au web sémantique*

Thierry Poibeau



hermes  
Science  
— publications —

EXTRAIT DU CATALOGUE GÉNÉRAL

- Le document audiovisuel – *procédures de description et exploitations pratiques*, Peter STOCKINGER, 2003
- Filtrage sémantique – *du résumé automatique à la fouille de texte*, Jean-Luc MINEL, 2002.
- Multilinguisme et traitement de l'information, Frédérique SEGOND (dir.), 2002.
- Structuration de terminologie, T. HAMON, A. NAZARENKO (dir.), 2002.
- Espaces numériques d'information et de coopération, C. SIMONE, N. MATTA, B. EYNARD (dir.), 2001.
- IHM et recherche d'information, Céline PAGANELLI (dir.), 2001.
- Ingénierie des langues, Jean-Marie PIERREL (dir.), 2001.
- Lexiques sémantiques, P. BOUILLON, E. VIEGAS (dir.), 2001.
- Logique pour le traitement de la langue naturelle – *application à la langue française*, Philippe DELSARTE, André THAYSE, 2001.
- Réingénierie des données et des documents pour le web, Jacques KOULOUMDJIAN (dir.), 2001.
- Traitement et contrôle de l'information, Peter STOCKINGER, 2001.
- L'archivage, Anne-Marie CHABIN (dir.), 2000.
- L'indexation, Jean-Michel JOLION (dir.), 2000.
- La publication en ligne, coordonnateurs, Charlotte NIKITENKO, Peter STOCKINGER (dir.), 2000.
- La recherche d'informations – *du texte intégral au thésaurus*, Philippe LEFÈVRE, 2000.
- Le management de l'archive, Anne-Marie CHABIN, 2000.
- Traitement automatique des langues pour la recherche d'information, Christian JACQUEMIN, 2000.
- Traitement automatique des noms propres, Denis MAUREL, Frantz GEUTHNER (dir.), 2000.
- Gestion des documents et gestion des connaissances, Gérard DUPOIRIER, Jean-Louis ERMINE (dir.), 1999.
- La recherche intelligente sur l'internet – *outils et méthodes*, 2<sup>e</sup> édition revue et augmentée, Henry SAMIER, Victor SANDOVAL, 1999.
- Les nouveaux produits d'information – *conception et sémiotique du document*, Peter STOCKINGER, 1999.

## Table des matières

<b>Préface</b> . . . . .	11
<b>Introduction</b> . . . . .	13
<b>Chapitre 1. Des systèmes de compréhension de textes aux systèmes d'extraction d'information</b> . . . . .	17
1.1. Les systèmes génériques de compréhension de textes : une approche trop ambitieuse . . . . .	17
1.1.1. Une approche ambitieuse de la compréhension . . . . .	18
1.1.2. Les limites de cette approche . . . . .	20
1.1.3. Des expériences instructives . . . . .	21
1.2. L'extraction d'information pour une compréhension locale . . . . .	22
1.2.1. Le renouveau des travaux en matière de compréhension de textes . . . . .	23
1.2.2. De la compréhension à l'extraction d'information . . . . .	23
1.2.3. Une approche guidée par le but . . . . .	25
1.2.4. Une approche locale . . . . .	27
1.3. Quelle généricité et quelle adaptabilité pour les systèmes d'extraction ? . . . . .	30
1.3.1. Des bases de patrons d'extraction très spécialisées . . . . .	30
1.3.2. Une technologie mature mais trop coûteuse . . . . .	31
1.3.3. L'émergence de modules réutilisables . . . . .	32
1.3.4. Le renouveau du web sémantique . . . . .	34
<b>Chapitre 2. Stratégies pour l'acquisition semi-automatique de ressources pour l'extraction</b> . . . . .	35
2.1. Un essai de classement des techniques d'apprentissage pour l'acquisition de ressources . . . . .	35
2.2. Apprendre à partir de données annotées . . . . .	37

## 6 Extraction automatique d'information

2.2.1. L'approche de Riloff . . . . .	37
2.2.2. D'autres expériences à base de corpus annotés. . . . .	38
2.2.3. Se fonder sur une base d'exemples . . . . .	39
2.2.4. Grandeur et décadence de l'annotation . . . . .	40
2.3. Limiter le volume de données à annoter . . . . .	41
2.3.1. Amorçage et coapprentissage . . . . .	41
2.3.2. Apprentissage par l'exemple. . . . .	42
2.3.3. Sélection automatique d'exemples à présenter au développeur ( <i>active learning</i> ). . . . .	43
2.3.4. Filtrage et repérage de portions de textes pertinentes . . . . .	44
2.3.5. Expansion sémantique d'une base de patrons existants . . . . .	45
2.4. Par-delà la diversité des expériences, des éléments communs . . . . .	47
2.4.1. Un appauvrissement de la tâche partiellement compensé par de nouveaux cadres d'application . . . . .	48
2.4.2. Vers un schéma opérationnel . . . . .	49

### Chapitre 3. Vers une mise en œuvre opérationnelle de l'extraction d'information . . . . .

3.1. Différentes applications, différents besoins . . . . .	51
3.1.1. Analyse d'un fonds documentaire de veille technologique. . . . .	52
3.1.2. Analyse de bases de données textuelles en génomique . . . . .	53
3.1.3. Analyse de courrier pour le support en ligne . . . . .	55
3.1.4. Analyse d'un fil d'agence de presse . . . . .	56
3.1.5. Synthèse sur les applications . . . . .	56
3.2. Cadre des expériences menées . . . . .	57
3.2.1. Corpus ayant servi de support aux expériences. . . . .	58
3.2.2. Ressources et outils utilisés . . . . .	61
3.2.2.1. Le système INTEX et la technologie à nombre fini d'états . . . . .	61
3.2.2.2. Les dictionnaires électroniques . . . . .	63
3.3. Techniques d'évaluation. . . . .	64
3.3.1. Techniques d'évaluation objectives : des métriques pour l'évaluation. . . . .	64
3.3.2. Techniques d'évaluation subjectives : mesurer l'ergonomie et l'utilisabilité des systèmes . . . . .	65

### Chapitre 4. SEMTEX : architecture du système et cadre applicatif. . . . .

4.1. Etude des besoins et de l'existant : diversité des contextes d'utilisation . . . . .	67
4.1.1. Préanalyse des textes . . . . .	68
4.1.2. La définition de la tâche : une étape difficile mais primordiale . . . . .	70
4.1.2.1. La mise au point du formulaire d'extraction . . . . .	70
4.1.2.2. Des modes de représentation plus complexes . . . . .	72

4.1.3. Les contraintes d'utilisabilité : cerner les attentes et le rôle de l'utilisateur . . . . .	73
4.1.3.1. Qui utilise le système ? . . . . .	74
4.1.3.2. Quelles sont les interactions entre l'utilisateur et le système ? . . . . .	75
4.1.3.3. Quels services le système doit-il fournir ? . . . . .	77
4.1.4. Disponibilité, élaboration et utilisation de corpus annotés . . . . .	78
4.2. Cadre applicatif . . . . .	81
4.2.1. Domaines et classes d'applications envisagés . . . . .	81
4.2.2. Généricité et adaptabilité du système visé. . . . .	82
4.3. Architecture du système. . . . .	82

## **Chapitre 5. Le repérage d'entités nommées : une approche à base de connaissances hybrides. . . . .**

5.1. Travaux antérieurs pour le repérage des entités nommées. . . . .	88
5.1.1. Travaux menés dans le cadre des conférences MUC . . . . .	88
5.1.2. Trois types de systèmes. . . . .	89
5.1.3. Quelle technique pour quel niveau de performance ? . . . . .	90
5.1.4. Corpus abordés dans ce chapitre . . . . .	91
5.2. L'apport des techniques d'apprentissage pour la reconnaissance des entités nommées. . . . .	92
5.2.1. Limites des approches fondées sur des corpus annotés . . . . .	92
5.2.2. Apport des techniques d'apprentissage sans corpus annoté en phase d'acquisition . . . . .	93
5.3. Un module de repérage des entités nommées à base de grammaires locales. . . . .	96
5.3.1. Description du module de reconnaissance des entités nommées . . . . .	96
5.3.1.1. Principes de fonctionnement. . . . .	96
5.3.1.2. Un exemple . . . . .	97
5.3.1.3. Evaluation. . . . .	98
5.3.2. Limites des bases de règles face à la diversité des corpus . . . . .	99
5.3.2.1. Expériences sur des corpus non journalistiques . . . . .	99
5.3.2.2. Constat initial : une chute de performances importante. . . . .	100
5.3.2.3. Une grammaire faite de variantes. . . . .	102
5.4. Vers des systèmes adaptables . . . . .	102
5.4.1. Composants du système de reconnaissance des entités nommées. . . . .	103
5.4.1.1. Les dictionnaires. . . . .	103
5.4.1.2. La grammaire. . . . .	105
5.4.1.3. Le processus d'apprentissage. . . . .	106
5.4.1.4. Les mécanismes de révision . . . . .	107
5.4.2. « Déconstruire » un système de reconnaissance des entités nommées . . . . .	109
5.4.2.1. Les dictionnaires de noms propres . . . . .	109

8 Extraction automatique d'information

5.4.2.2. La grammaire. . . . .	111
5.4.2.3. Inférence et généralisation . . . . .	112
5.4.2.4. Capacités de révision . . . . .	113
5.4.3. Analyse des erreurs restantes . . . . .	113
5.4.4. Prédire l'apport de l'apprentissage . . . . .	114
<b>Chapitre 6. La mise en relation des entités . . . . .</b>	<b>117</b>
6.1. Cadre de l'expérience . . . . .	117
6.1.1. Annotation des entités . . . . .	117
6.1.2. Définition du formulaire d'extraction . . . . .	119
6.2. Elaboration manuelle de classes sémantiques et de grammaires d'extraction . . . . .	121
6.3. Bilan et performances du système élaboré manuellement . . . . .	126
<b>Chapitre 7. Acquisition semi-automatique de classes sémantiques . . . . .</b>	<b>127</b>
7.1. Acquisition de classes sémantiques : ressources générales <i>versus</i> ressources spécifiques . . . . .	128
7.2. Acquisition automatique de familles sémantiques par apprentissage symbolique interactif . . . . .	131
7.2.1. Le système d'apprentissage ASIUM. . . . .	132
7.2.2. Apprentissage supervisé ou non supervisé ? . . . . .	135
7.2.3. Critères pour l'élaboration des classes . . . . .	136
7.2.4. Mesurer l'apport de l'apprentissage pour l'acquisition de ressources . . . . .	137
7.3. Utilisation d'une ressource linguistique générale : le réseau sémantique de Memodata . . . . .	139
7.3.1. Le réseau sémantique : le DICTIONNAIRE INTÉGRAL et les outils associés. . . . .	139
7.3.2. Critères pour l'élaboration des classes . . . . .	142
7.3.3. Mesurer l'apport de ressources générales pour l'acquisition de classes sémantiques . . . . .	143
7.4. Bilan . . . . .	145
7.4.1. Evaluation des deux approches proposées. . . . .	145
7.4.2. Combinaison de méthodes pour l'acquisition de ressources . . . . .	147
7.4.3. Les outils d'acquisition de ressources : une évaluation difficile. . . . .	148
<b>Chapitre 8. Acquisition semi-automatique de patrons d'extraction . . . . .</b>	<b>151</b>
8.1. Description de la tâche et de l'approche adoptée . . . . .	151
8.2. Eléments pour le repérage de prédicats en situation de paraphrase . . . . .	153
8.2.1. Calcul de la distance entre mots. . . . .	153
8.2.2. Pondération de la mesure de proximité sémantique . . . . .	158
8.2.3. Remarques sur les mesures proposées par le SÉMIOGRAPHE. . . . .	159

8.3. Stratégie de recherche de prédicats en situation de paraphrase . . . . .	160
8.3.1. Normalisation du corpus . . . . .	161
8.3.2. Filtrage de séquences potentiellement pertinentes . . . . .	162
8.3.2.1. Principes de fonctionnement. . . . .	162
8.3.2.2. Evaluation de l'outil de filtrage . . . . .	164
8.3.3. Sélection manuelle de structures prédictives caractéristiques . . .	166
8.3.4. Expansion sémantique de patrons. . . . .	166
8.3.4.1. Repérage de structures prédictives en situation de paraphrase .	166
8.3.4.2. Génération de grammaires d'extraction sous la forme de transducteurs à nombre fini d'états . . . . .	170
8.3.4.3. Evaluation de l'outil d'acquisition de structures prédictives en situation de paraphrase . . . . .	174
8.3.5. Mesure de l'utilisabilité . . . . .	178
8.4. Positionnement par rapport à d'autres travaux . . . . .	179

<b>Conclusion</b> . . . . .	181
-----------------------------	-----

<b>Annexe</b> . . . . .	187
-------------------------	-----

A.1. Les <i>Message Understanding Conferences</i> , tableaux récapitulatifs . . .	187
A.2. Expressions régulières, automates et transducteurs dans INTEX . . . .	189
A.3. Ressources pour la reconnaissance des entités nommées . . . . .	197
A.4. Acquisition de classes sémantiques : le cas de la classe « opération d'achat » . . . . .	205

<b>Bibliographie</b> . . . . .	211
--------------------------------	-----

<b>Glossaire</b> . . . . .	227
----------------------------	-----

<b>Index</b> . . . . .	235
------------------------	-----