Collège Doctoral du Languedoc-Roussillon

THÈSE

Pour obtenir le grade de

Docteur

Délivré par l'**Université Montpellier**

Préparée au sein de l'école doctorale **I2S***
Et de l'unité de recherche **LIRMM**

Spécialité: **Informatique**

Présentée par **Lionel Ramadier**

Indexation et apprentissage de termes et de relations à partir de comptes rendus de radiologie

Soutenue le <JJ/MM/AAAA> devant le jury composé de :

Nuria Gala	MdC, HDR	Univ. Marseille	Rapporteur
Hervé Blanchon	MdC, HDR	Univ. Grenoble	Rapporteur
Anne Laurent	Professeur	Univ. Montpellier	Examinateur
Marianne Huchard	Professeur	Univ. Montpellier	Examinateur
Roland Ducournau	Professeur	Univ. Montpellier	Examinateur
Patrice Bellot	Professeur	Univ. Marseille	Examinateur
Juan-Manuel Torres Moreno	MdC, HDR	Univ. Avignon	Examinateur
Mathieu Lafourcade	MdC, HDR	Univ. Montpellier	Directeur
Denis Hoa	PDG, IMAIOS	Montpellier	Invité

[Analyse et Traitement Informatique de la langue Française]

* **I2S :** École doctorale Information Structures Systèmes

Remerciements

Je voudrais remercier chaleureusement :

Mathieu Lafourcade , mon Directeur de thèse, pour sa disponibilité et son aide efficace. Je tiens à le remercier également pour sa grande disponibilité, son encouragement permanent ainsi que ces conseils précieux.

Denis Hoa et Antoine Micheau présidents fondateurs de l'entreprise Imaios - pour leur accueil , leur patience et leur grande liberté qu'ils m'ont accordé pour mon travail de recherche.

Mme Núria Gala et Hervé Blanchon qui m'ont fait l'honneur d'être les rapporteurs de cette thèse.

Je remercie également tous les membres de l'**équipe TEXTE** ainsi que les employés de la société **IMAIOS** qui m'ont soutenu et supporté pendant 3 ans. Je v remercie Jérôme pour son aide pour le code informatique ainsi que Vincent, Mickael, Julien et aussi Virginie pour son aide dans la préparation de mes présentations. Je voudrais aussi remercier chaleureusement mes camarades doctorants (Manel, Nadia, Guillaume, Kevin).

Et enfin , je tiens à remercier tous mes proches (mes parents, ma soeur) qui m'ont soutenu dans cette aventure.

Résumé

Dans le domaine médical, l'informatisation des professions de santé et le développement du dossier médical personnel (DMP) entraîne une progression rapide du volume d'information médicale numérique. Le besoin de convertir et de manipuler toute ces informations sous une forme structurée constitue un enjeu majeur. C'est le point de départ de la mise au point d'outils d'interrogation appropriés pour lesquels, les méthodes issues du traitement automatique du langage naturel (TALN) semblent bien adaptées. Les travaux de cette thèse s'inscrivent dans le domaine de l'analyse de documents médicaux et traitent de la problématique de la représentation de l'information biomédicale et de son accès. Nous proposons de construire une base de connaissances dédiée à la radiologie à l'intérieur d'une base de connaissance générale (réseau lexico-sémantique JeuxDeMots). Nous montrons l'intérêt de l'hypothèse de non séparation entre les différents types de connaissances dans le cadre d'une analyse de documents. Cette hypothèse est que l'utilisation de connaissances générales, en plus de celles de spécialité, permet d'améliorer significativement l'analyse des documents médicaux. Au niveau du réseau lexico-sémantique, l'ajout manuel et automatisé des méta-informations sur les annotations est particulièrement utile. Ce réseau combine poids et annotations sur des relations typées entre des termes et des concepts ainsi qu'un mécanisme d'inférence dont l'objet est d'améliorer la qualité et la couverture du réseau. Nous décrivons comment à partir d'informations sémantiques présentes dans le réseau, il est possible de définir une augmentation des index bruts construits pour chaque comptes rendus afin d'améliorer la recherche documentaire. Nous présentons, ensuite, une méthode d'extraction de relations sémantiques entre des termes ou concepts. Cette extraction est réalisée à l'aide de patrons linguistiques auxquels nous avons rajouté des contraintes sémantiques. Les résultats des évaluations montrent que l'hypothèse de non séparation entre les différents types de connaissances améliore la pertinence de l'indexation.

Abstract

In the medical field, the computerization of health professions and the development of the personal medical file (DMP) results in a fast increase in the volume of medical digital information. The need to convert and manipulate all this information in a structured form is a major challenge. This is the starting point for the development of appropriate tools where the methods from the natural language processing (NLP) seem well suited. The work of this thesis are within the field of analysis of medical documents and address the issue of representing biomedical information (especially the radiology area) and its access. We propose to build a knowledge base dedicated to radiology within a general knowledge base (lexical-semantic network JeuxDeMots). We show the interest of the hypothesis of no separation between different types of knowledge through a document analysis. This hypothesis is that the use of general knowledge, in addition to specialized, significantly improves the analysis of medical documents. At the level of lexical-semantic network, manual and automated addition of meta information on annotations (frequency information, pertinence, etc.) is particularly useful. This network combines weight and annotations on typed relationships between terms and concepts as well as an inference mechanism which aims to improve quality and network coverage. We describe how from semantic information in the network, it is possible to define an increase in gross index built for each records to improve information retrieval. We present then a method on extracting semantic relationships between terms or concepts. This extraction is performed using lexical patterns to which we added semantic constraints. The results show that the hypothesis of no separation between different types of knowledge improves the relevance of indexing. The index increase results in an improved return while semantic constraints improve the accuracy of the extraction of relations.

Table des matières

Introduction 3

1 Recherche d'informations 15

1.1 Quelques différents modèles de recherche d'information 16

1.1.1 Modèle booléen 17

1.1.2 Modèles vectoriels 20

1.1.3 Modèles connexionnistes 23

1.1.4 Modèles probabilistes 24

1.2 Évaluation des Systèmes de RI 28

1.2.1 Mesures de rappel/précision 29

1.2.2 Précision moyenne et gain cumulatif réduit 32

1.2.3 La courbe ROC 33

1.3 Les différentes étapes de la recherche d'information 34

1.3.1 Le processus d'indexation 34

1.3.2 Les requêtes 38

2 Indexation sémantique et bases de connaissance : un état de l'art 41

2.1 Définition de l'indexation sémantique 42

2.1.1 Indexation pour des textes généraux 43

2.1.2 Indexation dans le domaine médical 47

2.1.3 Indexation par propagation 49

2.2 Critère d'évaluation de l'indexation 50

2.2.1 Consistance de l'indexation 51

2.2.2 Exactitude 51

2.2.3 Qualité de l'indexation 52

2.3 Utilisation de bases de connaissances 52

2.3.1 Définition des ontologies 53

2.3.2 Définition d'un réseau sémantique 56

2.4 Extraction de relations et le TALN dans le domaine médical 71

2.4.1 Extraction de relations 72

2.4.2 Traitement automatique du langage naturel dans le domaine médical 74

3 Le réseau lexico-sémantique JDM et le domaine radiologique 79

3.1 Crowsourcing et Game With A Purpose 81

3.1.1 Crowdsourcing 81

3.1.2 Un outil contributif : Diko 83

3.2 Annotation de relations 85
3.2.1 Déduction 85
3.2.2 Principe des annotations de relation 87
3.2.3 Expérimentations sur la propagation des annotations 95
3.2.4 Exploitation des annotations 98

4 Constitution de la base connaissances spécialisées dans la base JeuxDeMots et indexation des comptes rendus radiologiques 103

4.1 Constitution de connaissances spécialisées 104
4.1.1 Corpus de comptes rendus de radiologie 108
4.1.2 Pré-traitement du corpus 109
4.2 Augmentation d'index par propagation à travers le réseau JDM . . . 113
4.2.1 Indexation standard des comptes rendus 114
4.2.2 Algorithme d'augmentation par propagation 119
4.2.3 Évaluation des index augmentés 122

5 Extraction de relations sémantiques 127

5.1 Patrons lexicaux 129
5.1.1 Définition de patron lexical 129
5.1.2 Exemples de patrons lexicaux 130
5.2 Contraintes sur les patrons 134
5.2.1 Patrons sémantiques 134

5.2.2 Algorithme d'identification des relations 136
5.3 Expérimentation et résultats 137
5.3.1 Expérience 137
5.3.2 Résultats 137
5.4 Modèle PMA (Patient, Modalité, Affection) 141
5.4.1 Patient 142
5.4.2 Modalité 143
5.4.3 Affection 144

Conclusion générale **147**

Annexes **173**

Table des figures

1.1 Exemples de recherche booléenne. 18

1.2 Modèle vectoriel 21

1.3 Modèle connexionniste 23

1.4 Exemple d'une courbe précision-rappel 31

1.5 Exemple d'un index inversé. 35

2.1 cycle de vie de l'indexation MTI 48

2.2 Relations dans WordNet(version 2006) d'après université de Princeton [1] 58

2.3 Vue globale de Babel Net pour le terme *balloon* 60

2.4 Exemple de sous réseau lexical pour le terme IRM dans JeuxDeMots. 61

2.5 Représentation de UMLS 64

2.6 Le réseau sémantique UMLS 65

2.7 Représentation d'une partie de SNOMED 66

2.8 Exemple de la représentation dans FMA 69

2.9 Représentation de Radlex 70

1. https ://www.cs.princeton.edu/courses/archive/fall14/cos226/assignments/wordnet.html

2.10 Représentation des connaissances dans Gamuts 71

3.1 Capture écran de la fenêtre de Diko du terme *cirrhose*. 84

3.2 Capture écran de la fenêtre de Diko du terme *signe du Mont Fuji*. . . 84

3.3 Schéma d'inférence déductive triangulaire 86

3.4 Exemple d'implémentation d'annotation. 90

3.5 Approche basée sur la hiérarchie utilisée pour choisir l'annotation la plus précise avec plusieurs termes centraux 95

4.1 schéma montrant les relations pour un signe d'imagerie médicale (signe du Mont Fuji) 107

4.2 Exemple de compte rendu original 108

4.3 Schéma général de l'ajout de connaissances spécialisées dans le réseau JDM. 113

4.4 Exemple d'un compte rendu et de son index brut 118

4.5 Exemple de comptes rendus et de leur index brut. 118

4.6 Index augmenté correspondant à la figure 4.4 120

4.7 Index augmenté correspondant à la figure 4.5. 120

4.8 Index augmenté correspondant à la figure 4.4 125

5.1 Modélisation de certaines relations sémantiques utilisées. 131

5.2 Exemple de relations sémantiques utilisées. 131

5.3 Devenir des relations extraites. 139

5.4 Exemple de mots clés pour découvrir le genre (homme ou femme) lorsque celui-ci n'est pas explicite. 142

5.5 Exemple de comptes rendus ou le genre du patient est déduit par les mots en gras. 142

5.6 Compte rendu original. Les termes en gras permettent de déduire la modalité (scanner dans cet exemple) 145

5.7 Extraction des motifs *PMA* 145

8 Capture écran du prototype du moteur de recherche Okapi 180

Liste des tableaux

1.1 Modèle booléen : exemple 17

1.2 Modèle booléen flou 20

1.3 Méthodes de lissage 27

1.4 Mesure du rappel et de la précision 31

2.1 Exemple de mots présents dans un anti-dictionnaire général. 43

2.2 Nombre de mots et de concepts dans WordNet 58

3.1 Nombres de liens entre termes dans JeuxDeMots pour certains termes clés du domaine médical. 85

3.2 Relations pertinentes en radiologie pour l'analyse de compte-rendu . . 89

3.3 Nombre d'annotations inférées après application du système d'annotation des relations sur celles existantes. 97

3.4 Exemples de termes remplacés 100

3.5 Score du test (*cloze test* pour les textes originaux et simplifiés 101

4.1 approche itérée pour la détection des multi-termes. 110

4.2 Exemple de page **Wikipédia** et des hyperliens extraits. 112

4.3 Résultats de l'algorithme de propagation. 122

4.4 Présentation des nouvelles valeurs *pert* **sans les raffinements** en fonction des paramètres NBI et S. 123

4.5 Apport des relations sémantiques prises séparément 124

4.6 Apport des relations sémantiques combinées 125

5.1 Liste des relations à détecter. 130

5.2 Exemples de patrons lexicaux 132

5.3 Résultats de l'extraction de relations sémantiques avec patrons linguistiques **sans** contraintes sémantiques 137

5.4 Résultats de l'extraction de relations sémantiques avec patrons linguistiques **avec** contraintes sémantiques 138

5.5 Comparaisons des résulats pour la relation *traitement* 139

5.6 Résultat de l'extraction sur différents corpus 140

5.7 Résultats de l'extraction de la variable *Patient* 143

5.8 Résultats de notre méthode pour la variable *modalité* 143

5.9 Résultats de notre méthode pour la variable affection 144

Liste des algorithmes

1 triDesGeneriques 93
2 rechercheDeMotComposés 116
3 DesambiguisationDesTermes 117
4 PropagationIndex 121
5 extractionDeRelations 136

Introduction générale

Informatisation des données médicales et contexte de cette recherche

Dans la société de médias dans laquelle nous vivons, l'information, au sens large, joue un rôle de plus en plus important. L'invention et le développement du support électronique ont permis le stockage de données de plus en plus considérables. Grâce à l'avènement de l'informatique, l'ordinateur permet de traiter des quantités importantes d'informations de toutes natures ainsi stockées. La quantité de données traitées par les différentes organisations est devenue si grande que leur manipulation serait impossible sans l'outil informatique. Pour le grand public, la plus grande source d'information disponible reste le web. En 2006, on estimait qu'il existait environ 60 milliards de pages en sachant que les principaux moteurs de recherches en indexaient 20 milliards. D'après l'union internationale des télécommunications, le nombre d'internautes en 2014 est d'environ 2,9 milliards, soit environ 40% de la population mondiale. Ces quelques chiffres permettent de comprendre et d'appréhender les défis majeurs que peuvent représenter la collecte, le stockage, la transmission de l'information, ainsi que la capacité à rechercher efficacement au sein de la masse de données qu'elle représente.

L'informatisation des données textuelles concerne non seulement le domaine général mais aussi les domaines de spécialité (domaine biomédical, domaine juridique,

domaine nucléaire, ...). Dans le domaine médical, l'informatisation des professions de santé et le développement du dossier médical personnel (DMP) entraîne une progression rapide du volume d'information médicale numérique. Les systèmes informatiques médicaux permettent de stocker de l'information (dossier médical, résultats d'examens complémentaires, images et comptes rendus radiologiques, par exemple), d'y accéder en vue de découvrir de nouvelles informations ou de fournir une aide à la décision pour l'amélioration de la qualité des soins. Ces informations constituent des banques de données, d'une grande importance, sur les plans économique, politique et sociétaux. Ils peuvent avoir un impact déterminant sur les décisions de santé publique.

L'information médicale à exploiter est pour une grande part sous forme textuelle, et il s'agit alors de pouvoir extraire de façon automatique des données sémantiques. Dans la plupart des cas, les textes médicaux sont écrits de façon libre, et non structurés. Le besoin de convertir toute cette information sous une forme structurée et automatiquement interprétable constitue un enjeu majeur. C'est le point de départ du développement et de la mise au point d'outils d'interrogations appropriés. Pour cela, les méthodes issues du traitement automatique du langage naturel (TALN) semblent bien adaptées.

Dans le domaine de l'imagerie médicale, cette recherche sémantique pourra être combinée avec la recherche par similitude (CBIR (Content Based Image Retrieval)) pour améliorer l'extraction d'images médicales. Cela permettra d'améliorer le suivi des patients, la communication entre praticiens, l'aide au diagnostic ainsi que l'aide pédagogique.

Cette thèse s'est déroulée dans le cadre d'une Convention Industrielle de Formation par la Recherche (CIFRE) en partenariat avec la société IMAIOS [2] . Son activité est dédiée à la recherche et au développement de nouvelles solutions de formation et d'aide au diagnostic dans le domaine médical. Elle s'appuie sur l'expertise des deux fondateurs de l'entreprise, médecins radiologues. La société IMAIOS développe des sites internet médicaux destinés aux professionnels de santé. Le caractère innovant de ses réalisations se fonde sur une technologie de visualisation d'examens radiologiques à partir d'une interface web. L'entreprise a étudié la faisabilité d'un système de recherche et d'identification de sous-ensembles d'une image pour l'identification

2. https ://www.imaios.com/fr

d'une maladie à partir d'une grande base de données d'images. Dans ce cadre, des recherches ont été effectuées sur les descripteurs visuels les plus adaptés pour caractériser les lésions présentes au niveau du foie (couleur, forme, texture,...). Toujours dans un objectif d'amélioration de l'aide au diagnostic, une partie sémantique liée à l'analyse des comptes rendus a été ajoutée en vu de réaliser un moteur de recherche (*projet IMAIOS*).

Recherche d'information biomédicale

La recherche d'information est la branche qui étudie la façon de retrouver une information dans un ensemble de documents. Le processus général de recherche d'information a été décrit par R.K.Belew [Belew, 2000]. Ainsi, l'objectif de la recherche d'information peut être défini comme étant de permettre l'exploitation des informations contenues dans des documents textuels.

Un paradoxe, noté dans la littérature biomédicale, réside dans le fait que l'augmentation de la disponibilité des comptes rendus médicaux n'est pas toujours synonyme de disponibilité de l'information [Christensen et Grimsmo, 2008]. Depuis quelques années la recherche d'information dans le domaine biomédical fait l'objet de plusieurs campagnes d'évaluation [Simpson *et al.*, 2014]. Par exemple, nous pouvons citer la campagne d'évaluation TREC (Text REtrieval Conference) [3].

Dans le domaine de l'imagerie médicale, l'objectif de la recherche d'information (pathologie, termes anatomiques, indication,...) au sein des comptes rendus de radiologie est de fournir aux professionnels de santé (radiologues, chirurgiens, orthopédistes...) la bonne information et surtout la plus précise possible. Les difficultés de l'extraction d'information contenue dans les comptes rendus médicaux ne doivent pas être sous-estimées [Edinger *et al.*, 2012]. Les différents éléments contribuant à la complexité de la tâche d'extraction d'information comprennent l'utilisation par les radiologues de beaucoup d'abréviations et d'acronymes [Barrows Jr *et al.*, 2000] (FO pour *foramen ovale* ou *fond d'oeil*), de raccourcis (*hépatite C* au lieu de *virus de l'hépatite C*), d'élisions (la *communicante antérieure* au lieu de l'*artère communicante antérieure*), de phrases agrammaticales (phrases sans verbes), d'incorrections

3. (http ://trec.nist.gov)

(fautes d'orthographe). La détermination du sens (ou des usages) des termes ou des abréviations [McInnes et Stevenson, 2014], la résolution des anaphores et des coréférences constituent aussi un enjeux supplémentaire. Les méthodes d'extraction automatique d'information (incluant les méthodes du traitement automatique du langage naturel) portent l'espoir de pouvoir transformer les comptes rendus écrits de façon libre en un format structuré plus facilement utilisable par la machine.

Indexation et bases de connaissances

L'indexation est le processus qui consiste à décrire et à caractériser un document à l'aide de la représentation du contenu. Le but de l'indexation est de préciser la teneur du document afin de permettre une recherche efficace des informations contenues dans un corpus de documents. Le contenu d'un document peut être décrit de façon plus ou moins détaillée. Une indexation idéale doit être à la fois pertinente, objective et cohérente. La pertinence signifie que les mots-clés de l'index doivent rendre compte du contenu du document et ces mots choisis doivent être désambiguïsés le plus précisément possible. L'objectivité consiste à respecter le plus possible la pensée de l'auteur du texte. La cohérence est liée au fait que les mêmes mots doivent désigner les mêmes notions.

Les mots-clés, parfois appelés *descripteurs* sont soit des mots simples, des mots composés, des lemmes ou tout autre unité d'information pouvant décrire le contenu d'un document. Cette indexation du document par ses mots-clés peut être réalisée de façon manuelle, semi-automatique ou automatique. Lorsqu'il s'agit d'indexer des documents du domaine médical, les indexeurs utilisent un vocabulaire contrôlé sous forme de thésaurus ou d'ontologie spécialisée dans le domaine de spécialité que n'utilise pas le praticien qui rédige le compte rendu. Dans les bases terminologiques, les termes sont liés par de simples relations hiérarchiques telles que la relation d'hyperonymie (*is-a*) ou par un ensemble de relations lexicales et sémantiques et on parle dans ce cas de réseau sémantique. Dans le domaine médical, les vocabulaires contrôlés, en particulier leur agrégation au sein du Metathesaurus de l'UMLS [Cimino *et al.*, 1994] sont actuellement très utilisés pour réaliser l'indexation de documents médicaux. D'autres ontologies existent et tentent de couvrir l'ensemble de la médecine. Dans le champ de la radiologie, nous pouvons citer RadLex

[Langlotz, 2006] qui n'est pas incorporé au sein de l'UMLS.

Questions et contributions de la thèse

Pour un utilisateur d'un domaine spécialisé, il existe un besoin de tenir compte non seulement des connaissances du domaine de spécialité mais aussi des connaissances du domaine général. En radiologie, le praticien peut être amené à écrire une requête contenant des termes d'ordre général (*accident de ski, encornage par un taureau,etc.*). Comme nous l'avons déjà mentionné ci-dessus, à l'heure actuelle, les systèmes existants dans le domaine médical utilisent des ontologies spécialisées et ne prennent pas en compte des informations d'ordre général ce qui peut constituer un premiere faiblesse dans la recherche d'information. Un deuxième aspect est celui de l'augmentation d'index qui est une forme d'inférence, pour permettre au système de retourner des documents où les termes de la requête ne sont pas présents dans le document, mais lui sont néanmoins liés par l'intermédiaire de diverses relations (synonymie, hyperonymie). Dans l'état de l'art, l'augmentation est souvent réalisée à un voisinage de 1 par rapport à un terme, c'est à dire, que pour une relation, par exemple celle d'hyperonymie, c'est un seul générique qui est choisi, souvent le plus spécifique [Fiorini *et al.*, 2014]. Nous proposons une augmentation d'index à un voisinage 3, 4 voire 5 et nous mettons dans l'index plusieurs génériques. Le travail que nous présentons traite de ces problématiques et essaie de proposer une solution en se basant sur un réseau lexico-sémantique de connaissances générales mais contenant aussi des domaines de spécialité (dont le domaine de la radiologie). La question à laquelle nous essayons de répondre dans ce travail est : dans quelle mesure la non-séparation des connaissances générales et de spécialité peut apporter un avantage à la tâche d'indexation et de recherche d'information ?

L'axe qui a été suivi au cours de ces travaux de thèse concerne l'utilisation d'un réseau lexico-sémantique qui combine des connaissances générales et des connaissances de spécialités. Nous pouvons utiliser cette ressource pour réaliser une indexation automatique des comptes rendus ainsi qu'une augmentation d'index à l'aide d'algorithmes de propagation. La deuxième étape a été l'extraction de relations, dans deux buts principaux. Le premier est d'alimenter la base de données, alors que le second permet d'affiner l'indexation en vue d'améliorer la recherche d'informa-

tion dans le domaine de la radiologie. Un objectif annexe est également la détection d'incohérences dans les comptes rendus ou la base de connaissance.

Dans le cadre de la thèse, un moteur de recherche a été implémenté à partir des travaux mentionnés ci-dessus. Étant donné qu'il valide les résultats de l'approche de non-séparation, il permet aux radiologues de réaliser des requêtes ouvertes.

Organisation de ce mémoire

Cette thèse est constituée de 5 chapitres répartis en deux parties. La première partie (chapitre 1 et 2) fait un état de l'art en lien avec le cadre de notre thèse.

Le premier chapitre de cette partie présente les notions et les concepts de base de la recherche d'information ainsi que les principaux modèles utilisés. Nous abordons les différentes étapes nécessaires à cette tâche. Les principales mesures de pertinence utilisées en RI sont décrites.

Le chapitre 2 est consacré à l'état de l'art concernant l'indexation sémantique. Après l'avoir réalisé dans le domaine général, nous abordons cette thématique dans le domaine biomédical. Ensuite, nous détaillons l'utilisation de bases de connaissances dans le cadre de l'indexation et de la recherche d'information. De la même manière, nous partons d'un cadre global avec des bases de connaissances d'ordre général et allons vers des bases de connaissance de spécialité biomédicale. Nous présentons le domaine de l'indexation par propagation en expliquant son principe. Ce chapitre se conclue par un état de l'art sur l'extraction de relations avec un double objectif. Le premier objectif est de contribuer à l'amélioration d'une base de connaissances et le second est de réaliser une indexation des relations entre termes dans un but de recherche d'information.

La deuxième partie de ce manuscrit comprend trois chapitres (chapitre 3, chapitre 4 et chapitre 5) et concerne notre contribution au sujet de la non-séparation des connaissances générales et de spécialités pour la recherche d'information. Le chapitre 3 présente le réseau lexico-sémantique utilisé dans nos travaux, à savoir le réseau JeuxDeMots (JDM). Nous en expliquons le principe et présentons un outil contributif (Diko) intégré au projet JDM. Ce chapitre abordera également la ques-

tion de l'annotation des relations à l'intérieur du réseau. Le chapitre 4 est consacré à la constitution de la base de connaissances de spécialité à l'intérieur du réseau de connaissances générales. Nous abordons ensuite la notion d'index augmenté grâce à l'implémentation d'un algorithme de propagation. Les différentes étapes du modèle sont détaillées et une évaluation est décrite.

Le chapitre 5 propose un modèle d'extraction de relations sémantiques basé sur les patrons linguistiques et sémantiques. Dans ce modèle, nous avons adjoint des contraintes sémantiques à certains patrons linguistiques. Une évaluation de ces contraintes est décrite comparativement à l'approche basée uniquement sur les patrons linguistiques. A la fin du chapitre, nous décrivons l'approche d'extraction *Patients, Modalité, Affection (PMA)*. Nous expliquons la méthode d'extraction ainsi que l'utilité d'un tel modèle.

En conclusion, nous dressons le bilan des travaux réalisés, dans le cadre général de l'indexation et de l'analyse de documents médicaux. Nous introduisons ensuite les perspectives d'évolution envisageables pour ces travaux.

PARTIE 1 : Etat de l'art

Chapitre 1

Recherche d'informations

Sommaire

1.1 Quelques différents modèles de recherche d'information 16

1.1.1 Modèle booléen 17

1.1.2 Modèles vectoriels 20

1.1.3 Modèles connexionnistes 23

1.1.4 Modèles probabilistes 24

1.2 Évaluation des Systèmes de RI 28

1.2.1 Mesures de rappel/précision 29

1.2.2 Précision moyenne et gain cumulatif réduit 32

1.2.3 La courbe ROC 33

1.3 Les différentes étapes de la recherche d'information . . . 34

1.3.1 Le processus d'indexation 34

1.3.2 Les requêtes 38

La recherche d'information (RI) [Grossman et Frieder, 2012] [Salton, 1971] est une des branches de l'informatique qui étudie la façon de sélectionner à partir d'un corpus de documents, ceux qui sont susceptibles de répondre à la requête de l'utilisateur. Traiter des textes nécessite de pouvoir les stocker, les rechercher, les explorer et de les sélectionner de façon pertinente. À partir de cette définition, nous pouvons introduire différentes notions :

— Corpus : un ensemble de documents.
— Document : l'objet élémentaire d'un corpus.
— Besoin d'information : les besoins de l'utilisateur (qui va chercher des documents).
— Requête : l'interface entre l'utilisateur et un système de recherche d'information.

1.1 Quelques différents modèles de recherche d'information

De façon générale, on distingue trois modèles principaux de recherche d'information :

— les modèles booléens : des méthodes ensemblistes de représentation du contenu d'un document. Il existe le modèle booléen pur (*boolean model*), le modèle

booléen étendu et un modèle basé sur les ensembles flous.
— les modèles vectoriels : le contenu d'un document est représenté selon une approche algébrique.
— les modèles probabilistes : ces modèles essaient d'inférer la probabilité de pertinence du document, connaissant la requête.

1.1.1 Modèle booléen

Le modèle booléen, une des premières méthodes utilisées en recherche d'information, est fondé sur la logique booléenne et la théorie des ensembles [Manning *et al.*, 2008]. Les documents sont représentés par des ensembles de termes grâce à un index inversé et les requêtes sont traitées par des expressions combinant des termes et des opérateurs logiques ET, OU et SAUF, selon le formalisme de l'algèbre de Boole. La recherche booléenne, en utilisant la structure d'index inversé, consiste à parcourir les listes de documents associés à la requête et à fusionner ces listes par rapport aux opérateurs logiques présents dans cette dernière (table1.1). Un document du corpus est considéré comme pertinent quand son contenu vérifie exactement l'expression de la requête.

le document contient					pertinence du document
cyclisme	natation	cyclisme OR natation	dopage	NOT dopage	
0	0	0	0	1	0
0	0	0	1	0	0
0	1	1	0	1	1

TABLE 1.1 – Modèle booléen : exemple
Requête Q : **(cyclisme OR natation) AND NOT dopage**

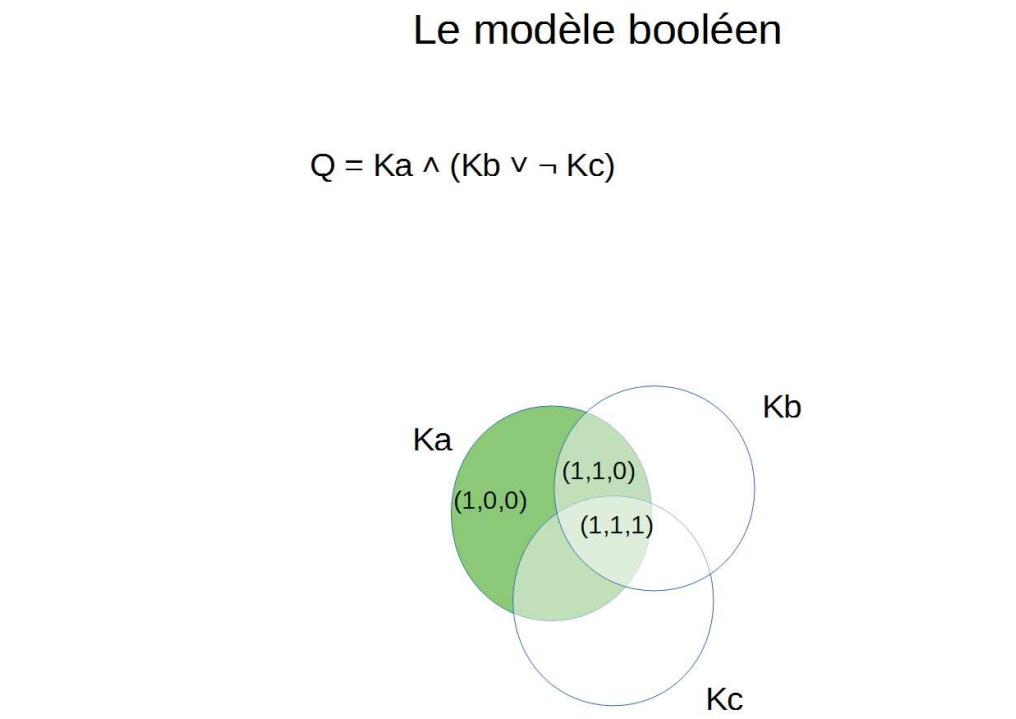


FIGURE 1.1 – Exemples de recherche booléenne.

Le modèle booléen standard présente l'avantage de la simplicité de sa mise en œuvre et de la clarté de l'expression de la requête (figure 1.1). Le modèle booléen peut être utile dans le cadre de corpus spécialisés où les utilisateurs possèdent une très bonne connaissance du vocabulaire. L'inconvénient de cette méthode est qu'elle effectue des appariements exacts entre les termes de la requête et les documents, ce qui ne permet pas de retrouver des documents pertinents ne contenant pas exactement les termes de la requête booléenne. Le deuxième inconvénient majeur de cette technique est qu'elle est incapable d'ordonner les documents par ordre de pertinence.

Pour contourner ces difficultés, Salton et al [Salton *et al.*, 1983] ont proposé le **modèle booléen étendu**. Il consiste à donner des poids aux termes des documents

et de la requête pour tenir compte de leur importance. Le poids des termes dans les documents est déterminé par des mesures statistiques comme par exemple la fréquence d'un terme dans le document (TF) et la fréquence inverse de documents (IDF). Le positionnement des documents se fait dans un espace euclidien dont les axes sont les termes de la requête. Dans le cas d'une requête composée de deux termes, une condition logique de type ET est représentée par la distance entre le document et les coordonnées (1,1) alors qu'une condition de type OU est calculée par la distance du document à l'origine (0,0). Cela permet d'ordonner les documents selon leur similarité avec la requête.

Pour modéliser les notions d'imprécision, d'incertitude de l'information, le **modèle booléen flou** a été proposé [Salton, 1989],[Dubois et Prade, 2012], [Paice, 1984], [Bordogna et Pasi, 2000], [Zadrozny et Kacprzyk, 2005]. Le modèle booléen flou, basé sur la théorie des ensembles flous ou la logique floue, est avant tout une extension du modèle booléen standard. L'objectif principal est l'introduction de la notion de degré d'appartenance d'un élément à un ensemble. Dans les ensembles flous, l'appartenance est mesurée par un degré variant entre 0 et 1. Ce modèle a pour objectif de caractériser un élément par son degré d'appartenance à un ensemble flou et de représenter un document donné par un ensemble flou de termes pondérés.

Ce modèle est utilisé pour traiter l'imprécision qui caractérise le processus d'indexation, de contrôler l'approximation de l'utilisateur dans sa requête ainsi que de traiter les réponses reflétant la pertinence partielle des documents par rapport aux requêtes. Des améliorations [Boughanem *et al.*, 2005] ont été proposées afin de palier l'inconvénient de ce modèle concernant le classement des documents pertinents, vu que les scores de pertinence sont déterminés par des fonctions min ou max. Chaque document du corpus d_i est vu comme un ensemble flou et chaque terme t_i comme objet de d_i. Une fonction d'appartenance $f_d(t_j)$ est créé. Un ensemble de termes pondérés représente un document d_i :

$$d_i = (t_1, w_{t1}..t_i, w_{ti})$$

avec $f_d(\mathrm{t}) = w_t$ qui est un entier compris entre 0 et 1. Cette fonction représente le degré d'appartenance du terme t_i dans l'ensemble flou d_i. Une requête q est une proposition flou dont nous calculons son degré d'appartenance $f_d(\mathrm{q})$ à chaque ensemble flou d_i. (table 1.2)

q	$f_{d_i}(q)$
t_i	$f_{d_i}(t_i)$
$t_i \wedge t_j$	$\min(f_{d_i}(t_i), f_{d_i}(t_j))$
$t_i \vee t_j$	$\max(f_{d_i}(t_i), f_{d_i}(t_j))$
$\neg t_i$	1 - $f_{d_i}(t_i)$

TABLE 1.2 – Modèle booléen flou
Exemple d'évaluation de requêtes.

1.1.2 Modèles vectoriels

Les modèles vectoriels englobent en pratique plusieurs modèles : notamment le modèle vectoriel proprement dit et l'indexation sémantique latente (LSI pour latent semantic indexation).

1.1.2.1 Le modèle vectoriel (Vector space Model)

L'hypothèse de base des modèles vectoriels est que les requêtes sont représentées dans le même espace vectoriel que les documents [Salton et McGill, 1986] et que comparer des vecteurs est relativement facile. L'espace est de dimension N (N étant le nombre de termes d'indexation de la collection de documents).L'indexation d'un document consiste à construire un vecteur le représentant tout en étant facilement manipulable. Le mécanisme de recherche de ce modèle consiste alors à retrouver les vecteurs documents qui sont les plus proches du vecteur requête. Ce modèle permet donc d'ordonner les documents selon une mesure de similarité avec la requête.

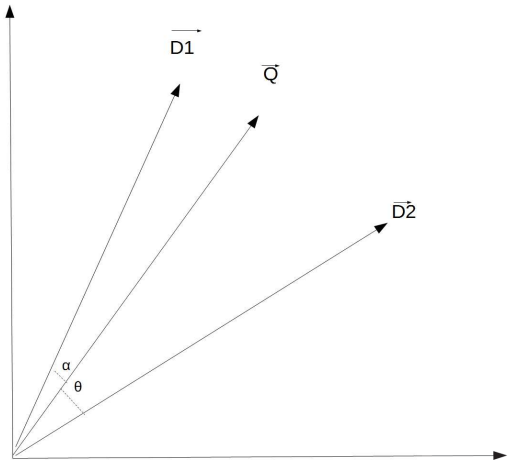


FIGURE 1.2 – Modèle vectoriel
Deux documents (D1 et D2) et une requête (Q) sont représentés dans un espace vectoriel (ici simplifié à l'extrême et n'ayant que deux dimensions). La distance de la requête avec les documents est représentée par les angles α et θ.

Une des mesures d'affinité la plus utilisée est la similarité cosinus [Singhal, 2001] qui permet de calculer la similarité entre deux vecteurs à n dimensions en déterminant le cosinus de l'angle entre eux (figure 1.2). Si on prend un document D du corpus et une requête Q, qui sont représentés respectivement par leur vecteur $\mathbf{D}$ et $\mathbf{Q}$.

$D_j = (d_{1j}, d_{2j}...d_{Nj})$

$\mathbf{Q} = (q_1, q_2...q_N)$

La mesure cosinus pour ces deux vecteurs est :

$$s(Q, D) = \frac{\sum_{i=1}^{N} q_i \times d_{ij}}{\sum_{i=1}^{N}\sqrt{q_i^2} \times \sum_{i=1}^{N}\sqrt{d_{ij}^2}} \tag{1.1}$$

avec d_{ij} qui représente le poids du terme t_i dans le document D_j et q_i est le poids du terme t_i de la requête Q. La pondération des termes composant la requête peut être soit la même que celle utilisée pour les documents du corpus, soit donnée par l'utilisateur lors de sa formulation.

Dans ce cas, plus l'angle entre les vecteurs est petit, plus le document est supposé être pertinent par rapport à la requête. Cette mesure n'est pas autre chose que la projection d'un vecteur sur l'autre (elle est symétrique).

Parmi les autres mesures de similarité, les plus utilisées sont les suivantes.

Mesure de Jaccard :

$$s(q, d_j) = \frac{\sum_{i=1}^{N} q_i \times d_{ij}}{\sum_{i=1}^{N} q^2 + \sum_{i=1}^{N} d_{ij}^2 - \sum_{i=1}^{N} q_i * d_{ij}} \tag{1.2}$$

Produit scalaire :

$$s(q, d_j) = \sum_{i=1}^{N} q_i * d_{ij} \tag{1.3}$$

1.1.2.2 L'indexation sémantique latente (LSI)

Ce modèle exploite les co-occurrences entre termes et réduit l'espace des termes, en regroupant les termes co-occurrents (similaires) dans les mêmes dimensions. Les documents et les requêtes sont alors représentés dans un espace plus petit [Berry *et al.*, 1995] [Deerwester *et al.*, 1990], [Papadimitriou *et al.*, 1998] composés de concepts de haut niveau. Cette méthode permet de sélectionner des documents pertinents même si le terme de la requête n'est pas présent dans les documents. LSI utilise une matrice qui décrit l'occurrence de certains termes dans les documents. LSI dépend de la décomposition de la matrice de termes et des documents par la technique de décomposition par valeurs singulières (SVD - *Singular Value Decomposition*). L'idée de SVD est que chaque matrice rectangulaire $A_{(m,n)}$ peut être écrite de la façon suivante :

$$A_{m,n} = U_{m,m}\ S_{m,n}\ V_{n,n}^T$$

où S représente la matrice diagonale qui contient les valeurs singulières de la matrice A, U représente les vecteurs propres orthonormés de la matrice AA^T , et V

représente ceux de la matrice A^TA. Les colonnes de ces trois matrices sont ordonnées d'une manière décroissante par rapport à leurs valeurs propres.

L'avantage de ce modèle est qu'il peut permettre de retrouver des documents même s'ils ne contiennent pas les mots de la requête. Son défaut majeur est qu'il est sensible à la quantité et à la qualité des données traitées.

1.1.3 Modèles connexionnistes

L'utilisation des réseaux de neurones dans le domaine de la recherche d'information est apparue dans les années 80 avec différents travaux [Mothe, 1994], [Mozer, 1984]. Ces modèles sont formés de neurones formels, le plus souvent organisés en couches et de liens pondérés reliant les neurones entre les différentes couches figure 1.3. Les neurones formels sont construits à partir des représentations initiales des documents. Le fonctionnement de ce réseau se fait par propagation de signaux de la couche d'entrée vers une couche de sortie. Les valeurs dans la couche de sortie servent de critères de décision (pertinence de documents). Ils sont formés d'une couche de termes qui sera activée par la requête de l'utilisateur et une couche de document. La requête active la couche des termes et cette activation se propage vers la couche de documents. La réponse du moteur de recherche est donnée par l'activation finale des documents.

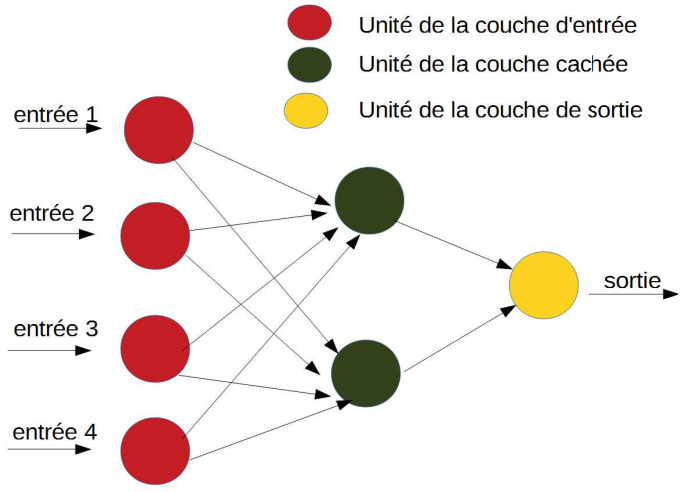


FIGURE 1.3 – Modèle connexionniste
Un réseau de neurones à couches.

Les systèmes de recherche d'information basés sur les réseaux de neurones permettent de représenter différentes associations comme par exemple les relations entre termes (synonymie, ...) ou entre documents (similitude, référence,...). Plusieurs modèles basés sur les réseaux de neurones ont été proposés dans le domaine de la recherche d'information [Crestani, 1994], [Kwok, 1989], [Crestani, 1997]. Récemment ont émergé des techniques d'apprentissage profond (*deep learning*) dans le domaine de la RI [Shen *et al.*, 2014], [Li et Lu, 2016].

1.1.4 Modèles probabilistes

Ces modèles essaient d'inférer la probabilité de pertinence du document sachant la probabilité qu'une requête soit associée à un document. Les modèles de recherche probabilistes se basent sur la théorie des probabilités et ont été proposés au début des années 60 [Maron et Kuhns, 1960]. Le principe de ces modèles considère chaque paire formée d'un document et d'une requête (d,q) comme une expérience aléatoire qui consiste à tirer de façon simultanée ces observations d'un ensemble prédéfini de documents et de requêtes. On associe, à chaque tirage (d,q) une variable aléatoire $R_{d,q}$ qui est égale à 1 si le document d est pertinent étant donné la requête q et 0 dans le cas contraire. Dans la suite de cette section, nous présentons les modèles probabilistes les plus courants, à savoir le modèle d'indépendance binaire, le modèle OKAPI BM25, les modèles de langue et les modèles informationnels.

Modèle d'indépendance binaire (MIB)

Ce modèle associé au principe d'ordonnancement probabiliste [Robertson, 1977] repose sur plusieurs hypothèses. Tout d'abord, les documents et les requêtes sont représentés sous forme de vecteurs binaires de même taille que le vocabulaire du corpus de document. La seconde hypothèse stipule que les termes présents dans les documents sont mutuellement indépendants. C'est l'hypothèse *Bayésienne Naïve* et elle est à l'origine d'une famille de modèles de classification portant le même nom. L'utilité d'un document pertinent ne dépend pas du nombre de documents pertinents que l'utilisateur a déjà obtenu. On modélise la pertinence comme un événement probabiliste.

Plus spécifiquement, un document est représenté par un vecteur d = (x_1, ...,x_m)

où x_t=1 si le terme t est présent dans le document et x_t=0 s'il n'est pas présent. Les requêtes sont représentées d'une façon similaire. En appliquant le théorème de Bayes, la probabilité P(R|d,q) qu'un document soit pertinent (ou non pertinent) par rapport à une requête, nous donne l'équation :

$$P(R|\overrightarrow{x}, \overrightarrow{q}) = \frac{P(\overrightarrow{x}|R, \overrightarrow{q}) * P(R|\overrightarrow{q})}{P(x|\overrightarrow{q})} \tag{1.4}$$

où P($\overrightarrow{x}$|R=1,$\overrightarrow{q}$) et P($\overrightarrow{x}$|R=0,$\overrightarrow{q}$) sont les probabilités d'extraire des documents pertinents ou non pertinent, respectivement. Les probabilités exactes ne peuvent pas être connues à l'avance, par conséquence, des estimations déterminées à partir du corpus doivent être utilisées. P(R=1|$\overrightarrow{q}$) et P(R=0|$\overrightarrow{q}$) indiquent la probabilité préalable d'un document pertinent (ou non pertinent, respectivement) pour une requête q. Si, par exemple, nous connaissons le pourcentage de documents pertinents dans la collection, alors nous pourrons l'utiliser pour estimer ces probabilités. Puisqu'un document est pertinent ou non pertinent à une requête, nous devons avoir ceci :

$$P(R = 1|\overrightarrow{x}, \overrightarrow{q}) + P(R = 0|\overrightarrow{x}, \overrightarrow{q}) = 1 \tag{1.5}$$

Modèle Okapi BM25

L'idée de départ de ce modèle [Robertson et Walker, 1994],[Zaragoza *et al.*, 2004] est qu'un bon descripteur du document est un terme fréquent dans ce document mais relativement rare dans l'ensemble du corpus. Ils sont partis du constat que certains termes apparaissent avec une fréquence basse dans beaucoup de documents alors qu'ils apparaissent avec une fréquence élevée dans un groupe particulier de documents (appelé groupe élite). Cela a amené à modéliser ces groupes de documents avec une loi de Poisson de paramètre *lambda pertinent* et *lambda non pertinent*. La probabilité d'apparition d'un terme apparaissant un certain nombre de fois dans un document du corpus peut être modélisée par une distribution mixte 2-poisson. Ainsi dans Okapi une hypothèse d'indépendance est faite sur l'élitisme et la fréquence d'un terme dans un document ne dépend que de l'appartenance du document à l'ensemble élite.

Nous obtenons alors les équations suivantes :

$$\begin{aligned} P(tf(t_i)|\overline{R})) &= p(tf(t_i)|E)p(E|\overline{R}) + p(tf(t_i)|\overline{E})p(\overline{E}|\overline{R}) \\ P(tf(t_i)|\overline{R})) &= p(tf(t_i)|E)p(E|R) + p(tf(t_i)|\overline{E})p(\overline{E}|R) \end{aligned} \tag{1.6}$$

où E représente l'ensemble des documents élites du terme t_i, P(R|d) est la probabilité qu'un document d soit pertinent (noté R pour relevant) et P($\overline{R}$|d) est la probabilité qu'un document d soit non pertinent (noté $\overline{R}$ pour non-relevant).

La pondération w_i du terme t_i est alors donnée par l'équation (1.6) qui indique le poids du terme t dans le document d (k_1 et b sont des constantes, dl la longueur du document, dl_{avg} la longueur moyenne des documents).

$$w_i = \frac{tf(t,d) * (k_1 + 1)}{tf(t,d) + k_1 * (1)} * log \frac{N - df(t) + 0.5}{df(t) + 0.5} \tag{1.7}$$

k_1 est une constante permettant d'accroître ou de diminuer l'influence de la fréquence du terme tf_i sur w_i Ce modèle peut être vu comme un tfidf prenant en compte la longueur des documents.

Modèle de langue

Une approche alternative au problème de recherche d'information consiste à considérer qu'un document et une requête sont proches si et seulement s'il est aisé de générer la requête à partir du document [Ponte et Croft, 1998],[Croft et Lafferty, 2013]. Ce modèle repose sur l'idée que l'utilisateur, quand il formule sa requête, a une idée du document idéal qu'il souhaite retrouver et que sa requête est formulée pour ce document idéal. C'est une approche probabiliste qui assigne une probabilité à toute séquence de mots. On suppose en général que les mots qui apparaissent dans la requête sont indépendants. Ainsi, pour une requête Q = t_1,t_2...t_n, cette probabilité de génération est estimée comme suit :

$$P(Q|D) = P(t_1, t_2...t_n|D) = \prod P(t|D)^{c(t;Q)} \tag{1.8}$$

où c(t ;Q) est la fréquence du terme t dans la requête Q.

Comme proposé par Lafferty et al [Lafferty et Zhai, 2001], la similarité entre un document et une requête peut être également exprimée par la mesure de la divergence de Kullback-Leibler (KL-divergence). La KL-divergence s'interprète comme la différence moyenne entre deux distributions probabilistes. Il s'agit de mesurer le coût supplémentaire nécessaire pour encoder la requête dans le modèle du document.

Afin de ne pas assigner une probabilité nulle aux documents ne contenant pas les termes de la requête mais qui seraient susceptibles d'être pertinents par rapport à cette dernière, on effectue une opération de lissage de ce modèle (table1.3). Ici, c_d est un facteur normalisateur assurant l'égalité $\sum_{t \in C} \mathrm{P(t|d)} = 1$. Une des stratégies utilisées est le lissage de Jelinek-Mercer [Jelinek, 1976]. Il a été montré que ce lissage [Zhai et Lafferty, 2001] a aussi pour effet de modéliser la spécificité des termes de la requête. La méthode du **décompte absolu** ressemble à celle de Jelinek-Mercer, sauf que la probabilité des mots présents dans le document est réduite en soustrayant le nombre d'occurrences de ces mots par une constance $\alpha \in [0, 1]$ caractérisant le décompte. La méthode du **maximun à priori** (MAP) avec a priori de Dirichlet présume que les termes présents dans le document sont issus d'un tirage aléatoire avec remise suivant une distribution multinomiale.

Methode	P(t|d)	c_d
Jelinek-Mercer	(1 - λ)P_{mv}(t|d) + λP_{mv}(t|C)	λ
Décompte absolu	(max($tf_{t,d}$ - α,0) / $\sum_{t \in V}\ tf_{t,d}$) + βP_{mv}(t|C)	α|d|/l_d
MAP	$tf_{t,d}$+ γP_{mv}(t|C)/$\sum_{t \in V}\ tf_{t,d}$ + (γ	$\gamma/\sum_{t \in V}\ tf_{t,d}$) + γ

TABLE 1.3 – Méthodes de lissage
pour l'estimation de la probabilité d'apparition d'un terme dans un document.

Modèles Informationnels

Ce modèle provient de la constatation que la distribution des termes significatifs d'un document s'éloignait de manière non négligeable de la distribution des termes non significatifs [Harter, 1975]. Si nous considérons un terme qui apparaît peu souvent et dans peu de documents du corpus, alors le fait de l'observer avec une fréquence élevée dans un document prouve que ce document est pertinent à l'égard de ce terme. Plusieurs modèles de probabilités ont été utilisés. L'article de [Clinchant et Gaussier, 2010] présente deux distributions de probabilité et donc

deux modèles informationnels. Ces modèles de distribution sont la distribution log-logistique (LL) et une distribution de puissance lissée (SPL pour Smoothed Power Law).

$$s_{LL}(q,d) = \sum tf_{t,q} ln(\frac{\lambda_t + ntf_{t,d}}{\lambda_t}) \tag{1.9}$$

$$s_{SPL}(q,d) = \sum tf_{t,q} ln(\frac{1-\lambda_t}{\lambda_t^{\frac{ntf_{t,d}}{ntf_{t,d}+l}}}) \tag{1.10}$$

— $ntf_{t,d}$ est une fonction de normalisation qui dépend du nombre d'occurrences du terme t dans le document$d$

— λ est un paramètre dépendant du corpus qui peut être estimé sur la collection.

Les modèles probabilistes sont plus efficaces que les modèles booléens. Des auteurs [Crestani et van Rijsbergen, 1998] ont soulevé le problème majeur de ces modèles à savoir de trouver des méthodes permettant d'estimer les probabilités utilisées pour évaluer la pertinence qui soit fondées d'un point de vu théorique et efficaces au calcul. L'hypothèse de l'indépendance des termes est utilisée en pratique pour implémenter ces modèles.

1.2 Évaluation des Systèmes de RI

L'évaluation des systèmes de recherche d'information constitue une étape importante dans l'élaboration d'un modèle de recherche d'information (moteur de recherche par exemple). Elle permet de caractériser le modèle et de réaliser une comparaison entre les modèles existants.

Il existe plusieurs collections de tests de référence, créées dans le cadre de campagnes d'évaluation dont la plus connue est le programme TREC *(Text REtrieval Conference*). Elle existe depuis 1992 [Harman, 1993] ; [Voorhees et Harman, 2001] et a permis le développement de collections de tests. Une autre campagne relativement importante est la campagne CLEF [1] *(Conference and Labs of the Evaluation*

1. www.clef-initiative.eu

Forum), elle a été lancée dans le but de développer des systèmes de RI sur d'autres langues que l'anglais.

D'une façon générale, un système de recherche d'information a deux objectifs principaux :

— Retrouver les documents pertinents

— Rejeter tous les documents non pertinents

Les deux mesures d'évaluation les plus utilisées en RI sont le rappel et la précision.

1.2.1 Mesures de rappel/précision

Pour des résultats non ordonnés de recherche, le rappel est la fraction des documents pertinents retournés par le système, par rapport au nombre de documents pertinents existant, pour une recherche d'information donnée.

Ainsi, si l'on note S l'ensemble des documents qu'un système automatique considère comme ayant une propriété recherchée, V l'ensemble des documents qui possèdent effectivement cette propriété, nous avons la formule suivante pour le rappel :

$$Rappel = \frac{S \cap V}{V} \tag{1.11}$$

La précision se définit comme la fraction de documents pertinents par rapport au nombre total de documents, dans les résultats retournés par le système.

$$Precision = \frac{S \cap V}{S} \tag{1.12}$$

Nous appelons *bruit* des instances non pertinentes qui apparaissent dans les résultats d'une recherche. Le taux de bruit peut être une mesure de l'efficacité d'un système de recherche. Il constitue la base pour le calcul de la précision.

Les *silences* sont des documents pertinents présents dans le corpus mais non présentés par le système à l'utilisateur. Ils servent de base pour calculer le rappel.

Suivant les situations, on préférera privilégier l'une ou l'autre mesure. Si l'on

effectue une recherche via un moteur sur le web, on souhaite, quelle que soit la requête, que tous les documents trouvés soient pertinents, puisqu'on n'a pas accès au nombre total de documents du web qui seraient pertinents par rapport à la requête. Dans ce cas, on privilégie la précision.

Dans le cas où l'on souhaite obtenir un système qui ne privilégie aucune des deux mesures, on utilise la moyenne harmonique pondérée entre le rappel et la précision (la F-mesure).

$$F_\beta = (1 + \beta^2)\frac{Precision * Rappel}{\beta^2 * Precision + Rappel} \tag{1.13}$$

Dans le cas où précision et rappel sont considérés comme étant d'égale importance la valeur de β sera égale à 1 et la mesure utilisée est appelée F_1.

$$F_1 = \frac{2 * precision * rappel}{precision + rappel} \tag{1.14}$$

Pour l'évaluation des résultats ordonnés, nous utilisons une courbe de précision-rappel. On obtient cette courbe en mesurant le rappel et la précision sur un sous-ensemble de documents, de taille croissante, obtenus à partir d'une liste de documents ordonnés. L'ensemble des mesures de rappel et de précision ainsi obtenues permet de tracer la courbe de la précision, P(r) en fonction du rappel r (tableau 1.4). Soit un moteur de recherche M répondant à une requête q. Dans la liste de réponse de M, 4 documents sont jugés pertinents sur les 10 documents retournés par le système.

A un rang *rg* donné sur la liste des documents ordonnés, si le document de rang (*rg + 1*) est considéré comme non pertinent, le rappel au rang *rg + 1* sera le même que celui du rang *rg*, alors que la précision sera diminuée. Dans le cas inverse, si le document de rang (*rg + 1*) est pertinent alors le rappel et la précision augmenteront. La figure 1.4 donne la courbe précision-rappel correspondante aux valeurs de la table 2.1 .

Document	Réponse de M	Pertinence	rappel	précision
D1	0.95	1	1/4	1
D2	0.82	0	1/4	1/2
D3	0.75	0	1/4	1/3
D4	0.7	1	1/2	1/2
D5	0.65	1	3/4	3/5
D6	0.5	0	3/4	1/2
D7	0.4	0	3/4	3/7
D8	0.35	1	1	1/2
D9	0.2	0	1	4/9
D10	0.1	0	1	2/5

TABLE 1.4 – Mesure du rappel et de la précision
sur un ensemble de 10 documents ordonnés. Si un document est considéré comme pertinent, il a une valeur de 1, 0 sinon.

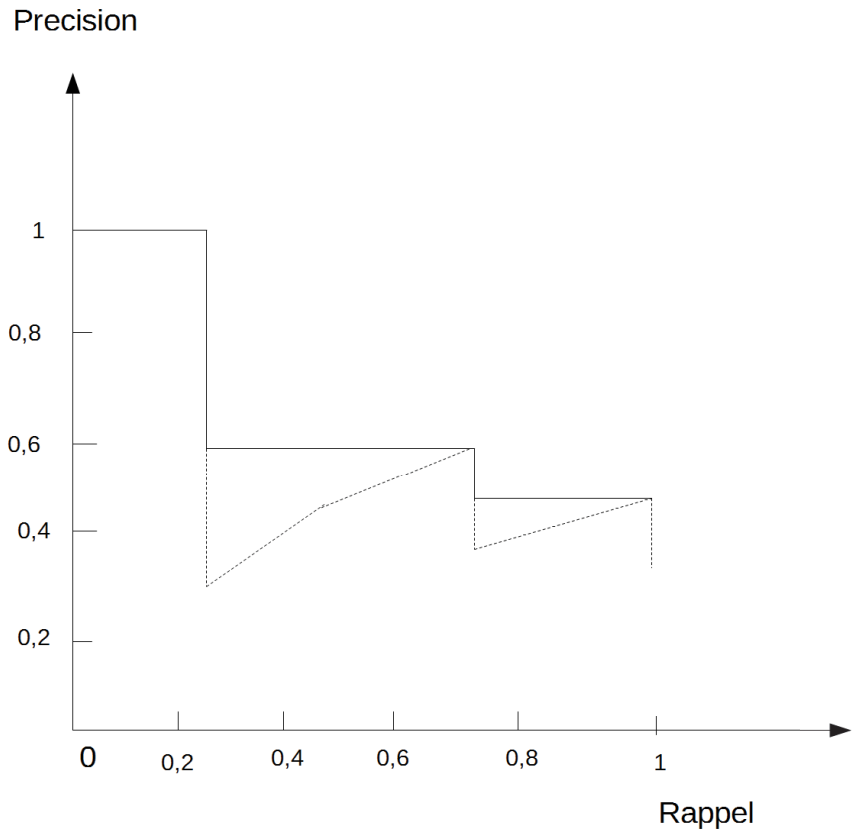


FIGURE 1.4 – Exemple d'une courbe précision-rappel
(en pointillé) et précision interpolée(en trait plein)

Pour enlever ces oscillations, on effectue un lissage de la courbe précision-rappel en remplissant un creux de la courbe par une ligne horizontale. Cette mesure est connue sous le nom de *précision interpolée*.

1.2.2 Précision moyenne et gain cumulatif réduit

Une autre mesure utilisée est la précision moyenne (AveP) (*Average Precision* en anglais). C'est la moyenne des valeurs de précision des documents pertinents par rapport à la requête q dans la liste ordonnée des réponses. La formule de la précision moyenne est la suivante :

$$AveP(q) = \frac{1}{n_+^q} \sum_{k=1}^{N} R_{d_{k,q}} XP@k(q) \tag{1.15}$$

avec $n_+^q = \sum_{k=1}^{N} R_{d_{k,q}}$ est le nombre total de documents pertinents par rapport à la requête q. P@k(q)est la précision moyenne à un rang k donné, définie comme :

$$P@k(q) = \frac{1}{k} \sum_{rg=1}^{k} R_{d_{rg,q}} \tag{1.16}$$

où $R_{d_{rg,q}}$ qualifie le jugement de pertinence du document d de rang rg dans la liste des documents retournés par rapport à une requête.

Dans les cas où on a un ensemble de requêtes Q = q_1....$q_{|Q|}$, il est possible de calculer la moyenne des précisions moyennes (MAP mean average precision) sur l'ensemble des requêtes :

$$MAP = \frac{1}{|Q|} \sum_{j=1}^{|Q|} AveP(q_j) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{n_+^q} \sum_{k=1}^{N} R_{d_{k,q}} XP@k(q) \tag{1.17}$$

Cette mesure peut être qualifiée de globale puisqu'elle combine différents points de mesure. Elle est moins sensible au nombre de documents que d'autres mesures (par exemple la précision à différents niveaux de coupes). C'est la moyenne des

précisions obtenues chaque fois qu'un document pertinent est retrouvé. Cette mesure a été introduite dans TREC2 [Moffat et Zobel, 2008] pour sa capacité à résumer les mesures de précision aux 11 points de rappel.

Dans le cas où le jugement de pertinence n'est pas binaire mais prend des valeurs discrète de N, il existe une mesure de l'efficacité du système de recherche appelée le gain cumulatif réduit (DCG pour *Discounted Cumulative Gain* en anglais). Il peut être souhaitable d'adopter une échelle de pertinence allant par exemple de 0 à 5 : 0 pour les documents non pertinents, 5 pour les documents parfaits, les degrés intermédiaires pour les documents plus ou moins pertinents. La mesure DCG par rapport à une requête q est calculée à un rang p.

$$DCG_p = \sum_{i=1}^{p} \frac{2^{R_{d_{rg},q}} - 1}{log_2(1 + rg)} \tag{1.18}$$

1.2.3 La courbe ROC

La courbe *ROC* (de l'anglais *Receiver Operating Characteristic*) mesure la capacité d'un moteur de recherche à ordonner les documents pertinents par rapport à une requête q avant les documents non pertinents. Cette courbe se construit, à partir de la liste de documents ordonnés, en mesurant le rappel à chaque rang de cette liste, en fonction du nombre de documents non pertinents ordonnés avant ce rang.

La courbe ROC est largement utilisée dans beaucoup de domaines comme la médecine, la biométrie, l'apprentissage machine, etc. L'aire sous la courbe (*Area Under Curve (AUC)*) est un indicateur de la capacité d'un système à ordonner les documents pertinents au dessus des non pertinents.

1.3 Les différentes étapes de la recherche d'information

Pour répondre aux besoins des utilisateurs, un système de recherche d'information doit mettre en œuvre un certain nombre de processus pour réaliser une mise en correspondance des informations contenues dans une collection de documents et des besoins en information des utilisateurs.

1.3.1 Le processus d'indexation

Pour assurer la recherche d'information dans des conditions acceptables, une étape importante consiste à analyser le document pour produire un ensemble de mots-clés, appelés aussi descripteurs, afin que le système puisse les gérer facilement dans le processus de recherche. C'est ce qu'on appelle l'indexation [Deerwester *et al.*, 1990]. L'utilisation d'une structure qui fait correspondre chaque terme du vocabulaire à la liste des documents du corpus qui le contiennent est la manière la plus pratique pour trouver le terme d'une requête donnée dans un corpus de documents. Cette structure est appelée index inversé [Zobel *et al.*, 1998]. Certains index indiquent le nombre de documents contenant les termes, leur nombre d'occurrences dans les documents ou la position des termes dans les documents (figure 1.5) [Mitra *et al.*, 1997] a comparé l'indexation à base de mots simples et de groupes de mots.

ID	Texte
1	L'hiver est propice à tomber malade
2	La grippe est une maladie
3	Cette maladie est bénigne et se guérit vite
4	Il a attrapé la grippe pendant l'hiver

Corpus

Termes	Freq	Document ID
hiver	2	[1][4]
propice	1	[1]
tomber	1	[1]
grippe	2	[2][4]
maladie	2	[2][3]
bénigne	1	[3]
guérit	1	[3]
vite	1	[3]
attrapé	1	[4]
pendant	1	[4]

Index inversé

FIGURE 1.5 – Exemple d'un index inversé.
A gauche le corpus et à droite les paires d'identifiants ainsi que la fréquence d'apparition dans le document.

La création d'un index inversé est constituée de différentes étapes :

1. l'extraction de paires d'identifiants (terme, documents).
2. le tri des paires suivant les clés d'identifiants des termes puis de celle des documents.
3. le regroupement des paires en établissant pour chaque identifiant de terme, la liste des paires identifiants de documents dans lequel le terme apparaît.

L'indexation peut prendre plusieurs formes :
— une indexation manuelle dans laquelle chaque document du corpus est indexé par un spécialiste du domaine.
— une indexation automatique à l'aide d'un processus entièrement informatisé.
— indexation semi-automatique où l'extraction se fait de manière automatique mais le choix final est laissé aux spécialistes du domaine.

L'indexation manuelle présente l'avantage d'assurer une meilleure correspondance entre les documents du corpus et les termes choisis par les indexeurs. Ceci donne une meilleure précision en ce qui concerne les documents retournés par un système de RI suite à une requête précise [Ren *et al.*, 1999]. L'inconvénient principal de cette méthode d'indexation est le temps et l'effort qu'elle exige. De plus, une part de subjectivité lié au facteur humain fait que pour un même document, des termes différents peuvent être sélectionnés par des indexeurs différents.

Dans l'indexation semi-automatique [Jacquemin *et al.*, 2002] les indexeurs utilisent en général un vocabulaire contrôlé sous forme d'ontologie ou de base terminologique. C'est le cas par exemple de l'indexation d'articles médicaux à l'aide du thésaurus MeSH. Les termes dans les bases terminologiques, sont préordonnés c'est-à-dire qu'ils peuvent être liés par des relations hiérarchiques tels que la relation d'hyperonymie ou par un ensemble plus riche de relations lexico-sémantiques.

L'indexation automatique comprend un ensemble de traitements sur les documents. Les différentes étapes sont, en général, l'extraction automatique des descripteurs, l'utilisation d'un anti-dictionnaire, la lemmatisation, le repérage des groupes de mots, la pondération des termes avant de constituer l'index.

Pondération des termes

Le poids d'un terme dans un document traduit l'importance de ce terme à l'intérieur du document. Une des lois importantes en recherche d'information est la loi de Zipf [Zipf, 1935]. Cette dernière spécifie que la fréquence d'occurrence d'un mot m dans un corpus de documents est inversement proportionnelle à son rang (r), ce dernier est obtenu quand on trie les mots du corpus par ordre décroissant des fréquences.

$$f_m = \frac{K}{r} \tag{1.19}$$

r représente le rang et K est une constante. Ce paramètre dépend de la collection considérée. Une conséquence de cette loi est que la fréquence des mots décroît rapidement avec leur rang. Dans le domaine de la recherche d'information, la loi de Zipf est utilisée pour réaliser un filtrage afin de déterminer les mots qui représentent au mieux le contenu d'un document.

Le poids d'un terme s'évalue par deux paramètres : le nombre d'occurrences de

ce terme dans le document (*term frequency*), et le nombre de documents dans lequel ce terme apparaît, (*document frequency*) (df) [Robertson, 2004].

Définitions : **Fréquence du terme (tf)** :
Nombre d'occurrences d'un terme dans un document du corpus. tf_{ij} correspond au nombre d'occurences du terme t_i dans le document D_j.

$$tf_{i,j} = \frac{n_{i,j}}{\sum n_{k,j}}$$

Définitions : **Fréquence inverse de document (IDF)** :
Mesure l'importance d'un terme dans l'ensemble du corpus.

$$idf_i = \log \frac{|D|}{|d_j : t_i \in d_j|}$$

où :

- |D| représente le nombre total de documents dans le corpus
- $|d_j : t_i \in d_j|$ est le nombre de documents où le terme t_i apparaît.

En recherche d'information, le codage le plus courant des documents est le codage *tfidf* :

$$tfidf_{ij} = tf_{ij} * idf_i \tag{1.20}$$

Des variantes à ce codage ont été proposées pour améliorer cette fonction de pondération [DeClaris *et al.*, 1994]. Si on triple la taille d'un document, les poids des termes seront triplés et ce document sera considéré comme plus pertinent à la suite d'une requête. Ces variantes utilisent souvent un terme de normalisation, dépendant de la taille du document considéré pour corriger ce biais. Certains auteurs ([Robertson et Walker, 1997] ont proposé de normaliser la fonction tf-idf de la manière suivante :

$$wd_{ij} = \frac{tf_{ij}.(K+1)}{K_1.((1-b) + b.\frac{dl_i}{\Delta l} + tf_{ij}} \tag{1.21}$$

Où wd_{ij} est le poids du terme t_i dans le document D_j. Le paramètre K contrôle l'influence de la fréquence du terme t dans le document d, sa valeur dépendant de la longueur des documents dans le corpus ; b ([0, 1]) est une constante contrôlant l'effet de la longueur du document. Δl représente la longueur moyenne des documents dans le corpus et dl représente la longueur du document d.

Depuis les années 1970, des chercheurs se sont penchés sur sur l'intérêt d'utiliser des ressources lexico-sémantiques dans le processus d'indexation. L'intérêt se justifie par le souci d'un meilleur contrôle et uniformisation du langage d'indexation. Nous reviendrons plus en détail sur ce point dans la partie 2.3 du manuscrit.

1.3.2 Les requêtes

Le processus d'appariement document-requête permet de mesurer la pertinence d'un document par rapport à une requête. Le système calcule un score traduisant la correspondance entre les documents retournés et la requête. Ce score représente le degré de pertinence vis-à-vis du document. On calcule cette valeur de pertinence grâce à une fonction de similarité RSV (Q, d)(*Retrieval Status Value* [Nottelmann et Fuhr, 2003]) dans laquelle Q représente une requête et d est un document du corpus. Cette mesure prend en compte le poids des termes.

Souvent, en recherche d'information il peut être intéressant de réaliser une reformulation de requêtes ou une expansion de requêtes [Efthimiadis, 1996]. La reformulation de requêtes a pour objectif de modifier la requête de l'utilisateur par ajout de certains termes (synonymes, hyperonymes...) ou par réestimation de leur poids. Les approches locales d'expansion de requêtes enrichissent le plus souvent la représentation vectorielle de la requête à partir de termes provenant des documents de la collection. Les deux méthodes les plus connues sont la boucle de rétropertinence (*relevance feedback*) [Hiemstra et Robertson, 2001] et la boucle de rétropertinence en aveugle (*pseudo relevance feedback*) [Boughanem *et al.*, 1999], [Cao *et al.*, 2008]. La méthode *boucle de rétropertinence* implique que l'utilisateur dans le processus de recherche, fournit un retour sur la pertinence des documents renvoyés lors d'un premier appariement entre la requête et les documents du corpus. L'utilisateur précise quels documents de cette liste sont pertinents ou non pertinents par rapport à la requête initiale et le système enrichit alors la représentation vectorielle de la requête en tenant compte de ce retour. Ces étapes sont répétées jusqu'à ce que l'utilisateur soit satisfait des résultats finaux. La méthode *boucle de rétropertinence en aveugle* permet d'éviter la phase d'annotation en supposant que les k (souvent fixés à 5, 10 ou 20) premiers documents fournis par le système sont pertinents par rapport à la requête. Cette hypothèse rend automatique l'approche précédente. Cette mo-

dification de requêtes peut aussi être basée sur l'utilisation de ressources externes telles que des ontologies ou des thésaurus. C'est ce qu'on appelle une reformulation basée sur les concepts (*Concept-based Query Reformulation*) [Fonseca *et al.*, 2005], [Qiu et Frei, 1993], [Järvelin *et al.*, 2001].

Conclusion

Nous avons évoqué dans ce chapitre les principales notions de la recherche d'information. Nous avons décrit les principales étapes d'un processus de recherche. Les principaux modèles existant dans la littérature ont été présentés ainsi que les différentes méthodes d'évaluation des performances d'un système de recherche d'information.

Les systèmes de recherche d'information (moteurs de recherche) ont été conçus, au départ, pour retrouver les documents contenant exactement les termes précis de la requête. Ceci est évidemment insuffisant dans la mesure où l'utilisateur n'emploiera pas exactement les mêmes termes que ceux du document (synonymes, termes génériques). Dans le chapitre 4, nous présenterons un modèle d'augmentation d'index par propagation à travers un réseau lexical, précisément conçu pour contourner cette faiblesse des systèmes de recherche.

Chapitre 2

Indexation sémantique et bases de connaissance : un état de l'art

Sommaire

2.1 Définition de l'indexation sémantique 42
2.1.1 Indexation pour des textes généraux 43
2.1.2 Indexation dans le domaine médical 47
2.1.3 Indexation par propagation 49
2.2 Critère d'évaluation de l'indexation 50
2.2.1 Consistance de l'indexation 51
2.2.2 Exactitude 51
2.2.3 Qualité de l'indexation 52
2.3 Utilisation de bases de connaissances 52
2.3.1 Définition des ontologies 53
2.3.2 Définition d'un réseau sémantique 56
2.4 Extraction de relations et le TALN dans le domaine médical 71
2.4.1 Extraction de relations 72
2.4.2 Traitement automatique du langage naturel dans le domaine médical 74

Dans le chapitre précédent, nous avons décrit les principaux modèles de recherche d'information qu'on trouve dans les systèmes actuels. Dans ce chapitre, nous consacrons une grande partie à l'état de l'art de l'indexation sémantique ainsi qu'à l'utilisation de bases de connaissances dans les processus de recherche et d'indexation. Dans un premier temps (2.1) nous définissons la notion d'indexation et plus particulièrement celle d'indexation sémantique à la fois dans le domaine général (2.1.1) et dans le domaine médical (2.1.2). La section (2.2) introduit les critères d'évaluation de l'indexation. Nous commençons par définir la consistance de l'indexation (2.2.1), ensuite nous parlerons de l'exactitude et de la qualité de l'indexation (2.2.2 et 2.2.3).

La section (2.3) introduit les ressources sémantiques externes. Nous décrivons les principaux types de ressources sémantiques à savoir les ontologies (2.3.1) et les réseaux sémantiques (2.3.2).

2.1 Définition de l'indexation sémantique

L'indexation est un traitement qui consiste à caractériser un document à partir de son propre contenu, des mots-clés qui le caractérisent. On fera correspondre ces mots à la liste des documents qui les contiennent. Cette structure s'appelle un index inversé et c'est une des façons les plus rapides pour retrouver un terme d'une requête dans un corpus de documents. Les mots clés ou descripteurs peuvent prendre

plusieurs formes : mots simples, mots composés, lemmes ou toute autre unité d'information décrivant le contenu du document. L'indexation peut se faire soit de façon manuelle (indexation manuelle) soit de manière automatique (indexation automatique). L'indexation manuelle fait appel à des spécialistes du domaine possédant une certaine expertise, qui vont sélectionner les descripteurs appropriés pour chaque document. Cette méthode est très coûteuse en temps et en ressources humaines. Au contraire, l'indexation automatique fait appel à des algorithmes pour sélectionner les descripteurs les plus pertinents. Elle comprend un ensemble de traitements des documents dont, entre autres, la lemmatisation (identifier le lemme (unité autonome constituante du lexique d'une langue) d'un mot), le repérage des groupes de mots, le filtrage (élimination des mots vides) par un anti-dictionnaire (*stopword list*), et l'extraction automatique des descripteurs. Un anti-dictionnaire ou liste de *mots vides* est une liste de mots qui sont présents avec une fréquence élevée dans un corpus de documents et qui n'apportent aucune information sur le contenu d'un document. Ces anti-dictionnaires peuvent être soit indépendants soit dépendants du domaine du corpus. Dans un corpus formé de documents de descriptifs de *projets*, le terme *projet* sera un exemple de mot vide dans un anti-dictionnaire dépendant du domaine. Dans le cas des anti-dictionnaires généraux, c'est à dire indépendant des domaines étudiés, on se repose plutôt sur la catégorie grammaticale des mots, l'anti-dictionnaire sera alors composé des mots autre que nom, adjectif, adverbe, verbe.

de	la	le	et	en	du	des
au	aux	avec	dans	par	pour	sur
quel	quelle	pas	s'	qu'	ou	se
ou	sont	plus	y	est	elle	ils

TABLE 2.1 – Exemple de mots présents dans un anti-dictionnaire général.

2.1.1 Indexation pour des textes généraux

2.1.1.1 Extraction des descripteurs

Le processus d'indexation consiste à extraire des descripteurs qui représentent au mieux le document [Salton *et al.*, 1983]. Ils peuvent indiquer le ou les thèmes objets du document [Laporte, 2000], mais également des informations sémantiques.

Les descripteurs peuvent prendre plusieurs formes :

— les *mots simples* du document à l'exclusion des mots de l'anti-dictionnaire.

— les *lemmes* ou les racines des mots (*racinisation*).

— les *lexèmes* qui sont des unités minimales de signification appartenant au lexique.

— souvent, les termes composés présents apportent plus de sens (sémantique) que les mots qui le composent pris séparément. En effet, comme exemple le terme composé *pied de biche* est plus précis que les trois mots *pied de biche* pris séparément.

  La reconnaissance de termes composés peut être une source de problèmes. Certains indexeurs utilisent une source de mots composés du langage pour les identifier comme ne formant qu'un seul mot [Habert et Jacquemin, 1993]. Dans la littérature trois principales approches existent :

  1. des approches basées sur la cooccurrence [Alvarez *et al.*, 2004], [Lauer, 1995]. La pondération tfidf est utilisée pour éliminer les groupes de mots fréquents ("dans ce cas là",...).

  2. des approches linguistiques qui se basent sur une analyse syntaxique partielle ou l'utilisation de patrons syntaxiques pour détecter les termes composés [Bourigault *et al.*, 1996].

— les concepts, qui sont des expressions contenant un ou plusieurs mots. La définition du terme concept varie légèrement selon les domaines (linguistique, informatique...). Nous reviendrons en détail sur la notion de concept plus loin.

— les N-grammes [Brown *et al.*, 1992], qui sont formés de sous-séquences de n éléments (unigramme, bigramme, trigramme...) à partir de séquences données. Cette approche dans la recherche d'information [Millar *et al.*, 2006] est retrouvé, par exemple pour certaines langues asiatiques [Lee *et al.*, 1999].

— l'augmentation d'index ou contexte où les descripteurs peuvent être des termes n'apparaissant pas dans le document mais possédant un lien sémantique ou de cooccurrence [Deerwester *et al.*, 1990] (*Latent Semantic Indexing*) avec les termes du document. L'ajout de termes liés sémantiquement aux mots du documents est rendu possible par la disponibilité de plus en plus importantes d'ontologies et autres ressources sémantiques externes (par exemple, un réseau lexico-sémantiques).

2.1.1.2 Indexation sémantique et conceptuelle

Jusqu'à présent, nous avons décrit une représentation des documents essentiellement basée sur des mots (ou mots composés) permettant de représenter l'information contenue dans un corpus de documents. C'est ce que l'on appelle une *indexation classique*. Certains travaux [Guarino *et al.*, 1999] ont montré l'insuffisance de la représentation basée sur les mots simples. Pour pallier cette insuffisance, ils proposent une exploitation sémantique des textes afin de mieux les représenter. Deux grandes familles de travaux permettent d'ajouter de l'information sémantique dans le processus de recherche d'information. Parmi ces travaux, il est possible de distinguer deux grandes tendances, à savoir l'indexation conceptuelle [Woods, 1997] et l'indexation sémantique [Mihalcea et Moldovan, 2000], [Biemann, 2005]. Cette catégorisation peut cependant susciter une certaine confusion dans la terminologie, parmi la communauté des chercheurs. On parle plutôt d'indexation sémantique lorsqu'on souhaite utiliser le sens des mots pour réaliser une indexation des documents [Sanderson, 1994]. Nous pouvons définir l'indexation conceptuelle comme recouvrant toute connaissance ajoutée à un document pouvant servir dans le cadre de "calculs" sous-tendus par l'exploitation de ces documents, pourvu que cette connaissance soit utilisable aussi bien par l'homme que par la machine [Prié, 2000]. Selon cette définition, l'indexation conceptuelle permet une représentation des documents via différentes relations sémantiques, alors que dans le cadre de l'indexation sémantique, on se limite le plus souvent à l'utilisation de la synonymie. Historiquement l'indexation sémantique utilise les techniques de désambiguïsation (WSD) [Krovetz, 1997], [Mihalcea, 2004] pour associer un sens à un mot alors que l'indexation conceptuelle concerne plutôt l'identification de concepts dans un corpus (utilisation dans le domaine biomédical par exemple).

Indexation sémantique

L'approche sémantique permet de détecter le sens d'un mot à partir de son contexte d'apparition (tâche de désambiguïsation). Pour désambiguïser le sens, certains auteurs [Leacock *et al.*, 1998] [Hirst et St-Onge, 1998] utilisent des représentations hiérarchiques pour calculer une distance sémantique ou une similarité sémantique entre les termes à comparer. D'autres auteurs [Mihalcea et Moldovan, 2000] ont utilisé une méthode de désambiguïsation basée sur un corpus pré-étiqueté et

les synsets de WordNet [Miller, 1995]. Un nouveau mot est désambiguïsé en tenant compte de sa relation avec les mots du corpus déjà désambiguïsés. Par cette méthode les auteurs ont noté une amélioration dans le rappel et la précision alors que dans [Gonzalo *et al.*, 1998], les auteurs ont réalisé des expériences d'indexation basées sur le sens et sur les synsets avec également des performances accrues. Cependant certains auteurs [Voorhees, 1993] [Katz *et al.*, 1998] ont noté que l'utilisation d'un désambiguïsateur automatique n'améliorait pas de façon sensible les performances. [Voorhees, 1993] a remarqué, dans le cadre d'une campagne TREC, que l'expansion de requêtes avec les synsets de WordNet améliorait quelques requêtes mais en dégradait d'autres. [Katz *et al.*, 1998] ont utilisé la notion de contexte pour effectuer la tâche de désambiguïsation. Ils utilisent un corpus d'entraînement ainsi que WordNet. Ils ont testé leur méthode sur le corpus Semcor [1] et ont obtenu une précision de 60%.

Indexation conceptuelle

Les approches conceptuelles consistent à rattacher les termes à des concepts sous-jacents. Les concepts sont issus d'une liste prédéfinie des termes d'index ou d'un vocabulaire contrôlé appelé système de classification pouvant prendre la forme d'ontologies, de thésaurus, de taxonomies [Stairmand et Black, 1996] [Stein *et al.*, 1997]. Les différentes méthodes utilisées ont pour objectif de rattacher les termes d'un document aux concepts d'une ontologie [Khan, 2000]. Il propose un algorithme fondé, dans une étape, sur la cooccurence entre termes et ensuite sur la proximité sémantique. Baziz [Baziz *et al.*, 2005] relie les termes d'un texte aux concepts de WordNet en se basant sur la notion de similarité sémantique entre concepts. Boubekeur [Boubekeur, 2008] représente les documents par un ensemble de concepts et de relations extraits d'une ontologie. Toujours pour une tâche de désambiguïsation [Hassell *et al.*, 2006] utilise les relations entre entités dans une ontologie pour déterminer le sens d'un mot.

Les travaux sur l'approche conceptuelle ont été appliqués dans plusieurs domaines comme par exemple le domaine légal [Stein *et al.*, 1997], ou encore le domaine du sport [Khan, 2000]. Le système **MetapMap** [Aronson, 2001] se basant sur le métathésaurus UMLS (Unified Medical Language System) [Bodenreider, 2004] est utilisé dans le domaine médical. Nous y reviendrons en détail dans la suite de ce

1. http ://web.eecs.umich.edu/ mihalcea/downloads.html#semcor

manuscrit. En ce qui concerne le domaine général, nous pouvons citer le systeme FERRET [Mauldin, 1991] ou encore les travaux de Woods [Woods, 1997]. Ces derniers apportent une évaluation relativement complète des approches d'indexation conceptuelle.

2.1.2 Indexation dans le domaine médical

Dans cette partie, nous allons nous focaliser sur les travaux d'indexation des documents biomédicaux. L'augmentation de la quantité d'information biomédicale, et en particulier des articles publiés dans MEDLINE [2] ont amené à la création d'outils d'indexation tel que MTI (*Medical Text Indexer*) [Aronson, 2001]. De plus, MEDLINE a été mis en ligne et est accessible librement à tous les utilisateurs d'internet par l'intermédiaire de PubMed [3]. Bien que le moteur de recherche de PubMed soit très utilisé, il ne permet pas de faire des recherches plus ciblées comme la recherche de gènes ou de protéines, ou encore de caractéristiques d'une maladie par exemple. En ce qui concerne l'indexation des ressources biomédicales francophones, elle implique souvent plusieurs ressources terminologiques médicales (MESH,ICD-10,...) [Névéol *et al.*, 2006] [Pereira *et al.*, 2008]. Dans ce qui suit, nous présentons quelques systèmes d'indexation qui peuvent utiliser une ou plusieurs ressources terminologiques.

Pour l'indexation de documents francophones, le système **MAIF** (*MEsh Automatic Indexing for French*) [Névéol *et al.*, 2006] permet l'indexation des articles en entier. A partir d'une adresse (URL) le système télécharge les documents et les indexe avec les termes français du MeSH. Les termes MeSH/qualificatifs sont identifiés en utilisant des patrons d'extraction définis dans le logiciel INTEX [Silberztein, 1999].

Le système **NOMINDEX** proposé par Pouliquen [Pouliquen, 2002] a pour objectif l'indexation de documents biomédicaux par des concepts issus d'une ressource termino-ontologique. Dans une première étape, les mots du document à indexer sont mis en correspondance avec les termes de l'Aide au Diagnostic Médical [Lenoir *et al.*, 1981]. Dans un second temps, les termes sont rattachés à leurs équivalents dans MeSH ainsi qu'à ceux de l'UMLS. Pour chaque concept identifié un

2. https ://www.nlm.nih.gov/pubs/factsheets/medline.html
3. http ://www.ncbi.nlm.nih.gov/pubmed

score basé sur *TFIDF* est calculé. Cette analyse est utilisée dans l'indexation, la recherche de documents similaires ainsi que dans la synthèse de documents.

Dans le cadre de l'indexation des articles de **MEDLINE**, l'outil MetaMap [Aronson, 2001] a été développé en utilisant une indexation multi-terminologique. C'est un analyseur morphosyntaxique lié à l'UMLS qui permet d'extraire des concepts à partir de documents (en anglais). Il applique des méthodes de recherche de variations de termes pour identifier les concepts du Metathésaurus qui s'apparient exactement ou approximativement à une expression donnée. [Aronson *et al.*, 2004] ont développé un outil d'indexation nommé *Medical Text Indexer* (MTI). Ils extraient les concepts issus de plusieurs terminologies de l'UMLS et ne retiennent que les concepts MeSH pour indexer les articles de MEDLINE (figure 2.1).

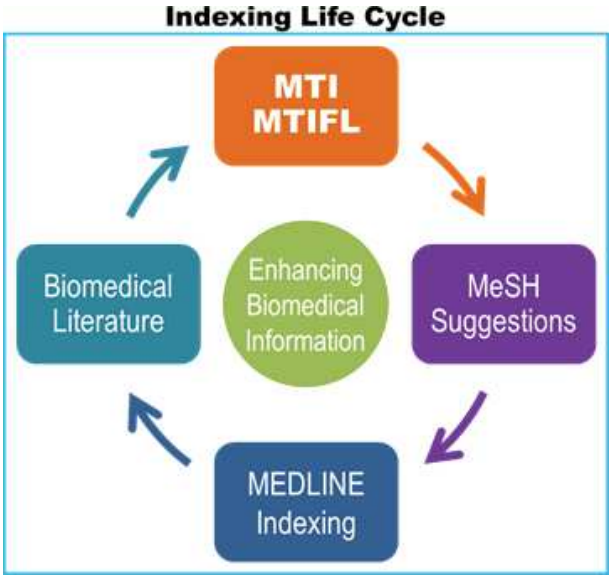


Figure 2.1 – cycle de vie de l'indexation MTI

Pour le français l'outil **F-MTI** (*French Multi-Terminology Indexer*) a été développé [Pereira *et al.*, 2009]. Ce système a été conçu pour réaliser l'indexation de dossiers médicaux en utilisant plusieurs terminologies à savoir la CIM-10, la CCAM (Classification Commune des Actes Médicaux), le thésaurus MeSH, la nomenclature

SNOMED [4] ainsi que la terminologie interne de la société Vidal [5].

Nous pouvons citer le système conçu par Avillach [Avillach *et al.*, 2007] qui pour indexer les résumés de sortie de l'hôpital a proposé une méthode d'indexation multi-terminologique. Les termes extraits sont classés afin de mettre en évidence leur importance dans le document. L'importance d'un terme est déterminée par le nombre de relations qu'il partage avec d'autres mots désignant des concepts. Les relations entre concepts peuvent être exploitées à partir du thésaurus UMLS mais aussi en exploitant des relations de cooccurrences entre les concepts issus d'une ou plusieurs terminologies (ICD-10). Les auteurs ont constaté une amélioration de la précision grâce à leur méthode d'indexation.

2.1.3 Indexation par propagation

La nécessité d'associer des métadonnées à des documents numériques a conduit à imaginer des mécanismes pour déduire de nouvelles indexations à partir d'indexations existantes [Lazaridis *et al.*, 2013]. Par ailleurs, [Marchiori, 1998] a proposé une propagation des métadonnées, en utilisant les hyperliens des pages Web comme support de cette propagation. Les métadonnées sont propagées à celles qui sont en lien avec elles, avec un poids d'affaiblissement. L'inconvénient de cette méthode, focalisée sur les liens de référencement et non sur le contenu des pages, est la propagation éventuelle d'erreurs (si une page est citée comme contre-exemple). L'indexation par propagation concerne principalement le domaine de l'image. Elle utilise à la fois l'analyse d'image et des indexations existantes pour en déduire, par diverses méthodes statistiques, des termes candidats pour indexer une nouvelle image [Jeon et Manmatha, 2004], [Zhang et Su, 2001]. Toujours dans le domaine du multimédia, [Pastorello Jr *et al.*, 2008] propose une propagation d'indexation sémantique reposant sur les techniques du Web Sémantique (RDF). Dans le domaine du TALN, il a été proposé une propagation dans un réseau pour réaliser une capture de mots-clés connexes [Lafourcade, 2011]. Les mots clés connexes sont définis comme étant fréquents dans la langue et ayant une distance angulaire faible d'au moins un des mots clés extraits à partir d'un texte. A chaque terme du texte est associé un en-

4. https ://www.nlm.nih.gov/research/umls/Snomed/
5. https ://www.vidal.fr/

semble pondéré de termes. Cet ensemble représente une approximation du voisinage du terme dans l'espace construit à partir du réseau. Une somme itérée des voisinages booléens des k mots clés trouvés au cours de la première étape est réalisée. Ils retiennent au plus k termes connexes en éliminant ceux déjà trouvés et ceux dont la valeur est inférieure ou égale à 1. Considérons l'exemple le segment de phrase : accident de moto, nous obtenons :

- motocyclette : 3
- accident de la voie publique : 2
- route : 2
- polytraumatisé : 2
- AVP : 2

En ce qui concerne la propagation d'index dans le domaine du TALN appliquée à la médecine, nous pouvons citer les travaux de [Fiorini *et al.*, 2014] dans lesquels ils appliquent une propagation d'indexation par l'intermédiaire d'une carte sémantique. La carte sémantique est une représentation visuelle par laquelle on identifie un réseau d'idées, d'informations ou de connaissances, ainsi que les relations qui unissent ces différents concepts entre eux. La carte sémantique peut permettre de retenir les idées importantes d'un document et les liens qui existent entre elles. Cette carte traduit une proximité sémantique des éléments dans l'espace d'indexation. Ils ont appliqué leur méthode sur un corpus de publications scientifiques dans le domaine de l'oncologie. L'indexation ne se fait qu'à une distance de un.

2.2 Critère d'évaluation de l'indexation

Les mesures de qualité de l'indexation peuvent être considérées selon deux points de vue :

- le point de vue du document : quelle est la qualité de l'indexation pour chaque document ?
- le point de vue du terme : est ce que chaque terme est correctement associé aux documents qu'il indexe ?

2.2.1 Consistance de l'indexation

Définitions : **consistance** :
La consistance vise à mesurer la concordance entre des indexations d'un même document par deux méthodes d'indexations différentes.

Que l'on raisonne d'un point de vue document ou terme, la consistance peut être mesurée en comparant les réponses :

— entre deux méthodes d'indexations A et B (consistance inter-indexeurs)
— entre deux indexations A et B différentes d'un même indexeur sur le même document à des moments différents (consistance intra-indexeur)
— entre deux indexations A et B différentes d'un même système d'indexation sur deux documents à sémantiques équivalente (consistance intra-indexeur).

Nous calculons :

consistance = nb termes affectés à D par les indexeurs A et B / nb termes affectés à D par les indexeurs A ou B

2.2.2 Exactitude

L'indexation est considérée comme exacte lorsqu'il n'y a aucune omission ou ajout d'éléments d'indexation. Un telle indexation est souvent impossible à générer. Il existe des mesures pour caractériser l'exactitude d'une indexation.

— La complétude : elle est liée à la présence des bons descripteurs. La complétude se définit comme le nombre de termes correctement affectés au document D sur le nombre de termes qui devraient être affectés à D.
— la pureté [Soergel, 1994] : elle est liée à l'absence de descripteurs erronés affectés aux documents.

2.2.3 Qualité de l'indexation

La consistance peut être considérer comme un critère de qualité de l'indexation. Mais ce critère seul ne peut pas être considéré comme suffisant pour évaluer la qualité d'une indexation. La difficulté majeure dans l'évaluation de la qualité d'une indexation est l'absence d'indexation de référence (*gold standard*) à laquelle comparer l'indexation à évaluer [Lancaster *et al.*, 1991]. Cependant nous pouvons utiliser deux méthodes pour évaluer cette qualité :

1) Comparer l'indexation à une indexation dite de référence élaborée par un expert.

2) Faire valider l'indexation par un expert

Un autre moyen indirect pour l'évaluation de l'indexation est de l'inclure dans un système de recherche d'information [Kim *et al.*, 2001].

2.3 Utilisation de bases de connaissances

Dans un premier temps, nous allons définir différentes notions utilisées dans ce manuscrit. Nous définissons, dans la thèse, la notion de **terme** comme un mot ou groupe de mots qui se réfère seulement à un objet ou une idée générale dans un domaine donné. Un terme formé d'un seul mot dans le sens de mot-forme [Melchuk et Gentilhomme, 2000] (*fièvre*, *fracture*, *os*, etc) est appelé **terme simple ou uniterme**, alors que celui formé de plusieurs mots est dit **terme complexe ou multi-terme** (*fracture ouverte*, *imagerie par résonance magnétique*, *moelle épinière*, etc).

La notion de concept est davantage sujette à caution suivant les communautés (sciences de la cognition, linguistiques, etc). Selon la définition du dictionnaire de l'Académie française le concept "regroupe les objets qu'il définit en une catégorie appelée classe". Selon Medin [Medin, 1989] un concept est une idée qui inclut tout ce qui est caractéristiquement associé à elle. En recherche d'information basé sur un processus d'indexation conceptuelle, les concepts sont le plus souvent prédéfinis

dans des structures conceptuelles comme dans des thésaurus ou les ontologies.

Par souci d'améliorer le contrôle et l'uniformisation du langage d'indexation, des chercheurs se sont posé la question d'utiliser des ressources lexico-sémantiques dans la phase d'indexation, ce pour améliorer les performances des systèmes de recherche d'information [Jones, 1987]. Depuis les années 90, la conception et le développement des ressources termino-ontologiques sont devenus un champ de recherche très actif en informatique, aussi bien dans le domaine de l'intelligence artificielle (IA) que celui de la recherche d'information. Ces ressources peuvent prendre plusieurs formes : hiérarchies de concept ou encore taxonomies comme dans MeSH [6], ontologies de domaine (Radlex [Langlotz, 2006]), ou ressources génériques comme par exemple WordNet [Miller, 1995],Cyc [Sowa, 1983], JeuxDeMots [Lafourcade, 2007].

2.3.1 Définition des ontologies

Guarino [Guarino, 1997] emploie le terme *ontologie* pour désigner une *compréhension partagée d'un domaine donné*. Gruber [Gruber, 1995] fournit quant à lui cette définition du terme ontologie :

Définitions : **ontologie** : Une ontologie est une spécification explicite et partielle rendant compte d'une conceptualisation.

Une ontologie organise les concepts indépendamment de la langue et des mots dans lesquels ces concepts s'incarnent. Dans ce cadre là, une ontologie peut être vue comme un modèle de données représentatif de la connaissance au sujet d'un monde ou d'une partie de ce monde. Un des objectifs d'une ontologie est de permettre aux machines (ordinateurs) de raisonner à propos des objets d'un domaine spécifique, et les ontologies doivent être partageables, cohérentes, normatives et consensuelles. Les divers éléments composant une ontologie sont d'après [Baziz *et al.*, 2005] :

— les concepts ;

— les relations entre les différents concepts ;

6. http ://www.nlm.gov :mesh/MBrowser.html

— les fonctions dans lesquelles le nième élément de la relation est défini de manière unique à partir des n-premiers ;
— les axiomes ;
— les instances qui sont utilisées pour structurer des éléments ;

A partir des définitions des ontologies, il est possible de distinguer plusieurs types de caractéristiques permettant de préciser ce qui peut être représenté dans une ontologie ainsi que la manière de la modéliser.

En ingénierie des connaissances, plusieurs types d'ontologies [Charlet *et al.*, 2004] liés à l'ensemble des objets conceptualisés sont répertoriés. On peut citer les ontologies de catégorie, générique et spécialisée.

2.3.1.1 L'ontologie de catégorie

Définitions : **ontologie de catégorie** :
L'ontologie de catégorie, aussi appelée Top-Ontologie est une ontologie structurant des connaissances de haut niveau rangées selon des notions philosophiques.

Dans cette catégorie on peut citer comme exemple SUMO (*Suggested Upper Merged Ontology*) [7] [Niles et Pease, 2001] dont l'objectif est de former un standard pour permettre l'interopérabilité sémantique entre tous les systèmes d'information, la recherche d'information, l'inférence automatisée et le traitement des langues naturelles [Pease *et al.*, 2002]. Un autre exemple est la méta-ontologie DOLCE (*Descriptive Ontology for Linguistic and Cognitive Engineering*) qui a pour but de concevoir des ontologies de domaines [Gangemi *et al.*, 2002].

2.3.1.2 L'ontologie générique

L'ontologie générique est une ontologie de type thésaurus qui fournit les concepts structurant un certain domaine et les relations entre les concepts. Elle présente la spécification d'un vocabulaire de référence (médecine, agriculture, etc). Par exemple,

7. http ://www.adampease.org/OP/

l'ontologie CIDOC (*Conceptual Reference Model*) [Doerr *et al.*, 2007] permet l'intégration des informations pour le patrimoine culturel et leur corrélation avec des données de l'archive.

Mikrokosmos [Mahesh *et al.*, 1996]quant à elle, est une ontologie qui entre dans le projet *Knowledge Base Machine Translation* (KBMT) ayant comme but de fournir un ensemble de concepts avec une abstraction de la langue.

2.3.1.3 L'ontologie spécialisée

Une ontologie spécialisée, encore appelée ontologie de domaine, présente les concepts d'un certain domaine (droit, finance, etc.) tels que définis par les experts du domaine.

On peut citer dans le domaine agricole AOS (*Agricultural Ontology Service*) [Lauser *et al.*, 2006]. Cette dernière est un centre fédéré pour les termes, définitions et relations dans les secteurs agricoles et domaines liés à la communauté agricole.

TOVE (TOronto Virtual Entreprise) [Fox, 1992] est une ontologie qui vise à créer des modèles d'entreprise et à modéliser des connaissances générales.

Mais une ontologie n'est pas seulement une classification de concepts, c'est aussi un ensemble de caractéristiques qui leur sont liées et qu'on appelle attributs (dans le contexte des langages à objets). La relation de subsomption *is-a*, qui définit un lien d'hyperonymie, est employée pour structurer les ontologies. Cependant, pour pouvoir exprimer la sémantique du domaine, elle doit être complétée par d'autres relations. Les relations relient les concepts pour construire des représentations conceptuelles.

Les ontologies les plus répandues sont UMLS (dans le domaine médical), EuroWordNet [Vossen, 1998], etc. En recherche d'information dans le domaine médical, UMLS [Zweigenbaum, 2004] et MeSH [Lowe et Barnett, 1994] occupent les premières places.

2.3.1.4 Les approches de construction d'ontologies

Vu l'importance des ontologies dans de nombreux domaines, différentes méthodes ont été proposées pour leur construction. Nous pouvons distinguer quatre catégories d'approches :

— construction de nouvelle ontologie à partir de rien ;
— construction d'ontologie à partir de textes ;
— construction d'ontologie basée sur des ontologies existantes ;
— construction d'ontologie basée sur le crowdsourcing ;

La méthodologie peut varier en fonction du domaine d'application et être manuelle ou semi-automatique. L'ingénierie ontologique a été marquée par l'approche de la construction d'ontologie à partir de texte à l'aide d'outils issus du traitement automatique du langage (TAL) [Philipp et Völker, 2005], [Aussenac-Gilles *et al.*, 2008]. Ces approches visent à alléger le processus de construction d'ontologie en automatisant certaines étapes grâce aux textes qui forment des sources de connaissances très riches.

2.3.2 Définition d'un réseau sémantique

Les réseaux sémantiques ont été développés au début des années 60 par Robert F. Simmons [Simmons, 1963] et M. Ross Quillian [Quillan, 1963]. Ce dernier définit un réseau sémantique comme "un format de représentation permettant de mémoriser le sens des mots, pour rendre possible leur utilisation à la manière de l'être humain" [Quillan, 1963]. Plusieurs modèles existent dont celui des *graphes conceptuels* [Sowa, 1983]. Un graphe conceptuel est un graphe bipartite orienté. Il contient deux types de nœuds : concepts et relations. L'auteur [Sowa, 1983] a montré la passerelle qu'il existait entre la logique des prédicats du premier ordre et les graphes conceptuels par l'existence d'un fonction bijective permettant d'associer une formule logique à tout graphe donné. Dans ce cadre nous pouvons citer ConceptNet [Liu et Singh, 2004] qui est une base de connaissances lexicales visant à capturer de l'information de *sens commun*. Ce dernier fait référence à l'information sémantique qui permet aux humains de comprendre les événements triviaux du quotidien. ConceptNet, contrairement à WordNet [Miller, 1995], est optimisé pour réaliser des inférences pratiques et basées sur le contexte. Dans un article sur les réseaux sé-

mantiques [Brachman, 1983], l'auteur souligne les problèmes qui surgissent lors de la définition des nœuds (concepts, relations, etc) et des liens (difficulté d'interprétation).

Nous pouvons distinguer deux grands types de réseaux :

— Les réseaux structurés comme des ontologies ayant une structure pyramidale (exemple : WordNet)

— Les réseaux à modélisation relationnelle. Le modèle adopté par ce type de réseaux est celui d'un grand graphe de relations sémantiques, syntaxiques, et lexicales reliant les nœuds entre eux.

2.3.2.1 Dans le domaine général

, Nous allons présenter dans un premier temps le réseau WordNet puis nous détaillerons le réseau lexico-sémantique JeuxDeMots(JDM).

WordNet

WordNet [Miller *et al.*, 1990] [Fellbaum, 2005][8] est le réseau lexical (disponible pour l'anglais) qui fait référence dans le domaine du TAL. Il a été initialement conçu dans le cadre d'un projet dont le but était de tester les déficits lexicaux dans des expériences de psychologie cognitive. Il contient environ 155 000 termes et 180 000 relations (table 2.2). Les principales relations sont taxonomiques (hyperonymie, hyponymie, holonymie et méronymie) (figure 2.2). L'hyperonymie est une relation sémantique hiérarchique dans laquelle une relation d'inclusion établie entre un terme général et un ou plusieurs termes spécifiques (par exemple la *grippe* est une *maladie infectieuse*). La relation d'hyponymie est l'inverse la relation d'hyperonymie. Un holonyme A d'un terme B est un terme dont le signifié désigne un ensemble comprenant le signifié de B (membre inférieur est un **holonyme** de fémur). La relation inverse de cette relation est la méronymie (portière est un **méronyme** de voiture). Citons d'autre relations comme la relation implication (entailment) qui est spécifique aux verbes : un verbe X nécessite (entails) Y si X ne peut être réalisé à moins que Y ne le soit (*dormir* est impliqué par *ronfler*). Toujours pour les verbes, si le verbe A précise l'action décrite par le verbe B, on parle de relation de troponymie (*décapiter*

8. https ://wordnet.princeton.edu/

et *égorger* sont des manières de *tuer*).

Catégorie	Mots	Concept	Total Paires Mot-Sens
nom	117798	82115	146312
verbe	11529	13767	25047
adjectif	21479	18156	30002
adverbe	4481	3621	5580
total	155287	117659	206941

Table 2.2 – Nombre de mots et de concepts dans WordNet

L'organisation des entrées est réalisée autour de synsets (figure 2.2), des regroupements de lexèmes supposés synonymes (ou quasi-synonymes). Des définitions sont également présentes.

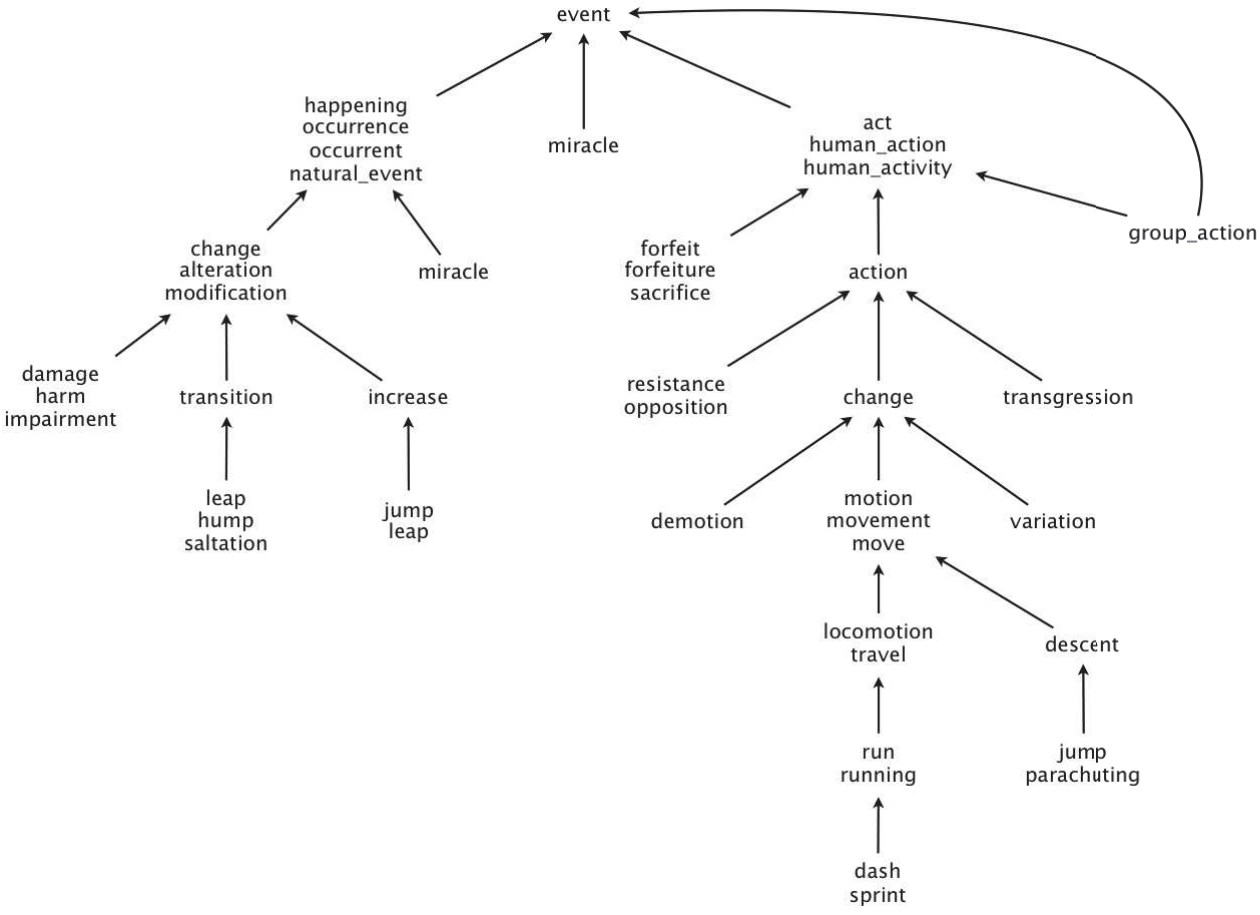


Figure 2.2 – Relations dans WordNet(version 2006) d'après université de Princeton [9]

WordNet est décliné en plusieurs langues (arabe, chinois, serbe, etc). Pour le français, WOLF [Sagot et Fišer, 2008], qui est un Wordnet Libre du français, a été construit à partir de Princeton WordNet et d'autres ressources multilingues. Les lexèmes polysémiques ont été traités au moyen d'une approche reposant sur l'alignement en mots d'un corpus parallèle en cinq langues.

Dans un réseau comme WordNet, la construction se fait de façon manuelle et demande donc du temps. D'autres systèmes sont construits de façon automatique à partir de corpus de textes [Lapata et Keller, 2005].

Hownet

HowNet [Dong et Dong, 2006], [Tang *et al.*, 2005] a été lancé en Chine pour créer un réseau sémantique reliant des entrées lexicales et conceptuelles des lexiques chinois avec leurs équivalents en anglais. Ce réseau, construit manuellement, s'appuie sur une base de connaissances conceptuelle. Hownet exploite deux idées principales :

— La composition : les concepts peuvent être composés pour former d'autres concepts.

— L'évolution : les entités ont des propriétés qui peuvent ne pas être partagées par spécialisation mais exprimées en utilisant les relations inter-attributives. La base de données HowNet est un base de sémèmes construite à la main. Un sémème est une unité de sens atomique. L'hypothèse de ce réseau est qu'un ensemble fermé de sémèmes doit être suffisant pour décrire un ensemble ouvert de concepts. Sachant qu'il est basé sur le système des alphabets chinois, la question se pose de savoir s'il est possible d'étendre HowNet vers d'autres langues.

BabelNet

[Navigli et Ponzetto, 2010] a introduit un réseau lexico-sémantique multilingue contenant plus de 6 millions de concepts appelé BabelNet. Cette ressource a été construite par la fusion automatique entre l'encyclopédie multilingue Wikipedia et le réseau populaire WordNet. Le comblement des lacunes lexicales pour les langues faibles en ressource est réalisé par traduction automatique des concepts et extraction automatique de sens et de relations de sens inter-langues. Ce réseau contient plus de 13 millions de Babel synsets sous forme de 6 millions de concepts et 7 millions

d'entités nommées, lexicalisées en différentes langues et liées par plus de 262 millions de relations lexico-sémantiques multilingues.

Ce réseau est structuré de façon similaire à WordNet : chaque *Babel Synset* représente un certain sens et contient tous les synonymes qui expriment ce sens en différentes langues (figure 2.3). Une méthode totalement automatique peut conduire à certaines erreurs.

Le réseau fournit, par exemple, de la connaissance lexicale pour un concept avec sa partie de discours, sa définition et les synonymes dans différentes langues, ainsi que de la connaissance encyclopédique. Lors de la construction, les auteurs ont pris de WordNet tous les sens de mots et toutes les liaisons sémantiques et lexicales entre les synsets (en guise de relations). Ils ont pris de Wikipedia toutes les entrées encyclopédiques et des relations non typées à partir des liens hypertextes.

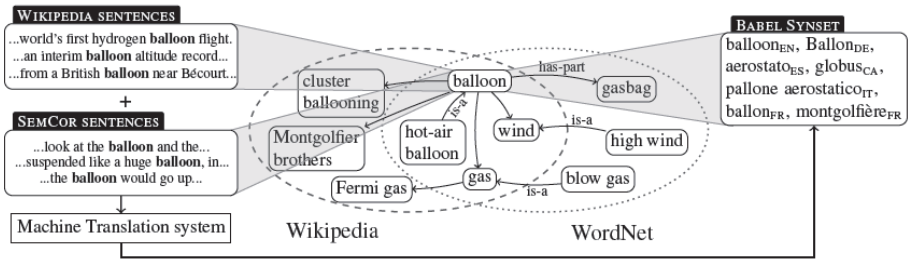


Figure 2.3 – Vue globale de Babel Net pour le terme *balloon* ([Navigli et Ponzetto, 2010]

FrameNet

FrameNet [Baker *et al.*, 1998] est un projet lancé à l'Institut International de Berkeley, et qui a produit une ressource lexicale se basant sur l'anglais. Cette ressource est fondée sur la théorie des cadres sémantiques (*frame semantics* [Fillmore, 1982]). Fillmore postule qu'on ne peut comprendre le sens de bien des mots, de façon optimale, qu'en tenant compte du contexte dans lequel ils s'inscrivent. Chaque cadre conceptuel est caractérisé par des rôles thématiques, des propriétés, etc, nommés éléments cadres. L'objectif principal de FrameNet est d'encoder de la connaissance sémantique sous une forme exploitable par les machines, ce qui a été effectué essentiellement en utilisant une méthode manuelle soutenue par des outils du TALN. Les cadres sont bien connus dans le domaine du TALN. Un cadre se définit comme un

ensemble d'attributs, de valeurs associées et de contraintes [Rich et Knight, 1991]. Le projet permet de décrire la diversité des possibilités combinatoires de chaque acceptation d'un mot, à la fois sur le plan syntaxique et sémantique grâce à un système semi-automatique d'exemples de phrases qui permet l'affichage des résultats de l'annotation [Ruppenhofer *et al.*, 2012].

JeuxDeMots

Lancé en septembre 2007, JeuxDeMots (JDM) [Lafourcade, 2007] est un projet ayant pour objectif de construire un grand réseau lexico-sémantique du français à l'aide de plusieurs GWAP (*game with a purpose*) [Thaler *et al.*, 2011].

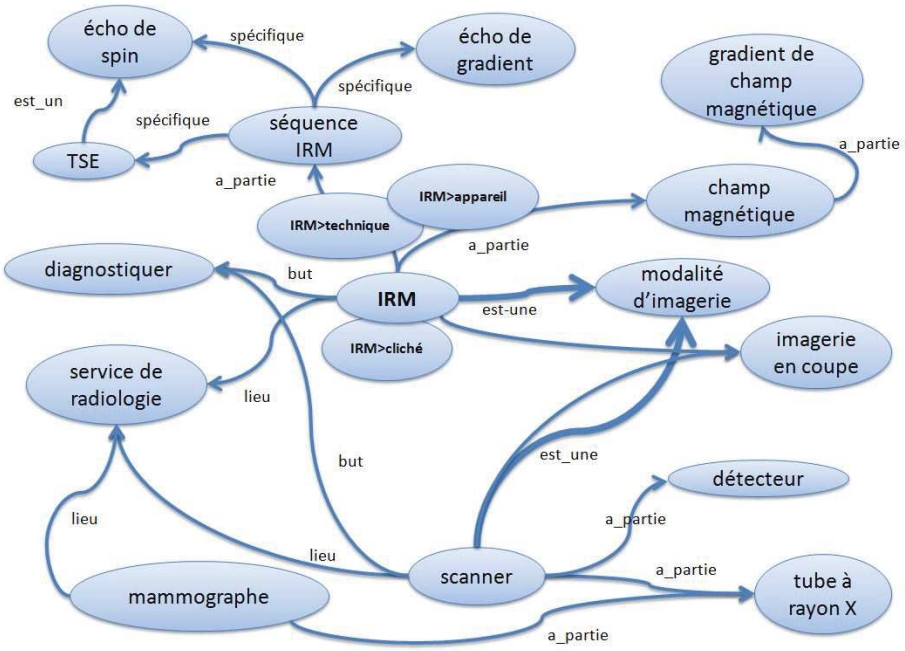


FIGURE 2.4 – Exemple de sous réseau lexical pour le terme IRM dans JeuxDeMots.

Le réseau JDM a une structure de graphe où les mots représentent des objets lexicaux et les arcs des relations entre ces objets. Il existe, dans le réseau, une centaine de relations lexico-sémantiques binaires pouvant relier deux termes. Ces relations sont pondérées et orientées, leur poids rendant compte de la force avec laquelle la relation considérée unit les deux termes. Le réseau est alimenté en partie par des GWAP, dont le jeu principal, appelé JeuxDeMots, repose sur l'idée suivante :

Au début d'une partie, une consigne concernant le type de relation (*idée associée,*

synonyme, caractéristique, etc) ainsi qu'un terme cible issu du réseau lexical (par exemple, donner des *idées associées* au terme *maladie*) sont présentés au joueur. Ce dernier dispose d'un temps limité pour saisir les termes qui lui semblent correspondre à la consigne. Par la suite, ce même couple terme/consigne est proposé à d'autres joueurs. Les réponses communes par paire de joueurs sont insérées dans le réseau lexical ou le renforce, c'est-à-dire au cas où la relation entre les deux termes existe déjà, son poids est augmenté (force d'association) [Lafourcade et Joubert, 2009]. Ce processus de validation est similaire à celui utilisé par [Von Ahn et Dabbish, 2008] pour l'indexation des images et par [Lieberman *et al.*, 2007] pour la collecte de la connaissance du *sens commun* et [Siorpaes et Hepp, 2008] pour l'extraction de connaissance. Une occurrence de relation peut éventuellement avoir un poids négatif ce qui indique que la relation est fausse bien que pertinente (par exemple : une autruche, bien qu'étant un oiseau, ne vole pas). Le calcul du score [Lafourcade et Joubert, 2008] est conçu pour augmenter à la fois la précision et le rappel lors de la construction du réseau. Les réponses présentées par les deux joueurs sont affichées, celles en commun sont mises en évidence, ainsi que leur score.

D'autres jeux, dans le cadre de JDM, existent en particulier ceux avec choix comme *AskIt* ou *LikeIt*.

Certaines relations s'avèrent compliquées à alimenter via ce mode de jeu, soit parce qu'elles requièrent une certaine expertise dans le domaine considéré (quand, par exemple, le terme n'appartient pas au vocabulaire général), ou lorsque la relation elle-même suscite très peu de réponses (par exemple, *donner le féminin de...*).

Pour répondre à ce problème, un outil contributif Diko[10] permet d'alimenter le réseau de façon négociée. Cette partie sera développée au chapitre 3.

Au 19 septembre 2016, le réseau contient **948 544** termes et **52 889 304** relations entre ces derniers. Les termes polysémiques sont raffinés en leurs différents usages.

10. http ://www.jeuxdemots.org/diko.php

2.3.2.2 Dans le domaine médical

Les ontologies générales du domaine médical, l'UMLS et SNOMED-CT, tentent de couvrir l'ensemble de la médecine. Les ontologies du domaine médical ne sont cependant pas toutes générales et nous considérerons aussi les ontologies spécialisées (en particulier RadLex pour le domaine de l'imagerie médicale).

UMLS

UMLS (Unified Medical Language System) [11] est développé à la bibliothèque nationale de médecine (United States National Library of Medicine) aux USA, structure dont le but principal est de faciliter la création de systèmes informatiques qui analysent des ressources médicales. UMLS combine dans une seule plate-forme un nombre important de terminologies. Il est formé de trois sources de connaissances :

— un métathésaurus, qui est une grande base de données de vocabulaire construite à partir d'un peu plus de 150 terminologies biomédicales (figure 2.5) (thésaurus, listes de termes, etc.) en une vingtaine de langues dont le français.

— un réseau sémantique, qui propose une catégorisation des concepts (figure 2.6 présents dans le métathésaurus d'UMLS et un ensemble de relations entre ces concepts. La version actuelle contient à peu près 135 types sémantiques et 54 relations possibles entre ces concepts. Les types sont définis par des descriptions textuelles, ainsi que par l'information héritée au sein de la hiérarchie. Le réseau peut être représenté par un graphe où les types sémantiques sont représentés par des nœuds et les relations par des arcs.

— un lexique (Specialist Lexicon) en langue anglaise contenant les termes n'apparaissant pas dans le métathésaurus. Pour chaque terme, il donne des informations syntaxiques, morphologiques et orthographiques. Chacune de ces entrées possède un lemme, une catégorie syntaxique, un identifiant et parfois des variantes orthographiques. Il est utilisé pour des tâches de traitement automatique de la langue anglaise.

11. http ://umlsks.mlm.nih.gov

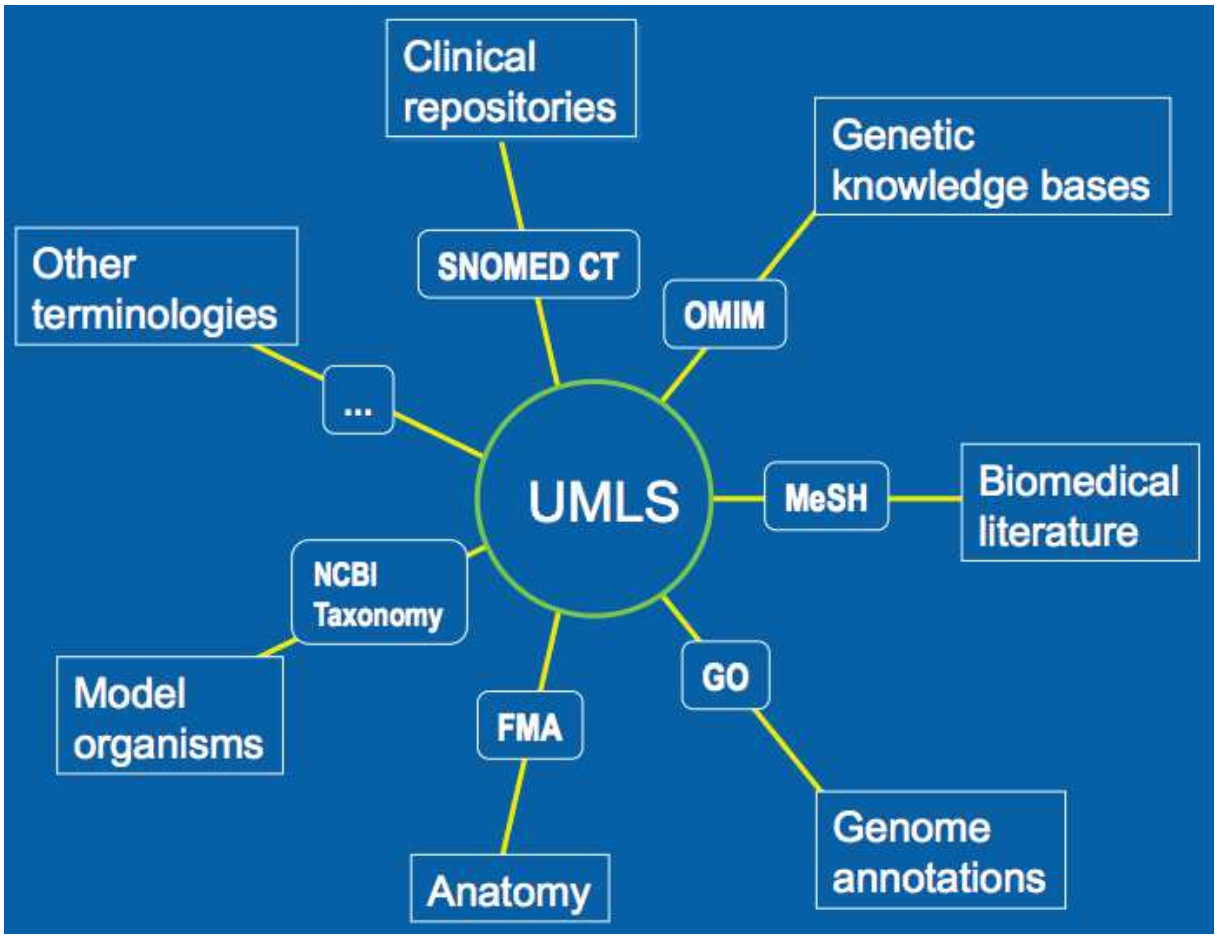


Figure 2.5 – Représentation de UMLS

Les travaux de [Bodenreider *et al.*, 2002] étudient l'utilité d'UMLS en explorant la couverture de ses termes et de ses relations par rapport aux besoins des applications biomédicales. Son utilisation couvre différents domaines telles que le traitement du langage, l'indexation [Aronson, 2001], étiquetage sémantique de textes [Ruch *et al.*, 1999], enrichissement de thésaurus [Le Duff *et al.*, 2000] ou de connaissances linguistiques [Zweigenbaum et Grabar, 2000]. La figure **??** représente un extrait du métathésaurus UMLS.

Une autre étude [Burgun et Bodenreider, 2001] analyse la compatibilité du réseau sémantique UMLS avec d'autres ontologies du domaine général comme WordNet. Les auteurs ont relevé deux difficultés importantes : les classes de même nom n'ont pas la même sémantique dans les différentes ontologies et deux classes peuvent avoir la même signification par intention, alors que leur extension dans les différentes

ontologies est différente.

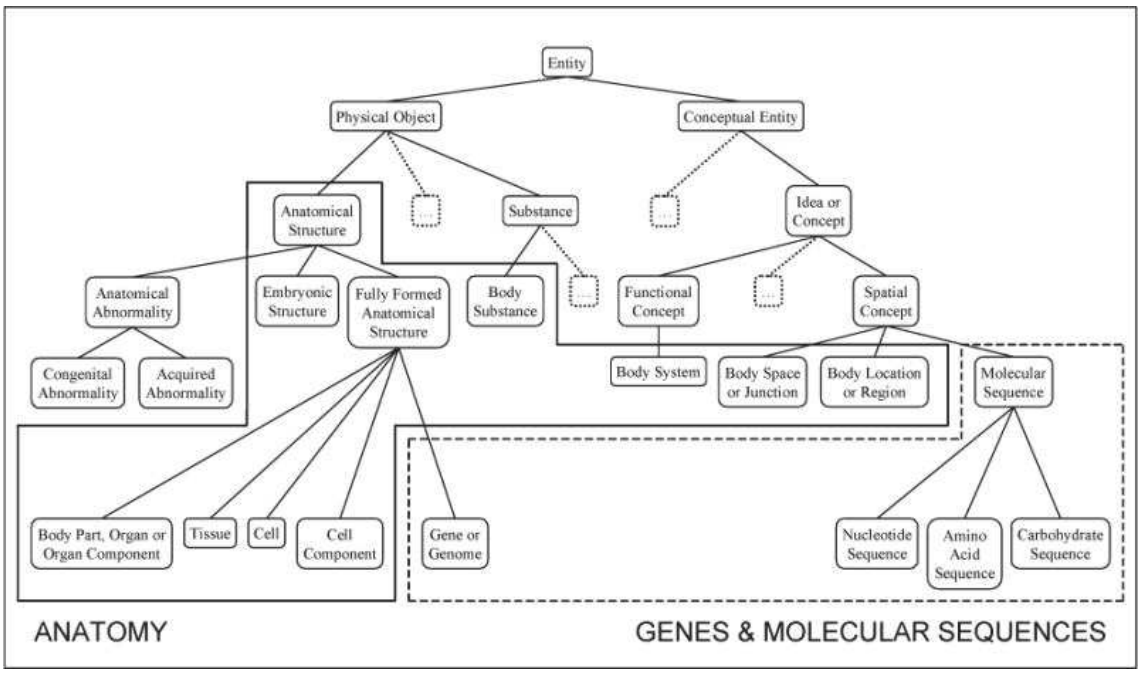


FIGURE 2.6 – Le réseau sémantique UMLS
regroupé en groupe sémantique d'après [McCray *et al.*, 2001]

Au delà de l'anglais, d'autres projets de lexiques médicaux ont été élaborés pour d'autres langues : par exemple le projet allemand [Weske-Heck *et al.*, 2002] ou encore celui en langue française [Zweigenbaum *et al.*, 2003].

En ce qui concerne le métathésaurus d'autres langues sont présentes avec une couverture variable suivant les langues.

SNOMED

SNOMED [12](Systematized Nomenclature of Medicine) est une terminologie médicale qui couvre les principaux champs de l'information clinique. La fusion de SNOMED-RT avec une terminologie britannique (clinical terms) a donné naissance à la SNOMED Clinical Terms (SNOMED CT) [Donnelly, 2006], une terminologie dynamique des soins de santé plus accessible pour les différentes spécialités médicales (figure 2.7). C'est une nomenclature pluri-axiale couvrant la médecine, la den-

12. http ://www.snomed.org

tisterie ainsi que la médecine vétérinaire. Il s'agit d'une structure hiérarchique de concepts désignés par des descriptions (termes) sur plus de 31 niveaux de subsomption. Actuellement SNOMED CT fait partie intégrante de l'UMLS. La traduction de la SNOMED en français est effectuée par l'équipe du Centre de Recherche en diagnostic médical informatisé (CRDMI).

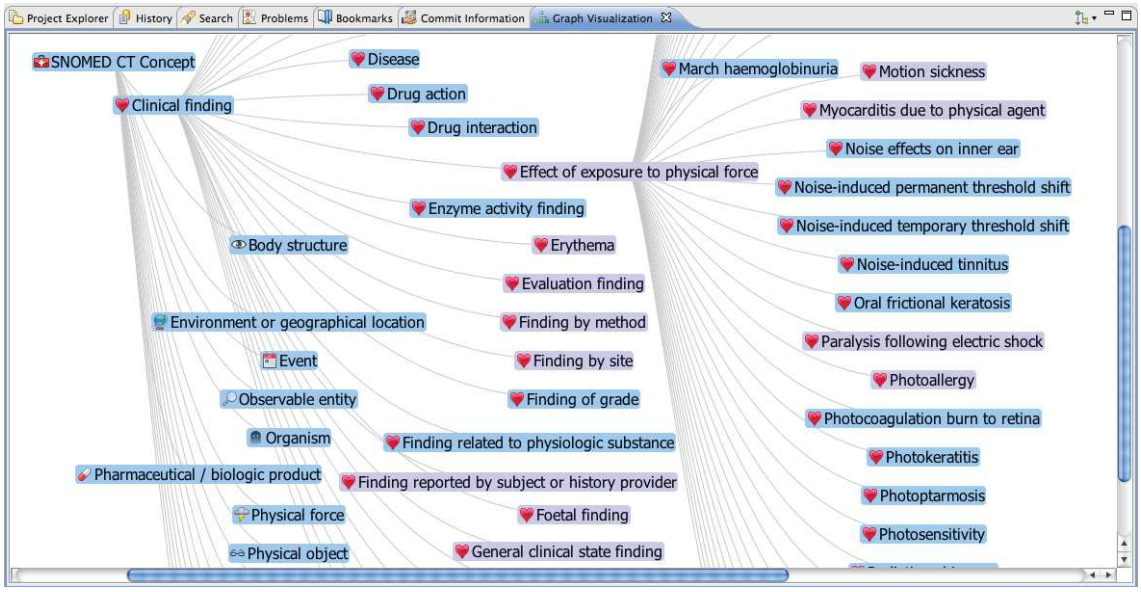


Figure 2.7 – Représentation d'une partie de SNOMED

La recherche en informatique médicale a montré que la SNOMED est une des terminologie les plus adaptées à l'indexation des informations du dossier patient [Pereira *et al.*, 2009].

Classification statistique internationale des maladies : CIM

Elle a été développée dans le but d'indexer les causes de mortalité et de morbidité dans un but statistique et épidémiologique. Elle est utilisée dans l'évaluation de l'activité hospitalière par exemple. C'est une classification divisée en 21 chapitres permettant de classer les différentes pathologies selon leur siège anatomique. Cette terminologie n'utilise pas des expressions de la langue naturelle mais plutôt un métalangage [Zweigenbaum, 1999]. Par exemple, l'emploi de sigles tel que SAI(Sans Autre Indiction) ou NCA (Non Classé Ailleurs) sont spécifiques d'un métalangage. Elle comporte environ 18 000 codes et environ 50 000 termes. Ci-dessous, nous présentons un exemple du CIM dans sa 10ème version.

Groupes :

C50_ Sein

(blocs)C60-C63 Organes génitaux de l'homme

C64-C68 Voies urinaires

Catégories

C64_ Tumeur maligne du rein, à l'exception du bassinet

C65_ Tumeur maligne du bassinet

C66_ Tumeur maligne de l'uretère

C67_ Tumeur maligne de la vessie

Sous-Catégories

C67.0 Trigone de la vessie

C67.1 Dôme de la vessie

C67.2 Paroi latérale de la vessie

Par ailleurs, des extensions de codes à usage national ont été créés pour mieux prendre en compte certains types de prise en charge .

Medical Subject Headings(MeSH) :

Le MeSH (Medical subject Heading) [13] est un thésaurus dans le domaine biomédical réalisé par le NLM (U.S. National Library of Medicine). La structure du MeSH est à trois niveaux : un ensemble de termes dont un préférentiel désigne un concept

13. http ://mesh.inserm.fr/mesh/

qui fait partie d'une classe de concepts appelée *descripteur*. Trois types de relations, ne reposant pas sur la logique formelle sont utilisées : hiérarchique, synonymique et de proximité sémantique. Nous détaillons ces élément et leur rôles.

— **Terme** : Un terme est un mot ou un ensemble de mots exprimant une notion particulière.

— **Concept** : Un concept comprend un ou plusieurs termes synonymes et porte le nom d'un de ces termes, désigné sous l'appellation *terme préféré*

— **Relation** : Il existe dans MeSH des relations de types hiérarchiques (hyperonymie, méronymie, hyponymie)
Il existe aussi une relation d'association exprimée par l'expression *voir aussi* qui relie deux concepts proches d'une façon libre et non parfaitement définie.

— **Descripteur** : Un descripteur est formé d'un ou plusieurs concepts de significations proches. Certains mots clefs sont des *descripteurs obligatoires* et doivent être obligatoirement utilisés au moment de l'indexation du document si les concepts auxquels ils renvoient apparaissent dans le document. Les autres concepts, appelés subordonnés, possèdent une relation sémantique de type hiérarchique ou associative.

— **Qualificatif** : Les qualificatifs sont des termes qui peuvent être associés à des mots clés afin d'en préciser le sens. Par exemple, *<la paire diabète de type II/thérapeutique>* désigne le traitement du diabète de type II. Toutes les associations mots clés/ qualificatifs ne sont pas autorisées.

Le MeSH est utilisé pour l'indexation par de nombreuses bibliothèques et institutions à travers le monde. L'indexation au sein de MeSH est effectuée manuellement par des indexeurs professionnels.

Foundational Model of Anatomy (FMA)

FMA [14](*Foundational Model of Anatomy*) est une ontologie de référence dans le domaine de l'anatomie humaine [Rosse *et al.*, 2003] (figure 2.8). Elle fut conçue initialement pour compléter la partie anatomie de l'UMLS. L'ontologie est maintenant devenue une référence [Rosse *et al.*, 2003]. Plusieurs travaux traitent de la représentation de l'ontologie FMA en une représentation en logique de description avec OWL (*Web Ontology Language*) [Golbreich *et al.*, 2005], [Dameron *et al.*, 2005]. Certains

14. http ://sig.biostr.washington.edu/projects/fm/

auteurs présentent l'ontologie FMA comme étant enracinée dans une ontologie de niveau supérieur [Rosse et Mejino Jr, 2008]. Elle contient 85 000 classes, 140 relations reliant les classes, et plus de 120 000 termes. La majorité des entités sont des structures anatomiques formées de plusieurs parties interconnectées de manière complexe, en termes de localisation, constituants, innervation, vaisseaux sanguins, limites, etc.

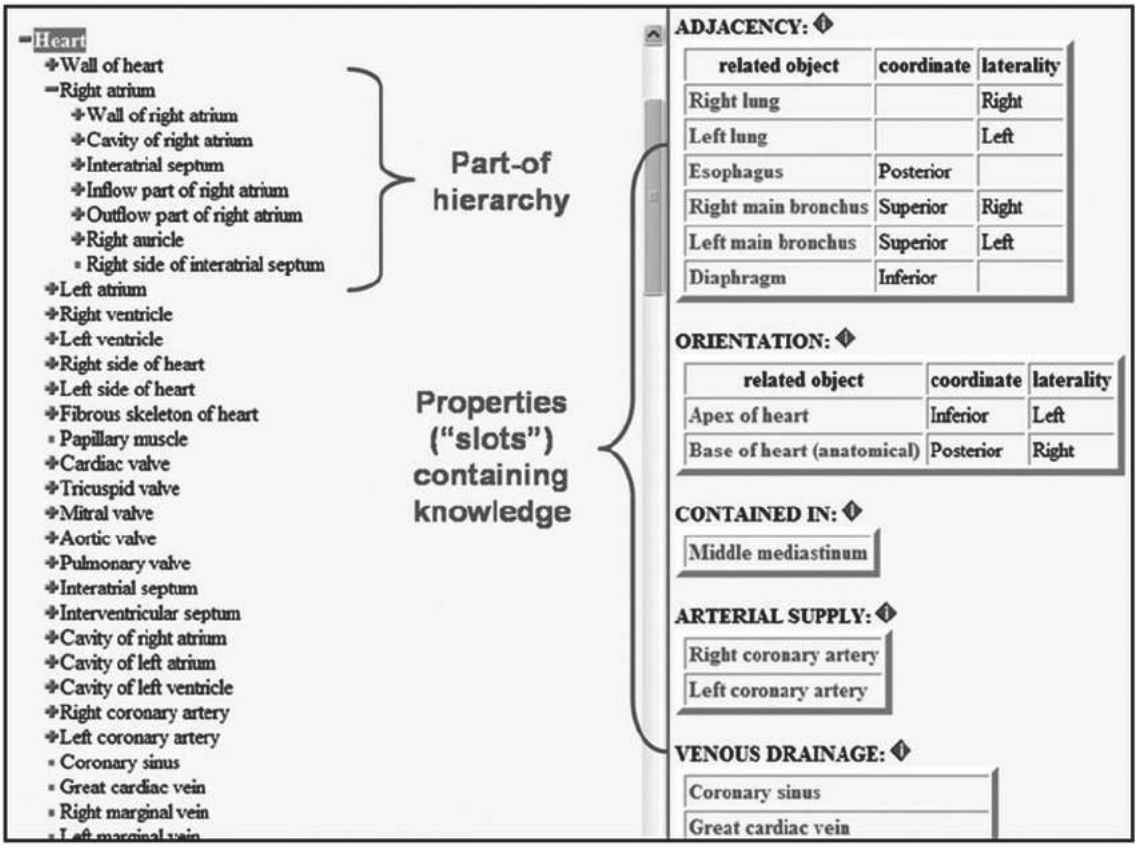


FIGURE 2.8 – Exemple de la représentation dans FMA

RadLex

Dans le domaine de la radiologie, la Radiological Society of North America (RSNA) a développé une ontologie dédiée à l'imagerie médicale : RadLex [Langlotz, 2006] (figure 2.9). Son objectif est d'offrir une structure hiérarchique unique pour indexer et retrouver des ressources en radiologie.

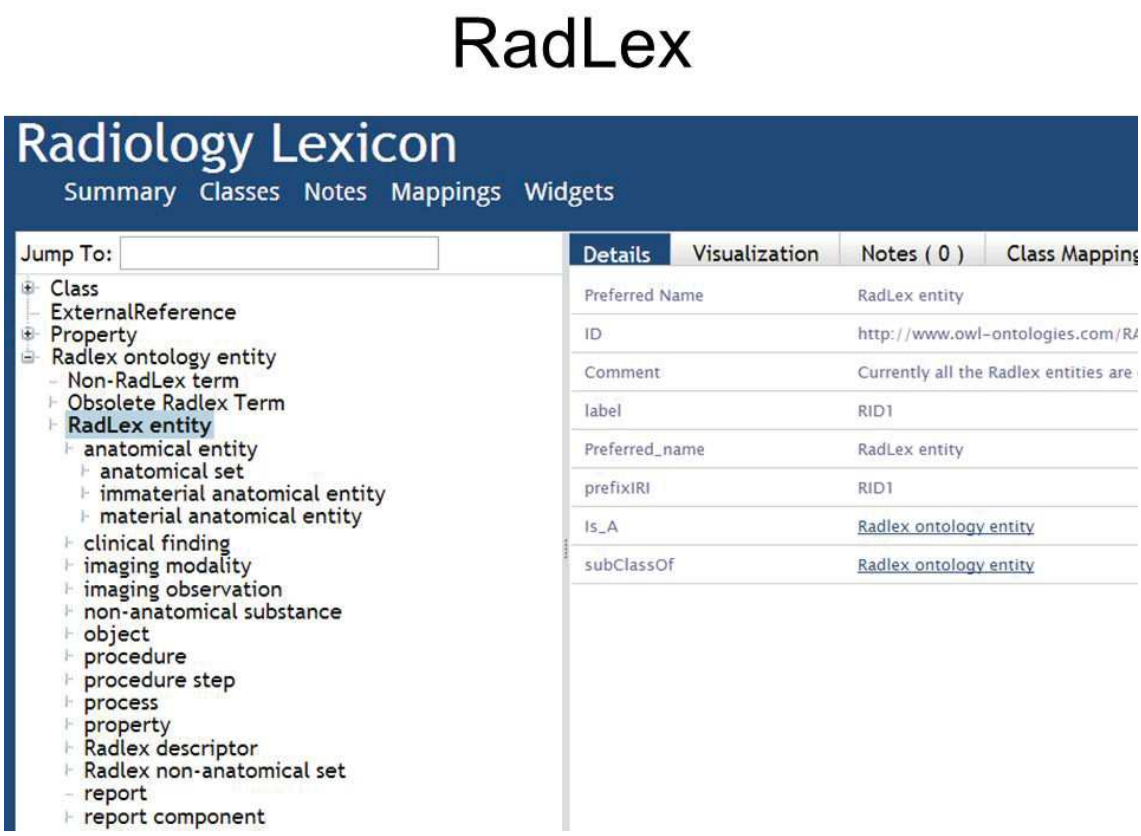


Figure 2.9 – Représentation de Radlex

RadLex contient 35 000 termes et partage certains sous-domaines plus transversaux (anatomie, signes cliniques...) avec d'autres terminologies de santé telles que la SNOMED CT et la FMA. RadLex [15] remplace l'ACR Index for radiological Diagnoses, une classification à deux axes permettant l'indexation des images et des comptes rendus. Les catégories sont générales (identifiants, patient), spécifiques à l'imagerie médicales (acquisition, qualité d'image, niveau de certitude), spécifique de la spécialité radiologiques (localisation anatomique, anatomie radiologique, signes, recommandations). A notre connaissance, c'est la seule terminologie où les signes radiologiques sont présents (par exemple : *le signe du mont Fuji*) [Shore *et al.*, 2012].

Gamuts

L'ontologie radiologique Gamuts (*Radiology Gamuts Ontology RGO*) est une base

15. http ://www.rsna.org/radlex/

de connaissances de diagnostic différentiel en radiologie qui inclut 1674 diagnostics différentiels, 19 017 termes et 52 976 liens entre ces termes [Budovec *et al.*, 2014]. Les seules relations présentes dans cette ontologie sont des relations de subsumption (is_a) et son inverse (*has subtype*) ainsi que des relations de causes et conséquences (figure 2.10).

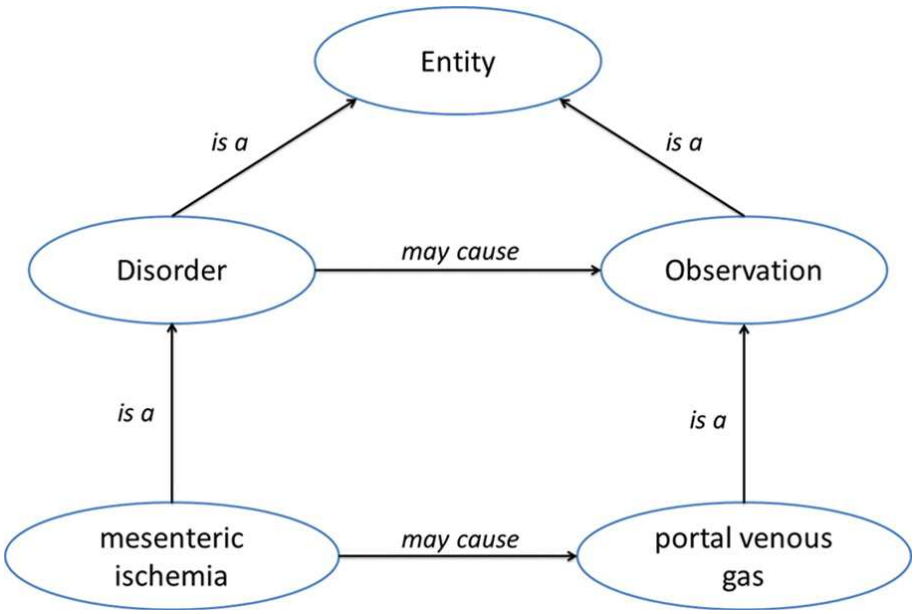


FIGURE 2.10 – Représentation des connaissances dans Gamuts
Schéma d'après [Budovec *et al.*, 2014]

2.4 Extraction de relations et le TALN dans le domaine médical

L'extraction de relations sémantiques à partir de textes peut avoir deux buts principaux. Le premier est d'aider à indexer des documents en vue de réaliser un système de recherche d'information. Le deuxième est d'améliorer qualitativement les bases de connaissances [Lee *et al.*, 2004] en les enrichissant en nouvelles relations extraites de textes.

2.4.1 Extraction de relations

La plupart des travaux concernant l'extraction de relations sémantiques se concentrent sur des relations indépendantes du domaine [Chklovski et Pantel, 2004], [Snow *et al.*, 2006]. Dans le domaine général, l'extraction entre entités utilise des approches statistiques [Nazar *et al.*, 2012] et des techniques d'apprentissage automatique [Wang *et al.*, 2006] ainsi que des approches basées sur l'utilisation de patrons ou des règles d'extraction [Hearst, 1992] voire des approches combinant ces deux techniques [Suchanek *et al.*, 2006]. L'apprentissage automatique consiste à définir un ensemble d'attributs, puis à entraîner un classifieur sur un corpus d'entraînement et enfin à utiliser le modèle appris grâce à ce corpus pour extraire les relations dans un corpus de test. L'approche linguistique [Hearst, 1992] est fondée sur l'utilisation de patrons. Un patron linguistique est un modèle qui identifie une forme d'expression de la relation à extraire. L'extraction, appelée *pattern matching* en anglais, consiste à rechercher des correspondances entre les patrons précédemment définis et les phrases du corpus. Dans l'approche combinant les deux techniques, il est possible de prendre les patrons comme attribut d'apprentissage ou bien de sélectionner l'intersection des résultats de chaque approche.

Au vu des difficultés rencontrées, selon le type de patron sélectionné, pour déterminer le type de relations entre deux termes, à cause de l'ambiguïté des patrons lexicaux, [Girju *et al.*, 2003] ont proposé d'ajouter des contraintes sémantiques pour la détection de relations de méronymie. A l'aide d'un algorithme d'apprentissage automatique, ils ont trouvé 20 contraintes. La précision obtenue est de 83%. D'autres contraintes lexicales et syntaxiques ont été appliquées sur des relations exprimées par des verbes [Fader *et al.*, 2011]. Leur système, appelé **ReVerb**, obtient une précision supérieure à 80% pour plus de 30% de relations extraites grâce à ces contraintes. Dans notre corpus, il est difficile d'exploiter les verbes, étant donnée leur absence relativement fréquente des comptes rendus radiologiques. Dans le domaine biomédical, nous pouvons distinguer plusieurs techniques.Une méthode basée sur le principe des cooccurrences a été proposée par [Stapley et Benoit, 2000] pour détecter des relations entre gènes en se basant sur des mesures statistiques de coocurrences entre mots. D'autres approches utilisent des patrons lexicaux ou des règles [Auger et Barrière, 2008], [Song *et al.*, 2015]. Cimino et Barnett [Bases, 1993] ont utilisé des patrons pour extraire des relations à partir des

titres d'article de Medline[16]. Ils ont exploités les descripteurs MeSH associés à ces articles dans Medline et la cooccurrence de termes cibles dans un même titre pour générer des règles d'extraction de relations sémantiques. Khoo et al. Certains auteurs [Embarek et Ferret, 2008] ont proposé un système fondé sur des patrons construits automatiquement en vue de l'extraction de quatre relations entre cinq types d'entités médicales (*maladie-traitement, maladie-médicament, maladie-examen, maladie-symptômes*). La construction des patrons se base sur une méthode automatique en utilisant une distance d'édition entre deux phrases et un algorithme d'alignement de parties de phrases en prenant en compte les différentes informations sur les mots. Dans les approches à base de règles nous pouvons aussi citer le système SemRep [Rindflesch *et al.*, 2000] qui permet de trouver les relations sémantiques dans des textes biomédicaux. De même, [Abacha et Zweigenbaum, 2011] et [Lee *et al.*, 2004] se sont intéressés à l'extraction de relations sémantiques reliant deux types d'entités médicales à savoir une maladie et un traitement en utilisant une méthode basée sur des patrons lexicaux. Lee et al [Lee *et al.*, 2004] ont appliqué leur méthode sur des résumés médicaux pour l'identification de relations de type traitement entre *médicament* et *maladie*. Ils obtiennent un rappel de 84.8% et une précision de 48.1%. Abacha et al [Abacha et Zweigenbaum, 2011] ont utilisé une méthode semi-automatique pour la génération de patrons alors que Lee et al ont utilisé une méthode manuelle. Les principales relations extraites sont les relations *traitement, remède, détection, signe*. Parallèlement, d'autres travaux se basent sur des approches d'apprentissage supervisé [Rink *et al.*, 2011]. Roberts et al [Roberts *et al.*, 2008] ont extrait des relations sémantiques dans les textes médicaux (*relation de lieu,d'indication,etc*) par des méthodes d'apprentissage supervisé en utilisant des classifieurs SVM. Plusieurs challenges [Uzuner *et al.*, 2011] ont proposé entre autre une tâche d'extraction de relations sémantiques. Xiao et al [Xiao *et al.*, 2005] ont proposé une technique pour l'extraction des interactions entre protéines avec une méthode à base d'apprentissage supervisé.

Plusieurs méthodes sont proposées pour l'extraction de relations sémantiques. Certaines approches privilégient le rappel alors que d'autres se concentrent sur la précision. Les méthodes linguistiques peuvent échouer à extraire certaines relations à cause de la grande variabilité des patrons linguistiques ainsi qu'à le généricité de certains patrons. Les approches par apprentissage supervisé obtiennent une grande

16. https ://www.medline.com/home.jsp

précision seulement si un grand nombre d'exemples annotés est disponible.

2.4.2 Traitement automatique du langage naturel dans le domaine médical

Jusqu'à présent, les applications nécessitant un traitement de la langue sont :

— indexation de textes médicaux divers (articles, comptes rendus ....) ;
— classification ;
— recherche d'information (RI) ;

Pour ces applications, il s'agit la plupart du temps de mettre en correspondance les termes d'un texte libre avec ceux d'un vocabulaire contrôlé (CIM-10, MeSH). Il est également souhaitable que différents systèmes automatiques puissent être inter-opérables. Par exemple, il est, parfois, intéressant d'effectuer une recherche d'information contextuelle (une requête en termes MeSH) à partir d'un dossier patient (codé à l'aide des termes du CIM10).Dans certains cas, cela permet une re-formulation d'une requête utilisateur en langue naturelle dans un but de désambiguïsation.

Plusieurs travaux se sont attachés à traiter certaines problématiques du traitement naturel du langage médical. Pour la langue française, plusieurs projets ont tenté de fournir des ressources contrôlées aux acteurs de l'informatique pour la santé. Le projet VUMEF [Darmoni *et al.*, 2003] a pour but d'augmenter la part du français dans l'UMLS et de gérer la traduction de certaines terminologies comme le MesH, par exemple. Ce projet permet d'effectuer un aide au codage des affections dans le cadre du dossier patient. Le projet UMLF [Zweigenbaum *et al.*, 2003] propose de constituer une ressource pour le médical français. Il propose d'intégrer les variantes flexionnelles et dérivationnelles des termes médicaux. Ces informations sont encodées dans un format informatique standard afin de favoriser leur intégration dans des systèmes de traitement automatique du langage médical.

Les problèmes liés aux variations lexicales (casse, ordre des mots, orthographe, accentuation (*epanchement intrathoracique/épanchement intra-thoracique*), etc) ont reçu une certaine attention [Grabar, 2004] de même que ceux liés aux variations mor-

phologiques [Namer, 2005] (*maladie cardiaque/maladie du cœur*). D'autres travaux se sont intéressés à l'insertion et suppression d'éléments dans les termes [Jacquemin, 1997] ainsi qu'au traitement des périphrases [Namer, 2005] (*myocardite/ inflammation du cœur*). Le traitement de la métonymie a aussi été par aborder par [Bouaud *et al.*, 1996] (*le fond de l'œil confirme le diagnostic / l'examen du fond d'œil confirme le diagnostic*) . D'autres approches ont exploité les spécificités de la langue médicale [Claveau et Zweigenbaum, 2005], [Namer, 2005]. Sager et al [Sager *et al.*, 1987] présentent également des travaux centrés sur la langue médicale.

Conclusion

Dans ce chapitre, nous avons présenté un état de l'art sur l'indexation et sur l'extraction de relations sémantiques. Nous avons présenté les ontologies nécessaires souvent à une tâche d'indexation conceptuelle. Nous avons décrit cette tâche à la fois dans le domaine général et dans le domaine biomédical. En général, jusqu'à présent, les deux domaines sont séparé surtout du point de vue des bases de connaissances. De plus dans le réseau JDM seront présent les différentes variations morphologique, des périphrase, etc dans le but de résoudre les difficultés liés au langage médical quotidien. Nous allons dans la suite du manuscrit présenter notre méthode d'indexation dans le domaine médical adossée à un réseau lexico-sémantique général contenant des connaissances spécialisées.

PARTIE 2 : Notre approche de l'indexation de comptes rendus de radiologie

Dans cette deuxième partie, nous abordons l'indexation des comptes rendus radiologiques en vue de l'implémentation d'un moteur de recherche. L'approche que nous proposons consiste à réaliser une indexation (à la fois de termes et de relations) qui s'appuiera sur un base de connaissance de sens commun dans laquelle nous incluons des connaissances de spécialité (hypothèse de non séparation). Dans ces travaux, nous ne nous contentons pas seulement d'extraire des concepts ou des relations entre termes, mais nous essayons aussi d'expliciter les informations implicites notamment par l'utilisation d'annotations présentes dans un réseau lexico-sémantique ainsi que par une augmentation d'index à l'aide d'un algorithme de propagation à travers ce réseaux.

Le chapitre 3 présentera le crowdsourcing, le réseau lexico-sémantique (JeuxDeMots) ainsi que l'implémentation des annotations de relations.

Le chapitre 4 présente notre méthode de constitution d'une base de connaissance spécialisée à l'intérieur d'une base de sens commun. Nous détaillerons notre méthode d'indexation (index brut et index augmenté) des documents du corpus.

Le chapitre 5 présente une méthode d'extraction de relations sémantiques ainsi qu'un système d'extraction des variables ciblées *Patients, Modalité et Affection*, le modèle *PMA*.

Chapitre 3

Le réseau lexico-sémantique JDM et le domaine radiologique

Sommaire

3.1 Crowsourcing et Game With A Purpose 81
3.1.1 Crowdsourcing 81
3.1.2 Un outil contributif : Diko 83
3.2 Annotation de relations 85
3.2.1 Déduction 85
3.2.2 Principe des annotations de relation 87
3.2.3 Expérimentations sur la propagation des annotations . . . 95
3.2.4 Exploitation des annotations 98

Dans le domaine de la radiologie où il est intéressant d'extraire des termes pertinents des comptes rendus, les relations pertinentes entre les termes sont cruciales. La taxonomie indique seulement la hiérarchie entre les termes (relation d'hyperonymie *is-a*). Il peut être utile pour le médecin de disposer également plus facilement de relations non hiérarchiques (*caractéristique, cause, conséquence, ...*). Il est pertinent de donner pour une maladie, la liste des symptômes (*douleur, fièvre, céphalée, etc*), des cibles potentielles (*enfants, adultes, adulte jeune, diabétique, alcoolique ...*), des localisations anatomiques et cela indépendamment de toute hiérarchie. Ceci peut être modélisé par un réseau sémantique. L'utilisation d'un réseau sémantique de sens commun dans lequel est présente une base de spécialité peut jouer un rôle important dans l'analyse de comptes rendus radiologiques. En effet, dans la section *Indication* du rapport de radiologie, le texte est souvent rédigé avec des termes courants (*chute dans la baignoire, encornage, lame d'un couteau, bouteille de bière, boule de pétanque, etc*) alors que la section *Résultats* comporte davantage de termes spécialisés (*glioblastome multiforme, ectasier, hypersignal FLAIR, prise de contraste lobulée, aspect en nid d'abeille*, etc). Dans l'optique de la création de ce réseau spécialisé dans le domaine de la radiologie en langue française, nous avons décidé d'utiliser comme réseau sémantique JDM [Lafourcade, 2007] comme base pour le réseau de sens commun. Nous ajoutons ainsi des informations de spécialités (radiologie) à ce réseau sémantique.

Dans ce chapitre, nous consacrons une grande partie aux annotations de relations

à l'intérieur du réseau lexico-sémantique JDM. L'organisation de ce chapitre est la suivante : nous commençons par rappeler brièvement quelques notions sur le crowdsourcing et les Games With A Purpose (*GWAP*). Dans la section suivante, après un bref rappel sur l'outil contributif Diko, nous nous focaliserons sur le principe des annotations (mécanisme endogène au réseau) de relations et l'intérêt à ajouter ces informations. L'évaluation de l'approche sera enfin présentée et discutée. Les travaux sur l'annotation de relations présentés dans ce chapitre s'appuient sur les articles [Ramadier *et al.*, 2014a], [Ramadier *et al.*, 2014b].

3.1 Crowsourcing et Game With A Purpose

3.1.1 Crowdsourcing

La construction d'un réseau lexical collaboratif peut être envisagée selon deux stratégies. Premièrement, comme un système contributif de type Wikipedia où des volontaires complètent les entrées. Dans un second cas, les contributions sont faites indirectement par l'entremise de jeux, connus sous le nom de GWAP (*Game with a purpose*) [Von Ahn et Dabbish, 2008].

Il existe un grand nombre de GWAP, couvrant une grande variété de domaines comme la biologie avec Foldit [Eiben *et al.*, 2012] (repliement des protéines), la médecine (*Reverse The Odds*[1]), ou encore des connaissances générales comme par exemple le jeu OnToGalaxy [Krause *et al.*, 2010] ou encore le projet Wordrobe[2] [Venhuizen *et al.*, 2013]. JeuxDeMots (voir paragraphe 2.3.2) est un GWAP permettant la construction d'un réseau lexico-sémantique de grande taille pour la langue française.

La myriadisation (*crowdsourcing*) [Sagot *et al.*, 2011] consiste à externaliser une activité, à la transférer vers la foule (*crowd*), qui représente un grand nombre d'acteurs potentiels (cf.définitions). Son essor est fortement lié au développement des nouvelles technologies de l'information et de la communication et, plus particuliè-

1. http ://www.cancerresearchuk.org/support-us/citizen-science-apps-and-games-from-cancer-research-uk/reverse-the-odds
2. http ://wordrobe.housing.rug.nl/Wordrobe/public/HomePage.aspx

rement, du Web 2.0 qui facilite la mise en relation d'un grand nombre d'acteurs dispersés. La participation de ces internautes peut être bénévole ou rétribuée, suivant les tâches et les systèmes. Nous pouvons citer l'encyclopédie Wikipedia comme système de contribution bénévole alors que RentACoder [3] fait partie de la catégorie des systèmes où la collaboration est rémunérée. Il existe aussi des systèmes pour lesquels il n'est pas demandé aux internautes d'avoir une connaissance particulière, mais d'effectuer des tâches élémentaires telles qu'attribuer une étiquette sur une image. La tâche demandée à l'internaute peut être découpable en micro-tâches élémentaires. Souvent, dans ce type de système il y a rétribution soit sous forme de points gagnés dans le cadre d'un GWAP (comme JDM) soit sous forme monétaire (comme dans le cas d'Amazon Mechanical Turk (*AMT*) [4]) [Fort *et al.*, 2011]. Cependant, certains chercheurs [Bhardwaj *et al.*, 2010], [Gillick et Liu, 2010] ont observé que, quand la complexité de la tâche augmente, alors la qualité des données produites avec le système AMT est insuffisante. Par ailleurs, le paiement à la tâche entraîne comme conséquence de privilégier le nombre de tâches réalisées au détriment de la qualité de la réalisation des données produites. Quand elle est monétaire, la rétribution soulève divers problèmes législatifs et éthiques concernant le travail [Fort *et al.*, 2011]. Il est intéressant de noter que, toujours dans le cas d'AMT, les relations entre les participants, les donneurs d'ordre et Amazon sont très douteuses en regard du droit français. En effet, en France, le travail à la tâche est illégal.

Définitions : **Myriadisation (crowdsourcing)** :
Utilisation du savoir-faire d'un grand nombre de personnes pour réaliser certaines tâches

Certains auteurs [Mortensen, 2013] considèrent que le *crowdsourcing* peut être un moyen de pallier les difficultés que pose le développement d'ontologies larges et complexes. Ayant utilisé la myriadisation pour tester la qualité de l'ontologie (et non pour la construction de la ressource), ils montrent que leur méthode pour cette tâche de vérification donne une précision de 82%. Le crowdsourcing existe aussi dans les domaines de spécialité comme le droit [Getman et Karasiuk, 2014] ou la bio-informatique [Good et Su, 2013]. Dans ce dernier domaine, bien que les

3. http ://www.rent-acoder.com/
4. http ://www.mturk.com

participants ne soient pas des spécialistes, les résultats obtenus sont tout à fait corrects [Mortensen *et al.*, 2015]. Dans leur expérience, ils rapportent un accord inter-annotateurs (coefficient Kappa-Cohen) entre l'expert et la *foule* de 0.58 alors que celui entre experts est de 0.59. Ces résultats suggèrent, que pour la tâche demandée, à savoir la détection d'erreurs dans l'ontologie SNOMED-CT, il n'existe pas de différence entre experts et la *foule*. L'utilisation de la foule est donc utile même dans des tâches concernant des domaines de spécialités.

En plus de la construction d'ontologies, la myriadisation a été utilisé dans l'alignement d'ontologie [Sarasua *et al.*, 2012], et dans leur enrichissement [Sajous *et al.*, 2013].

Ces différents travaux montrent le potentiel du recours au crowdsourcing dans l'ingénierie des ontologies [Lafourcade *et al.*, 2015].

3.1.2 Un outil contributif : Diko

En complément des jeux présents dans le réseaux JDM, il a été développé un outil contributif *Diko* car certaines relations sont difficiles à renseigner sous forme ludique. Le système de contribution peut être utilisé pour développer un domaine de spécialité (radiologie, botanique, zoologie, etc) à l'intérieur du réseau. Une proposition faite par un joueur dans Diko sera soumise aux votes des autres joueurs, alors que via le jeu JDM, elle est validée automatiquement, pour peu qu'elle soit faite simultanément par une paire de joueurs, sur le même couple terme/relation. Lorsqu'un certain nombre de votes donnés est atteint pour une proposition, un expert validateur est averti et finit par inclure (ou éventuellement exclure) la relation proposée dans le réseau. L'expert peut rejeter totalement une proposition de relation ou l'inclure dans le réseau avec un poids négatif si cela est pertinent. Par exemple, il est pertinent de savoir qu'une *autruche* ne peut pas *voler* bien que ce soit un *oiseau* :

autruche r_agent-1 voler

Nous présentons ci-dessous des exemples d'entrées de l'interface Diko (figure 3.1, figure 3.2) :



FIGURE 3.1 – Capture écran de la fenêtre de Diko du terme *cirrhose*.
Nous voyons les différentes relations du terme *cirrhose* présentes dans le réseau.

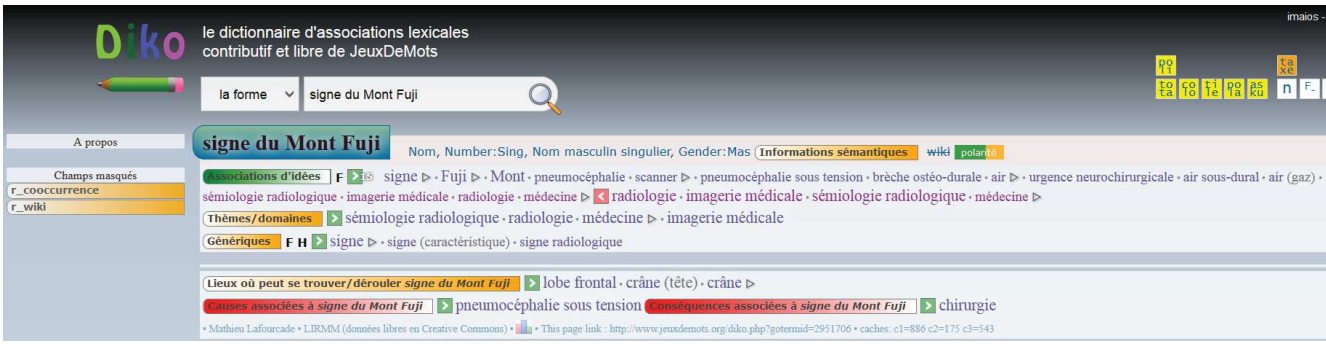


FIGURE 3.2 – Capture écran de la fenêtre de Diko du terme *signe du Mont Fuji*.
Terme utilisé spécifiquement pour la radiologie.

Concernant le domaine de spécialité (médecine) pouvant nous intéresser pour le projet IMAIOS, le tableau (table 3.1) donne une idée de l'ordre de grandeur relatif à la quantité d'informations dont nous disposons. Ces nombres de liens sont en perpétuelle évolution.

terme	nombre de liens sortants	nombre de liens entrants
médecine	34 734	37 190
anatomie	24 252	25 978
radiologie	581	729
accident	965	1280
sémiologie	96	86
maladie	15 016	14 802
IRM	354	801

TABLE 3.1 – Nombres de liens entre termes dans JeuxDeMots pour certains termes clés du domaine médical.

3.2 Annotation de relations

En général, et surtout dans le domaine des connaissances spécialisées, la corrélation entre la force d'association de la relation (poids de la relation) et son importance conceptuelle n'est pas toujours assurée. Par exemple, pour le terme *carcinome hépatocellulaire*, la relation *caractéristique* avec le terme *wash-out* est très spécifique à la radiologie, par conséquent le poids de la relation sera faible dans le cadre général de la médecine alors que pour le radiologue cette relation est particulièrement importante. Un autre cas est la relation *diagnostic* entre la *sclérose en plaques* et l'*IRM*. Nous avons affaire à une relation, là encore, spécifique au domaine de l'imagerie médicale, qui sera particulièrement pertinente pour le radiologue mais avec une faible force d'association. Pour remédier à ce problème, nous introduisons la notion d'annotation de relations. Ces annotations sont, dans un premier temps, renseignées de manière manuelle, puis dans un second temps nous appliquons un mécanisme d'inférence afin de les propager. Avant de détailler le principe des annotations, nous décrivons brièvement un schéma d'inférence [Zarrouk *et al.*, 2013] à savoir la déduction qui nous servira pour la propagation des annotations à travers le réseau JDM.

3.2.1 Déduction

Le schéma déductif est basé sur la transitivité de la relation ontologique *is-a* (hyperonyme). Si un terme A est un type de B et B a une relation R avec le terme C, alors on peut proposer que A entretienne la même relation avec C. Le schéma

peut être représenté comme suit :

$$\exists A \xrightarrow{is-a} B \wedge \exists B \xrightarrow{R} C \Rightarrow A \xrightarrow{R} C$$

Définitions : **Transitivité des relations** :
Une relation de sous-classe est transitive lorsque B est une sous-classe de A et C est une sous-classes de B, alors C est une sous-classe de A

Le moteur d'inférence est appliqué sur les termes ayant au minimum un hyperonyme. Si un terme T possède un ensemble d'hyperonymes pondérés, le moteur d'inférence déduit un ensemble d'inférences. Ces hyperonymes vont être classés selon un ordre hiérarchique. Le poids d'une inférence proposée est la moyenne géométrique incrémentale de chaque occurrence (c'est-à-dire, que la présence d'un poids négatif suffit à rendre la moyenne invalide). Le schéma présenté ci-dessus est très simple ; en effet, si le terme A est polysémique, l'inférence proposée sera fausse. Nous utilisons alors un blocage logique (figure 3.3).

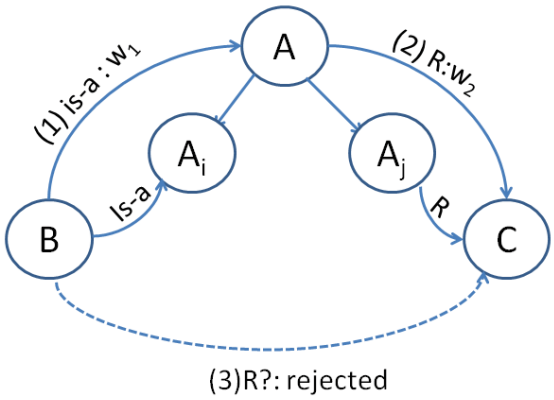


FIGURE 3.3 – Schéma d'inférence déductive triangulaire avec un blocage logique dû à la polysémie du terme A du milieu. Les termes A_i etA_j sont des raffinements/usage de A d'après [Zarrouk *et al.*, 2013] .

Nous évaluons le niveau de confiance de chaque inférence produite de façon à filtrer les inférences douteuses par l'intermédiaire d'un filtrage statistique. Le poids w d'une relation inférée est la moyenne géométrique (et non arithmétique) des poids des prémisses. Si la seconde prémisse a un poids de valeur négative, alors le poids de l'inférence n'est pas un nombre et cette dernière est rejetée.

Comme la moyenne géométrique est plus sensible aux petites valeurs que la moyenne arithmétique, les inférences qui ne sont pas basées sur des prémisses fortement valables ne passeront pas ce filtre.

En cas d'invalidation, un agent réconciliateur est invoqué pour essayer de comprendre pourquoi la relation a été considérée fausse : erreur dans les prémisses, polysémie (inférence faite en partant d'un terme central polysémique). Pour la propagation des annotations, c'est ce type d'inférence que nous allons considérer. Pour augmenter la précision des résultats et pour empêcher d'inférer certaines relations peu pertinentes mais vraies (*homme a pour partie protons*), nous avons bloqué les inférences sur les relations qui avaient été annotées comme *non pertinent* ou *exceptionnel*.

3.2.2 Principe des annotations de relation

Lors d'une recherche de documents, si plusieurs comptes rendus correspondent exactement à la requête, il pourrait être intéressant de les classer du plus fréquent au plus rare en ce qui concerne les caractéristiques de la pathologie. C'est pourquoi, il est apparu utile d'introduire la notion d'annotation pour certaines relations. Dans le réseau lexical, une relation est notée par un triplet : <nœud$_{source}$, type de relation/annotation, nœud$_{cible}$>.

3.2.2.1 Le principe des annotations

Dans le champ de la radiologie, les relations les plus utiles pour le radiologue (données établies suivant leur pratique quotidienne) sont indiquées dans le tableau 3.2. Parmi les relations pertinentes pour le domaine qui nous intéresse, où trois ont été ajoutées (*symptômes, diagnostic et cibles*), toutes les autres sont des relations ap-

partenant au domaine général. Dans les ontologies existantes dédiées à la radiologie comme RadLex, il n'existe pas autant de types de relations potentiellement utiles pour l'analyse des comptes rendus. Dans la recherche d'informations radiologiques (comptes rendus et images), les annotations de relations peuvent apporter un complément d'information et permettre de classer les réponses par ordre de pertinence. Par exemple, cela peut aider les radiologues devant une image anormale pour savoir si une caractéristique est rare ou fréquente et ainsi leur apporter une aide au diagnostic.

D'autres types d'annotations peuvent exister comme par exemple la pertinence ou non d'une relation entre deux termes. Ce type d'information est généralement absent d'un réseau sémantique ou d'une ontologie. Par exemple la relation *cible* entre *rougeole* et *adulte* est annotée comme *rare* et cette information sera directement disponible dans le réseau (figure 3.4).

Un autre exemple, pouvant être utile aux radiologues, est la relation *lieu* entre *muscle supra-épineux* et *trochiter* qui sera annotée comme *terminaison*. En ce qui concerne les vaisseaux sanguins, une annotation de type informative a été ajoutée (vascularisation) sur la relation lieu. Par exemple :

— *artère cérébrale moyenne* r_lieu *aire de Broca (vascularisation)*
— *artère cérébrale moyenne* r_lieu *ganglions de la base (vascularisation)*

Ces informations peuvent permettre une analyse plus précise des comptes rendus en permettant d'avoir des informations implicites.

Par exemple dans la phrase :

Aspect IRM compatible avec un infarctus dans le territoire sylvien gauche au stade subaigu.

le système pourra expliciter les zones atteintes (aire de Broca, aire de Wernicke, ganglions de la base, etc) grâce aux annotations.

is-a	Hyperonymes du terme. Exemple : *IRM* est une *modalité d'imagerie* (possible)
synonyme	Synonyme du terme. Exemple *hepatocarcinome* synonyme *carcinome hépatocellulaire* (préféré)
synonyme strict	Synonyme exact du terme. Exemple *patella* synonyme strict *rotule* (langage courant)
lieu-1	A partir d'un lieu, termes qu'on peut trouver dans ce lieu. Exemple dans le lieu *cuisse>jambe* on peut trouver *artère fémorale (vascularisation)*
partie-de	Parties, constituants, éléments du mot cible. Exemple : *foie* a comme partie *segment I* (toujours vrai)
caractéristique	Caractéristiques (adjectifs) possibles,typiques. Exemple : *carcinome hépatocellulaire* carac *hypervasculaire* (fréquent)
localisation	Lieux typiques où peut se trouver le terme/objet en question. Exemple : *sclérose en plaque* loc *système nerveux central*
cible	Population affecté par le terme. Exemple : *rougeole* cible *enfant* (fréquent)
diagnostic	Examen. Exemple : *sclérose en plaque* diag *IRM* (fréquent, crucial)
symptôme	Symptômes d'une maladie, *rougeole* symptôme *fièvre* (fréquent)
cause	B est une cause de A. Exemple : *cirrhose* cause *alcoolisme*
conséquence	B est une conséquence possible de A. Exemple : *accident vasculaire cérébral* peut avoir comme conséquence une *hémiplégie*
prédecesseur(spatial)	Qu'est ce qui peut précéder spatialement le terme A. Exemple *artère cérébrale antérieure* nait de (est précéder par) l'*artère carotide interne*
successeur (spatial)	Qu'est ce qui peut suivre spatialement le terme A. Exemple l'*aorte* se termine (est suivi par) par les *artères iliaques communes*
tributaire	affluent ou collatérale. Exemple l'*artérère cérébrale antérieure* a comme collatérale (tributaire) *artère communicante antérieure*
dérivé morphologique	Des dérivés morphologiques sont demandés. Exemple *artère* a comme dérivé morphologique *artérite*

TABLE 3.2 – Relations pertinentes en radiologie pour l'analyse de compte rendu

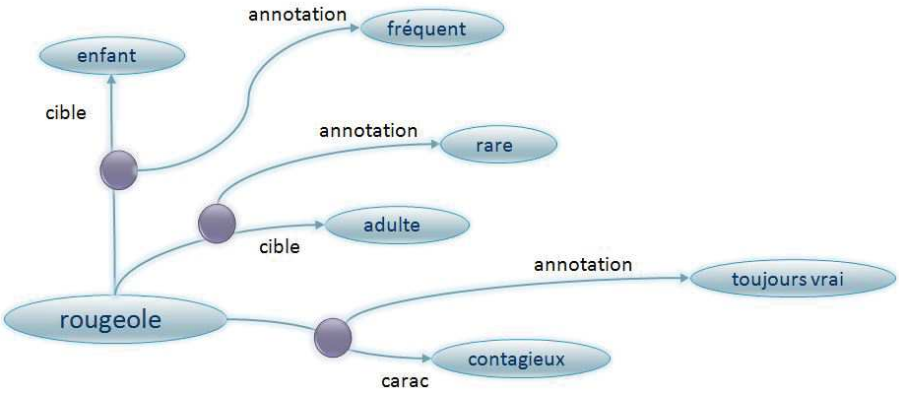


FIGURE 3.4 – Exemple d'implémentation d'annotation.
L'implémentation d'une annotation se fait par réification de la relation à annoter dans le réseau lexical. Le nœud relation ainsi créé peut être associé à d'autres termes. La relation annotation n'est qu'un type de relation parmi d'autres. Les valeurs d'annotation sont des termes standard.

3.2.2.2 Les différents types d'annotation

Les types d'annotations peuvent être de natures différentes (information fréquentielle d'usage ou de pertinence). Ci-dessous, nous présentons les principaux types d'annotations :

- annotations fréquentielles : *très rare*, *rare*, *possible*, *fréquent*, *toujours vrai* ;
- annotations d'usage : *souvent crû vrai*, *abus de langage*, *langage courant*, *terme préféré*, *élision* ;
- annotations quantitatives : *un nombre (1, 2, 4, ...)*, *beaucoup*, *peu*, etc. ;
- annotations d'exception : *exception* ;
- annotations qualitatives : *pertinent*, *non pertinent*, *inférable*.
- annotations informatives : *vascularisation*, *innervation*, *origine*, *terminaison* ;

L'annotation d'usage concerne la manière dont les médecins radiologues s'expriment dans leur comptes rendus. Un médecin peut utiliser le terme grippe au lieu de virus de la grippe ou bien hépatite C au lieu de virus de l'hépatite C : c'est un **abus de langage**, le praticien fait simplement un raccourci de langage sans pour autant faire de confusion dans son esprit. Il semble évident pour lui que ces deux expressions sont différentes. L'annotation **souvent crû vrai** s'applique pour une fausse relation (avec un poids négatif) qui est souvent considérée comme vraie, par exemple *araignée*(*is-*

a/souvent crû vrai) *insecte*. L'annotation **élision** permet de détecter les termes que les médecins écrivent sous forme de raccourcis. Par exemple, dans les comptes rendus il sera noté *rupture de la coiffe* au lieu de *rupture de la coiffe des rotateurs* qui est l'appellation complète présente dans les ontologies. L'annotation *origine (ou terminaison)*, elle, concerne surtout l'anatomie (par exemple pour les muscles, les nerfs) dans le cadre de cette thèse. Les exceptions sont également renseignées et prennent la forme d'une relation ayant un poids négatif.

Nous avons aussi ajouté des annotations de type qualitatif. L'annotation de nature **qualitative** est liée au statut inférable de la relation, particulièrement concernant l'inférence. L'annotation **pertinent** se rapporte à un niveau ontologique adéquat pour une relation donnée. Par exemple, *être vivant* (*carac/pertinent*) *vivant* ou *être vivant* (*carac/pertinent*) *mort*. L'annotation **inférable** est supposée être ajoutée quand une relation est inférable (ou a été inférée) à partir d'une relation existante, par exemple : *chien* (*carac/inférable*) *vivant* car *chien* (is-a) *être vivant*. L'annotation **non pertinent** est ajoutée aux relations vraies mais qui sont très éloignées du niveau pertinent, par exemple animal *possède/non pertinent atomes*. Le caractère pertinent ou non pertinent d'une annotation peut varier pour une même relation suivant le domaine de spécialité. Par exemple, foie>anatomie (*possède/ non pertinent*) *cellules* pour la connaissance générale mais cette information sera pertinente dans le domaine de l'anatomie-pathologie. La notion de pertinence peut également varier d'une spécialité médicale à l'autre.

L'annotation **quantitative** représente le nombre de parties d'un objet. Un être humain a généralement deux poumons alors nous annotons *être humain* (*has-part/2*) *poumon*. Ce type d'annotation n'est pas nécessairement numérique mais peut également utiliser des adverbes de quantité comme *beaucoup*, *peu*, etc.

3.2.2.3 Tri des génériques

Pour avoir l'annotation la plus précise possible, nous avons besoin d'ordonner les termes centraux du plus spécifique au moins spécifique. Nous entendons par termes centraux les termes génériques c'est-à-dire ceux liés par une relation d'hyperonymie. Pour les termes *sclérose en plaques*, *imagerie par résonance magnétique* et *artère cérébrale antérieure*, la hiérarchie sera :

sclérose en plaques
< maladie du système nerveux central < neuropathie < maladie dégénérative < maladie

imagerie par résonance magnétique
< modalité d'imagerie médicale < examen radiologique < imagerie < examen médical < examen

artère cérébrale antérieure
< artère cérébrale < artère < vaisseau sanguin < vaisseau < structure vasculaire < structure anatomique

Pour choisir la bonne annotation de la nouvelle relation inférée, la hiérarchie ontologique joue un rôle important. L'annotation du terme le plus spécifique doit avoir plus d'influence que celle du terme le moins spécifique. Nous prenons en compte ce fait pour les mécanismes d'inférences avec annotations. Nous expliquons l'algorithme qui permet d'obtenir ce résultat.

Au cours de la première étape, nous extrayons une liste de génériques non ordonnés du réseau lexical JDM. Par exemple, pour le terme *méningiome*, nous avons :

tumeur, maladie (médecine), tumeur bénigne, maladie neurologique, tumeur bénigne du cerveau, tumeur du cerveau, tumeur cérébrale, tumeur intracrânienne, tumeur bénigne du système nerveux central, tumeur bénigne de la moelle épinière, tumeur du système nerveux central, tumeur des méninges, tumeur maligne du système nerveux central, tumeur maligne, tumeur de la moelle épinière

Nous éliminons de la liste les variantes orthographiques ainsi que les synonymes stricts. Par exemple dans la liste ci-dessus *tumeur du cerveau* et *tumeur cérébrale*

sont synonymes stricts, donc un seul des deux est conservé pour la suite de l'algorithme. Nous notons cette liste filtrée FT.

La deuxième étape est d'ordonner les termes. Il faut remarquer que le résultat obtenu est un ensemble de chemins P = p_1...p_n où chaque p_i représente un chemin. Un chemin est un ensemble ordonné de termes qui appartient à FT. Nous initialisons P à l'ensemble vide. Pour chaque terme t appartenant à FT (t $\in$ FT), nous l'insérons à sa place appropriée dans chaque p de P (comme dans n'importe quel algorithme de tri par insertion qui est un tri stable). Si le terme t ne peut pas être inclus, nous ajoutons une nouvelle liste à P(P=P $\cup$ t). Pour un chemin donné p, un terme t peut être inséré entre deux termes consécutifs t_a et t_b si et seulement si (fonction test) t_a $<$ t et t $<$ t_b , où x $<$ y signifie que y est un générique de x.

Quand tous les termes t ont été ajoutés à un ou plusieurs chemins de P, l'algorithme se termine et le résultat est P. P est un ensemble de chemins (séquence ordonnée de termes) qui couvre complètement le graphe. Le processus a une complexité cubique selon le nombre de termes.

algorithme 1 triDesGeneriques

```
Entrée FT := liste des génériques filtrés d'un terme T
P : ensembles des chemins p
Initialisation : P <- ensemble vide
pour i de 1 à FT faire
  x ← [i]
  j ←i
  tantque (j > 0 AND P[j - 1] > x) faire
    P[j] ← P[j - 1]
    j ← j - 1
  fin tantque
fin pour
```

La complexité polynomiale $O(n^m)$ est largement raisonnable en pratique car le nombre de termes génériques dépasse rarement 100. Les chemins hiérarchiques peuvent être calculés instantanément. Même pour les termes monosémiques, la plupart du temps, les vues multiples liées au terme conduisent à une hiérarchie, qui prend la forme d'un graphe acyclique direct.

Nous présentons les différents chemins obtenus pour le terme *méningiome*, qui est monosémique (un seul sens),

— méningiome → tumeur bénigne du cerveau → tumeur bénigne du système nerveux central → tumeur bénigne → tumeur → maladie (médecine)
— méningiome → tumeur bénigne du cerveau → tumeur du cerveau → maladie neurologique → maladie(médecine)
— méningiome → tumeur bénigne de la moelle épinière → tumeur de la moelle épinière → tumeur du système nerveux central → tumeur → maladie (médecine)
— méningiome → tumeur des méninges → tumeur du système nerveux central → tumeur → maladie (médecine)

et pour le terme *carcinome hépatocellulaire* :

— carcinome hépatocellulaire → carcinome → tumeur maligne → tumeur → maladie(médecine)
— carcinome hépatocellulaire → tumeur maligne hépatique → tumeur maligne → tumeur → grosseur
— carcinome hépatocellulaire → tumeur maligne hépatique → tumeur hépatique → tumeur → grosseur
— carcinome hépatocellulaire → maladie du foie et des voies biliaires → maladie(médecine)
— carcinome hépatocellulaire → tumeur de l'appareil digestif → tumeur → grosseur
— carcinome hépatocellulaire→ carcinome → tumeur maligne → cancer → maladie(médecine)

Dans le mécanisme d'inférence, le terme B (le terme central, c'est-à-dire le générique) joue un rôle primordial. Nous inspectons la hiérarchie des termes B selon laquelle une relation spécifique a été inférée plusieurs fois et nous gardons la plus spécifique. Si nous obtenons deux termes ou plus ayant le même niveau sémantique, nous appliquons la règle du maximum aux valeurs correspondant à chaque annotation (toujours vrai : 5, fréquent : 4, possible : 3, rare : 2, très rare : 1) (figure 3.5).

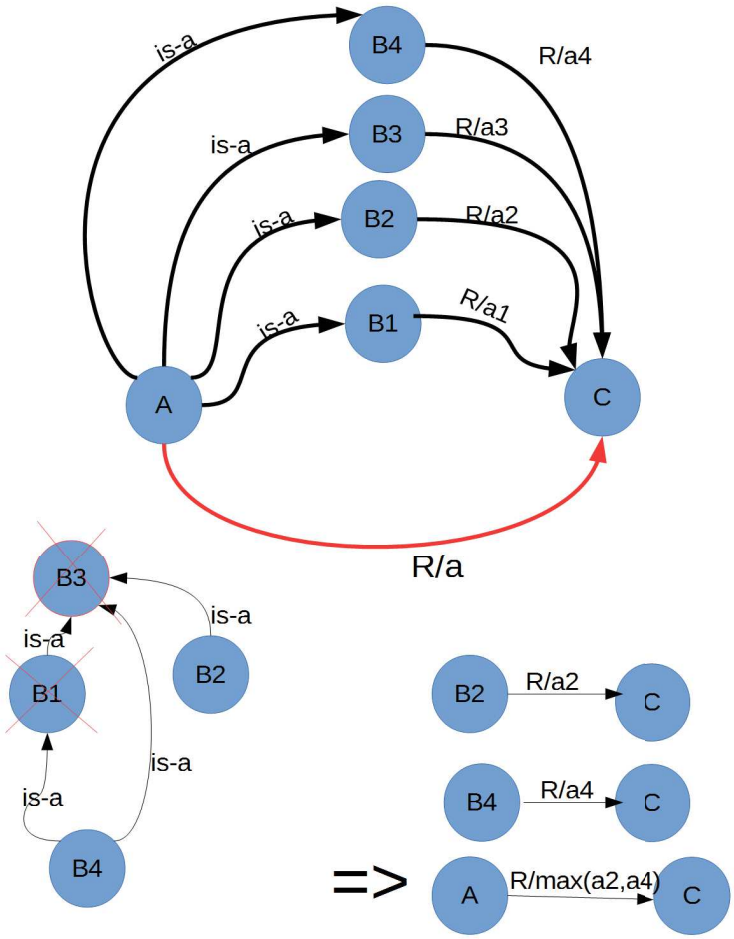


FIGURE 3.5 – Approche basée sur la hiérarchie utilisée pour choisir l'annotation la plus précise avec plusieurs termes centraux

.

3.2.3 Expérimentations sur la propagation des annotations

La propagation des annotations est réalisée par l'intermédiaire du mécanisme d'inférence. Dans le but de formuler de nouvelles conclusions (c'est-à-dire, des relations entre les termes) à partir des prémisses (des relations préexistantes), un moteur d'inférence a été proposé [Zarrouk *et al.*, 2013]. Le moteur d'inférence propose des relations, à l'image d'un contributeur, qui vont être soumises aux votes d'autres

contributeurs et validées par un expert dans le domaine de l'imagerie médicale. Dans le cadre de la thèse nous ne décrivons qu'un seul type d'inférence : le schéma déductif.

Nous avons appliqué le moteur d'inférences pour les annotations. Il permet d'ajouter des annotations aux relations de manière automatique à partir d'annotations de relations déjà existantes. Il est lancé sur la base des relations déjà enrichies avec le mécanisme d'inférence déductive. Contrairement au moteur d'inférences de relations , nous autorisons la redondance en vue d'améliorer la précision des résultats du système de propagation d'annotations de relations. Voici un exemple :

Prémisses : *accident vasculaire cérébral(is-a)infarctus cérébral*
& infarctus cérébral(diagnostic/fréquent) IRM
→ **relation inférée** : *accident vasculaire cérébral (diagnostic/fréquent) IRM* (1)

Prémisses : *accident vasculaire cérébral(is-a) maladie cérébrovasculaire*
& maladie cérébrovasculaire(diagnostic/possible) IRM
→ **relation inférée** : *accident vasculaire cérébral(diagnostic/possible) IRM* (2)

Le système d'annotations produit deux occurrences (1) et (2) de la même relation *accident vasculaire cérébral* (diagnostic) *IRM*, avec deux annotations différentes (possible, fréquent), nous décidons de garder celui avec la plus forte valeur (fréquent). Le système d'inférence des annotations appliqué sur la base de relations provenant des résultats du moteur d'inférence déductive, a annoté 10 085 relations à partir d'une amorce de 72 relations annotées (table 3.3).

Le nombre de relations annotées par type d'annotation ne dépend pas du nombre de relations existantes au départ mais simplement du nombre de relations d'hyperonymie existantes pour le terme central. Le schéma d'inférence est le suivant :

$$\mathbf{A}(\text{is-a})\mathbf{B} \text{ et } \mathbf{B}(\text{R/annot})\mathbf{C} \rightarrow \mathbf{A}(\text{R/annot})\mathbf{C}$$

Par exemple :

$$\underbrace{\begin{array}{c}\text{cancer du poumon non à petites cellules}\\ \text{carcinome hépatocellulaire}\\ \text{glioblastome}\end{array}} \tag{3.1}$$

Type d'annotation	**annotation existante**	**annotation inférée**
Fréquentiel :toujours vrai	20	8092
Fréquentiel :fréquent	18	1
Fréquentiel :possible	16	150
Fréquentiel :rare et très rare	7	35
Qualitatif :souvent crû vrai	1	7
Qualitatif :non pertinent	5	1604
Quantificateur	5	178
Total	72	10085

TABLE 3.3 – Nombre d'annotations inférées après application du système d'annotation des relations sur celles existantes.

(is-a) tumeur maligne
& tumeur maligne (carac/fréquent) mauvais pronostic

A partir des prémisses présentées ci-dessus 3 relations seront annotées comme fréquentes, et cela grâce aux multiples hyperonymes du terme *tumeur maligne*.

Plus le nombre de relations d'hyperonymie vers le terme B (*tumeur maligne*) qui a une relation annotée (*tumeur maligne(carac/fréquent)*) est important, plus le nombre de relations annotées est élevé. Supposons que le terme *carcinome hépatocellulaire* n'ait pas de relations d'hyperonymie, dans ce cas l'annotation *fréquent* ne générera pas d'autre annotation. Ceci peut expliquer la raison pour laquelle il y a peu d'annotations inférées pour le type d'annotation fréquentiel *fréquent*. Notons que l'absence de certaines relations ou certains termes est due à l'aspect de progression continue du réseau qui fait qu'il est possible qu'à un instant précis un terme ou une relation manque. Nous avons évalué le nombre d'annotations inférées, et il apparaît que 87% d'entre elles sont correctes, 5 % incorrectes et le reste 8% est discutable (les experts discuteront non pas leur validité mais plutôt leurs valeurs fréquentielles pour savoir si elles doivent être modifiées). Le système d'inférence et d'annotations tourne à présent en continu dans le but de consolider le réseau léxico-sémantique JDM.

3.2.4 Exploitation des annotations

L'exploitation des annotations dans le cadre de cette thèse rentre dans le domaine de l'analyse de texte. En effet, l'utilisation des annotations de relations présentes dans le réseau permet d'expliciter le texte c'est-à-dire d'expliciter des informations non présentes explicitement dans le compte rendu.

Pour l'ordonnancement de la pertinence des documents retournés par le système en réponse à une requête q, les annotations permettent une classification plus précise en fonction du type d'annotations utilisées. Par exemple nous utiliserons les annotations de types fréquentiels (*fréquent, rare*) pour ordonner les résultats retournés par le moteur de recherche.

Soit deux documents contenant les phrases suivantes :

D1 = *bilan d'un CHC chez une patiente post-parturiente.*

D2 = *suspicion d'un CHC chez un patient cirrhotique.*

A la requête *CHC*, les deux documents obtiendront le même score (avec la méthode BM25 par exemple). Mais grâce au réseau, nous avons l'information que la relation caractéristique entre *CHC* $\overset{r_carac}{\longrightarrow}$ *patiente post-parturiente* est annotée comme *rare* alors que la relation *CHC* $\overset{r_carac}{\longrightarrow}$ *patient cirrhotique* est annotée comme *fréquent*. Dans la réponse des documents retournée par le système *D2* apparaîtra en premier car la présence de l'annotation *fréquent* ajoute 1 (choisi de façon empirique) au score calculé du BM25.

Une autre application possible de l'utilisation des annotations de relations est la simplification de termes médicaux des documents médicaux. Dans le cadre cette tâche c'est-à-dire rendre le compte rendu radiologique intelligible pour les patients, nous remplaçons les termes médicaux difficiles (difficulté d'un terme calculée par leur tfidf) [Grabar et Hamon, 2016] possédant une relation de synonymie ou d'hyperonymie avec l'annotation *langage courant* par un terme plus familier pour les patients.

Exemple : *Trait de fracture au niveau de la patella* traduit par *Trait de fracture au*

niveau de la rotule.

Dans le réseau JDM, nous avons la relation suivante avec l'annotation *langage courant* :

$$patella \xrightarrow{r_syn} rotule\ (langage\ courant)$$

Exemples : *Ces caractéristiques évoquent fortement un hépatocarcinome* traduit par *Ces caractéristiques évoquent fortement cancer du foie.*

Dans JDM nous avons :

$$hépatocarcinome \xrightarrow{r_isa} cancer\ du\ foie\ (langage\ courant)$$

Exemple : *patiente se plaignant d'un prurit* traduit par patiente se plaignant d'une *démangeaison*

Dans JDM nous avons :

$$prurit \xrightarrow{r_syn} démangeaison\ (langage\ courant)$$

Le terme composé *carcinome hépatocellulaire* sera remplacé par un des ses hyperonyme *cancer du foie* (dans le réseau JDM nous avons *carcinome hépatocellulaire* r_isa *cancer du foie* (*langage courant*)). Dans ce cas, nous avons choisi la relation d'hyperonymie parce que les différents synonymes (*CHC, hépatocarcinome*) ne sont pas plus faciles à comprendre pour le patient. Pour évaluer notre approche, nous avons sélectionné un échantillon de notre corpus qui contenait 120 220 tokens. Pour ces derniers, nous cherchons dans le réseau JDM les synonymes et hyperonymes qui possédaient l'annotation *langage courant*. Dans le tableau (table 3.4) nous présentons quelques termes médicaux remplacés par des termes du langage courant grâce aux annotations. Nous les remplaçons ensuite dans le texte.

Terme original	**Remplacé par**	**Relation**
aphasie	mutisme	syn
cephalée	mal de tête	syn
carcinome hépatocellulaire	cancer du foie	is_a
glioblastoama	brain tumor	is_a
prurit	démangeaison	syn

Table 3.4 – Exemples de termes remplacés

le patient a une **aphasie** depuis deux jours (*phrase originale*)
Le patient a un **mutisme** depuis deux jours (*phrase simplifiée*)

Exemple de simplication. Le terme remplacé est en gras

L'annotation de relations (langage courant) n'est pas la seule façon de procéder pour simplifier un compte rendu radiologique pour le rendre compréhensible par les patients. En effet si le terme composé ne possède pas d'annotation pour la relation de synonymie ou d'hyperonymie, alors nous appliquons une autre approche. Nous extrayons l'information sémantique de JDM pour chaque mot composant le terme composé. L'information lexicale indique si un mot fait partie du langage commun ou pas. Par exemple *gliosblastome* est une *tumeur du cerveau*. Comme cette relation d'hyperonymie n'a pas d'annotation, nous extrayons des informations sémantiques pour *tumeur* et *cerveau*. Chaque mot appartient au langage commun, alors nous remplaçons *glioblastome* par *tumeur du cerveau*.

Pour l'évaluation manuelle, nous avons sélectionné 250 phrases de manière aléatoire. Pour environ 12 % de ces phrases, la substitution a entraîné une légère différence de sens. Ceci peut s'expliquer car parfois les synonymes ne sont stricts. Pour l'évaluation, nous avons utilisé un texte à trou (*cloze test*) qui est une procédure utilisée pour la compréhension d'un texte [Zeng-Treitler *et al.*, 2007]. Si le score du test est compris entre 50-60%, alors le texte est compréhensible. Nous avons recruté 3 personnes n'appartenant pas au domaine médical. Chaque sujet a évalué le texte original et celui simplifié.

compte rendu original	compte rendu simplifié
18%	57%

TABLE 3.5 – Score du test (*cloze test* pour les textes originaux et simplifiés

35 % des termes difficiles n'ont pas été remplacés car ils ne possèdent pas l'annotation *langage courant* dans le réseau JDM. Nos résultats sont proches de ceux de [Zeng-Treitler *et al.*, 2007] bien que notre corpus soit plus important et soit constitué seulement de comptes rendus radiologiques.

Conclusion

Pour améliorer la qualité du réseau et sa couverture, nous avons proposé une approche de consolidation du réseau lexico-sémantique exploitant un moteur d'inférences agissant sur des relations annotées. Le système d'annotation peut être vu comme un complément du système de consolidation du réseau JDM. Ce système propage, grâce à la procédure d'annotation, des informations sémantiques ou d'usages importantes, qui peuvent être utilisées non seulement dans le domaine de la radiologie mais aussi dans d'autres domaines. Nous devons, pour améliorer le système des annotations, penser à améliorer la diffusion des annotations de relations à travers le réseau avec l'aide d'experts, mais aussi de non-experts. Au niveau du réseau JDM, les annotations peuvent être utilisées pour limiter la portée des inférences. Par exemple, nous pouvons limiter la portée du schéma d'inférence par les annotations *possible*. L'utilisation des annotations permet dans le cadre de la recherche d'information, une meilleure présentation (*rang*) des résultats par rapport à la requête formulée par l'utilisateur. Une seconde application des annotations est la simplification des documents médicaux en particulier les comptes rendus radiologiques.

D'autres applications peuvent être envisagées non seulement dans le domaine de la médecine mais aussi dans le domaine général.

Chapitre 4

Constitution de la base connaissances spécialisées dans la base JeuxDeMots et indexation des comptes rendus radiologiques

Sommaire

4.1 Constitution de connaissances spécialisées 104

4.1.1 Corpus de comptes rendus de radiologie 108

4.1.2 Pré-traitement du corpus 109

4.2 Augmentation d'index par propagation à travers le réseau JDM 113

4.2.1 Indexation standard des comptes rendus 114

4.2.2 Algorithme d'augmentation par propagation 119

4.2.3 Évaluation des index augmentés 122

Dans cette partie, nous décrivons notre méthode de constitution de notre base de spécialité à l'intérieur du réseau lexico-sémantique général JDM. A partir cette base, nous expliquons la création de l'index augmenté grâce à un algorithme de propagation à travers ce réseau. La différence principale avec les approches habituelles est l'utilisation simultanée de connaissance générale et de spécialité. Nous utilisons qu'une seule base de connaissance pour la création de cet index augmenté.

Cette partie se décompose comme suit. Dans un premier temps, nous présentons notre approche de constitution de la base de spécialité dans la base de JDM (4.1) à partir de notre corpus de comptes rendus de radiologie (4.1.1). Dans un second temps, nous détaillons l'indexation standard des comptes rendus (4.2.1) avant de présenter l'algorithme d'augmentation d'index par propagation (4.2.2). L'évaluation de l'approche est présentée dans la section (4.2.3). L'augmentation d'index par propagation à travers un réseau lexico-sémantique a fait l'objet de publications [Ramadier et Lafourcade, 2015], [Lafourcade *et al.*, 2015].

4.1 Constitution de connaissances spécialisées

Dans le domaine de la radiologie il n'existe qu'une ressource en langue anglaise, RadLex, mais dont la traduction française est très partielle [Névéol *et al.*, 2014]. Nous pouvons aussi signaler Gamuts (en anglais) qui traite plutôt des diagnos-

tics différentiels[1]. La plupart des travaux concernant le domaine de la radiologie sont en anglais. Le domaine de l'imagerie comporte en outre une certaine spécificité. Outre les termes purement médicaux (termes anatomiques, maladies, termes techniques liés aux différentes modalités) il existe de nombreuses métaphores pour décrire les images radiologiques (par exemple *image en rayon de miel*). Nous appelons métaphore, dans un sens très large, toute image visuelle, toute comparaison exprimée linguistiquement, tout glissement d'un concept à l'autre, tout déplacement sémantique. La métaphore opère par analogie et substitue un référent à un autre en établissant un lien sémantique entre les deux. Pour les domaines scientifiques, des métaphores sont obtenues quand le rédacteur quitte le champ sémantique scientifique pour utiliser des vocables appartenant à d'autres champs. Dans notre domaine d'imagerie médicale, il est intéressant de noter le nombre significatif de métaphores couramment utilisées dans les comptes rendus radiologiques.

Définitions : **Métaphore** :
Figure qui consiste à désigner un objet ou une idée par un mot qui convient pour un autre objet ou une autre idée liée aux précédents par une analogie.

En voici quelques exemples dans le domaine de la radiologie :

— aspect en *verre dépoli*
— perfusion en *mosaïque*
— calcifications pleurales en *os de seiche*
— rein en *fer à cheval*
— aspect en *rayon de miel*
— image en *lâcher de ballons*
— fracture en *motte de beurre*
— fragment méniscal luxé réalisant l'*anse de seau*
— image en *cocarde*
— image en *pelote de laine*
— aspect en *bague à chaton*
— aspect en *crosse de St Nicolas*

1. Le diagnostic différentiel est l'identification d'une pathologie grâce à la comparaison entre eux des symptômes dus à plusieurs affections voisines que l'on cherche à différencier les unes des autres en utilisant un processus d'élimination logique.

Ces termes extraits des comptes rendus ont été ajoutés à la base de données JDM au cours du prétraitement du corpus (cf.4.1.2). A notre connaissance, il n'existe pas d'ontologie médicale en langue française qui intègre ces métaphores. En particulier, la version française de UMLS ne contient pas les métaphores utilisées en imagerie médicale. Une première tentative d'intégrer les métaphores a été réalisée dans Radlex [Shore *et al.*, 2012] mais seulement pour la langue anglaise. Nous avons introduit dans le réseau les signes radiologiques, indépendamment de leur spécificité par rapport à une maladie. Les relations utilisées sont :

- la relation de *cause* (*pneumocéphalie sous tension* r_cause *signe du Mont Fuji*)
- la relation diagnostique qui relie le signe à la modalité d'imagerie (*signe du Mont Fuji* r_diag *scanner*) (figure 4.1)

Les signes radiologiques non spécifiques à une maladie ou à un lieu anatomique sont également renseignés et liés à tous les termes médicaux avec indication fréquentielle sous forme d'annotations, contrairement à la stratégie adoptée par les concepteurs de RadLex [Shore *et al.*, 2012]. Par exemple, le signe *en nid d'abeille* peut être rencontré dans le cadre d'une maladie pneumologique (fibrose pulmonaire idiopathique) ou bien dans certains types d'hémangiomes osseux. Si un signe, fréquent pour une pathologie, peut être présent plus rarement dans une autre, nous faisons la distinction grâce aux annotations de relations.

La difficulté liée à ces métaphores est que dans le cadre de cette thèse, elles proviennent d'un seul corpus. Or, certains radiologues utilisent d'autres métaphores différentes de celles utilisées dans nos comptes rendus.

Par conséquent, pour la réalisation d'une base de connaissances dans le domaine de la radiologie, nous avons utilisé le réseau JDM. Cela nous a permis de construire une base de connaissances de spécialité à l'intérieur d'un réseau de connaissances générales (**hypothèse de non-séparation** entre un réseau de connaissances de sens commun et un réseau de spécialité). Nous émettons l'hypothèse que pour l'analyse fine des comptes rendus de radiologie (c'est-à-dire la découverte d'informations implicites) ou pour une simplification des termes à destination des patients, l'utilisation d'un réseau possédant des connaissances générales et de spécialité apporte un gain significatif de performance. La construction du domaine de spécialité (la radiologie) a été réalisée à partir d'un corpus de comptes rendus radiologiques et

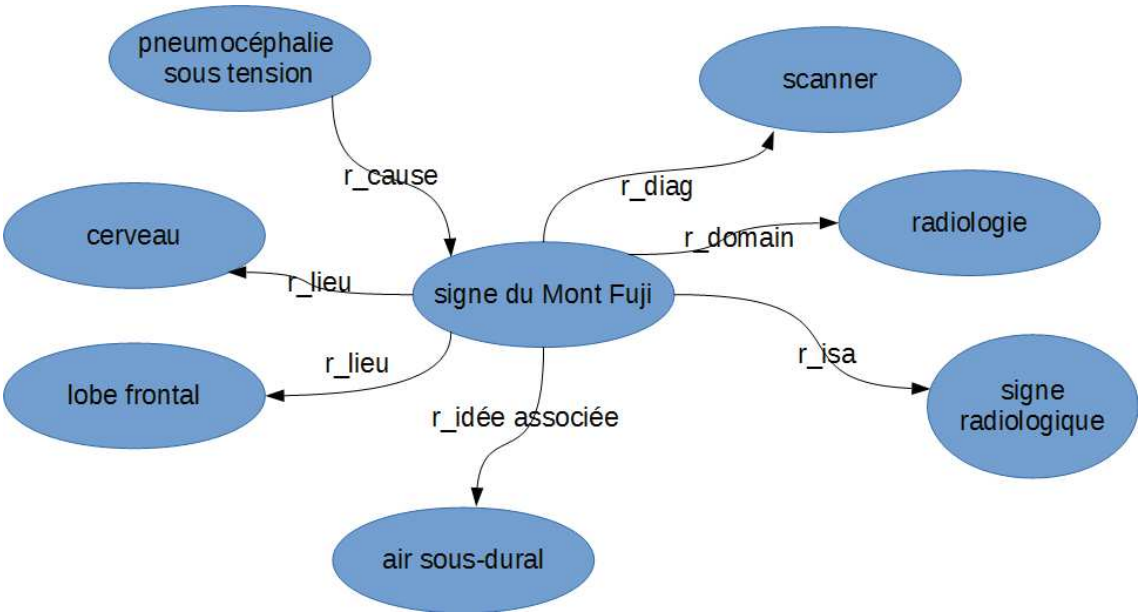


FIGURE 4.1 – schéma montrant les relations pour un signe d'imagerie médicale (signe du Mont Fuji)

d'autres sources (Orphanet, Radlex, Wikipédia).

Le réseau JDM contient 80 types de relations différentes (des relations hiérarchiques comme l'hyperonymie, l'hyponymie, des relations non taxonomiques (cause, conséquence, caractéristique,...) des relations lexicales, etc..). Dans la majorité des cas, nous avons utilisé les relations déjà existantes dans le réseau car elles pouvaient correspondre à une notion médicale. Par exemple la relation r_against-1 (qu'est-ce qui s'oppose à) correspond à la relation *traitement* utilisée dans les ontologies biomédicales, les relations *causes* et *conséquences* sont aussi bien explicites dans le domaine médical que pour le domaine général.

Pour les relations concernant la vascularisation des vaisseaux sanguins ou l'innervation, nous nous servons de relations spatiales (par exemple r_predecesseur_space) qui peuvent aussi servir à d'autres domaines, comme la géographie (*affluents*), par exemple. Nous avons, cependant, créé des relations spécifiques à la médecine : la relations *signe* (r_symptom), la relation *cible* (r_target) ou encore la relation *moyen de diagnostic* (r_diagnostic). La relation *signe* (r_symptom) correspond aux différends symptômes d'une maladie (*fièvre, douleurs abdominales, toux, etc*), la relation *cible* désigne le type de patients qui peut être atteint préférentiellement par une

embolisation
motif de l'examen : patient de 32 ans avec antécédent d'angiomyolipomes bilatéraux et hémo-rétro-péritoine avec déglobulisation.
technique et résultat : Cathétérisme Seldinger droit 4f après anesthésie locale et asepsie cutanée.
Mise en place d'une sonde cobra dans une artère polaire supérieure droite qui vascularise exclusivement l'angiome myolipome du rein droit.
Absence d'extravasation active de produit de contraste. par contre, on visualise de multiples faux anévrismes en regard de la lésion.
En conséquence, occlusion de cette artère à l'aide de 3 coils fibrés de type cook de diamètre 3mm.
L'opacification de l'artère rénale droite située 1 cm en dessous montre une bonne parenchymographie rénale sans extravasation de produit de contraste.
En conséquence, la procédure est arrêtée.
Fermeture par compression manuelle.
En conclusion : Embolisation en urgence d'un hémo-rétropéritoine sur saignement d'angiomyolipome polaire supérieur droit avec occlusion sélective par coils d'une artère polaire supérieure droite.

FIGURE 4.2 – Exemple de compte rendu original

affection (*enfant, nouveau-né, adulte jeune, alcoolique, toxicomane,etc*) et la relation *moyen de diagnostic* évoque les examens médicaux réalisés en vue d'aider au diagnostic (*IRM, scanner, échographie, analyse de sang, etc*).

4.1.1 Corpus de comptes rendus de radiologie

Notre corpus est constitué de 35 000 comptes rendus radiologiques quotidiens (figure 4.2) auxquels on a ajouté 1000 textes provenant du site d'e-learning d'IMAIOS. Ces documents englobent les différentes modalités d'imagerie médicale (imagerie par résonance magnétique, tomographie à rayon X, échographie, radiologie conventionnelle, radiologie vasculaire). Les comptes rendus sont écrits de manière semi-structurée, c'est-à-dire qu'ils sont généralement divisés en quatre parties distinctes (*indications, technique, résultats* et une *conclusion* optionnelle) (voir Annexe A).

Comme il sera décrit dans le paragraphe 4.2.1, nous réalisons un prétraitement en vue de reconnaître les mots composés (*cathétérisme Seldinger, aseptie cutanée, angiomyolipome du rein, parenchymographie rénale, ...*). .

Chaque partie est rédigée par le médecin radiologue sous une forme très libre, avec souvent une profusion d'acronymes (*ATCD* pour *antécédent*, *ACR* pour *american college of radiology*, *tt* pour *traitement*, *SA* pour *semaine d'aménorrhée*), *UH* pour *unité Hounsfield*, d'élisions (par exemple, la *communicante antérieure* au lieu de l'*artère communicante antérieure*), des variantes orthographiques (*hémorétropéritoine et hémo-rétro-péritoine*) et toutes sortes d'incorrections diverses (fautes d'orthographes, et parfois création d'un nouveau terme lexical).

Dans le traitement de ce corpus, nous n'avons pas utilisé les outils d'analyse grammaticale ou syntaxique car notre corpus possède un langage assez pauvre grammaticalement (absence de verbes, omission des articles,...). De même pour la lemmatisation, nous nous sommes servis du réseau lexico-sémantique JDM où cette information est présente.

Il a déjà été montré [Abeillé et Blache, 1997] que l'on peut se passer de syntaxe dans le cas où l'on travaille sur des domaines spécialisés (la radiologie dans notre travail) et en utilisant des bases de connaissances relativement détaillées. Le réseau lexico-sémantique JDM possède suffisamment d'informations implicites pour lever les éventuelles ambiguïtés (grâce notamment aux raffinements). Par exemple, le terme *coup de poignard* est ambigu car il peut s'agir d'un *coup de poignard* (arme blanche) ou bien douleurs en *coup de poignard* et il sera désambiguïsé car l'information est contenue dans JDM. Il est important d'avoir une base qui puisse distinguer les deux usages parce que les deux significations peuvent être présentes dans les comptes rendus.

4.1.2 Pré-traitement du corpus

La première étape a consisté en une série de prétraitements (transformation PDF en format texte, anonymisation). L'anonymisation n'a pas été extrême car le corpus utilisé n'a pas été mis à la disposition du public. Nous avons, cependant, automatiquement supprimé le nom, le prénom du patient ainsi que le nom du médecin et le nom de l'institution. Le numéro patient a aussi été enlevé. Par contre, nous avons gardé l'âge si il était renseigné ainsi que les indications concernant un appareil médical implantable (numéro d'identification). La difficulté de l'anonymisation

dans notre corpus provient du fait que les documents provenaient de différentes institutions. Les informations liées aux patients n'étaient pas indiquées de manière homogène. Vu que le corpus n'a pas été rendu public, nous n'avons pas réalisé une évaluation précise de la qualité de l'anonymisation.

4.1.2.1 Constitution de la base de spécialité à partir du corpus de radiologie

Après ce traitement, nous sommes passés à la phase de construction proprement dite de la base de spécialité. Dans un premier temps, nous avons réalisé une indexation de tous les termes simples médicaux (*glioblastome*, *adénocarcinome*, *sarcome*, etc) et non médicaux (*encornage*, *couteau*, *baignoire*, etc) présent dans les comptes rendus. Si les termes ne sont pas présents dans la base JDM alors ces derniers sont soumis à validation par des experts en vue de leur incorporation. Cette base de spécialité a ensuite été complétée par indexation des termes composés ou concepts (multi-termes => bigrammes (artère_rénale), trigrammes (artère_rénale_polaire), quadrigrammes (artère_cérébrale_antérieure_droite), ...) en utilisant un algorithme d'index inversé (approche itérée pour le recherche multi-termes) (table 4.1). Cette approche permet un repérage des termes voisins (cooccurrences) et ainsi l'extraction de termes. Les informations pertinentes qui sont indexées sont celles qui sont susceptibles de faire l'objet de requêtes par les utilisateurs. Nous présentons le schéma général de notre système (figure 4.3) :

nombre d'occurrence	**termes**
791	artère
137	artère_rénale
6	artère_rénale_accessoire
1217	muscle
61	quadriceps
4	muscle_quadriceps
1	muscle_quadriceps_fémoral

TABLE 4.1 – approche itérée pour la détection des multi-termes.

Nous pouvons noter que le terme *muscle quadriceps* apparaît seulement 4 fois, ce qui peut paraître peu élevé alors que le terme quadriceps seul apparaît 61 fois. Ceci s'explique par le fait que les médecins radiologues font souvent des ellipses (*rupture des croisés* au lieu de *rupture des ligaments croisés du genou*) voire des apocopes (*trauma de la face* au lieu de *traumatisme de la face*, *scan* au lieu de *scanner*, *angio* au lieu de *angiographie*). Toutes ces informations sont présentes dans le réseau afin de faciliter l'indexation. En ce qui concerne les problèmes liés aux variations de la morphologie des mots, nous ne cherchons pas à les normaliser et les intégrons tels quels dans le réseau. Par exemple dans la base, *IRM du cœur* et *IRM myocardique* seront présent et reliés entre eux par la relation *synonyme strict* (IRM du cœur *r_syn_strict* IRM myocardique). Concernant la combinaison des mots, grâce à notre algorithme itérée tous les concepts différents seront représentés dans la base. Si nous prenons l'exemple *muscle oblique externe de l'abdomen*, les différents concepts seront capturés, c'est-à-dire que *muscle_oblique*, *muscle_oblique_externe* et *muscle_oblique_externe_de_l'abdomen* seront identifié comme concepts et par conséquent présents dans la base. Un autre exemple est *lobe_pulmonaire*, *lobe_pulmonaire_inférieur*, *lobe_pulmonaire_inférieur_gauche* où tous ces termes seront présent dans la base puisqu'il peuvent apparaître sous ces différentes formes dans les comptes rendus.

L'extraction de mots composés est une étape importante car elle facilitera, entre autres, la désambiguïsation dans l'analyse de texte. Les termes composés ont, en général, un seul sens alors que les mots le composant peuvent être polysémiques. Si nous considérons le terme lame_d'épanchement, il est bien monosémique alors que le mot *lame* et *épanchement* pris séparément sont polysémiques. Le terme *lame* possède 10 raffinements et épanchement 4 raffinements sémantiques dans le réseau. Nous avons ainsi, à partir des comptes rendus, indexé 10 000 termes liés à *médecine* et 5 000 termes d'*anatomie*.

Cette indexation a été complétée, dans un deuxième temps, par différentes sources pour améliorer le réseau. Nous avons utilisé comme source le site Orphanet [2] qui est un site d'informations sur les maladies rares et les médicaments orphelins en libre accès. Si une entrée n'était pas présente dans le réseaux alors nous l'avons invoquée et nous avons ajouté les termes de la description de la même façon que

2. http ://www.orpha.net/consor/cgi-bin/index.php ?lng=FR

précédemment.

4.1.2.2 Sources complémentaires pour la constitution de la base de spécialité

Nous avons également complété avec la version française de RadLex (même si elle n'est que très partielle). Ce travail a été effectué par des spécialistes de l'imagerie médicale afin de vérifier la pertinence des informations. Dans un premier temps, ce travail a été réalisé manuellement au vu de la structure de Radlex et celle de JDM. Enfin, nous avons aussi utilisé la structure des hyperliens des pages Wikipédia pour enrichir le réseau et en particulier les articles liés au domaine médical (table 4.2) . Dans un premier temps tous les hyperliens pour une page, dont une entrée existe dans *Diko*, sont récupérés et ajoutés à la relation *idée associée* du réseau JDM. Ces nouvelles contributions sont, elles aussi, soumises à une validation par un expert. En effet, certains hyperliens ne sont pas pertinents par rapport à l'article précédent. Si nous considérons l'article *Technique de Seldinger*[3] les hyperliens liés aux dates de naissance et de décès de l'inventeur de la technique ne sont pas pertinentes et ne seront pas validés par l'expert.

Titre Wikipedia	**hyperliens**
myosite	maladie de Lyme, myosite ossifiante, etc
diverticulite du sigmoïde	diverticule, sigmoïde, côlon, etc
muscle deltoïde	muscle, épaule,clavicule, acromion, etc
muscle triceps brachial	muscle, os humérus, développé couché

TABLE 4.2 – Exemple de page **Wikipédia** et des hyperliens extraits.
Les termes sont rajoutés au réseau après validation par un expert. Ils sont rajoutés à la relation idée associée

Ceci nous permet d'enrichir la base de connaissance JDM. Un des travaux ultérieurs sera de déterminer la bonne relation entre les hyperliens et le terme principal (entrée Wikipedia) pour pouvoir ventiler de façon plus efficace les termes (les relations correctes).

3. https ://fr.wikipedia.org/wiki/Technique_de_Seldinger

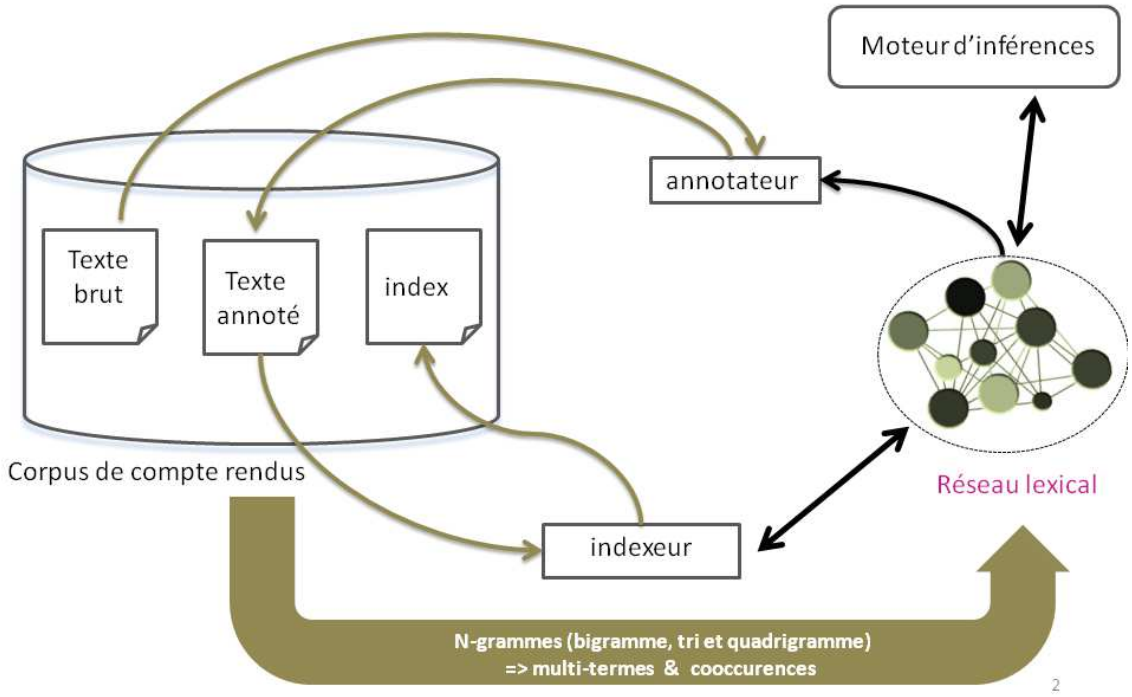


FIGURE 4.3 – Schéma général de l'ajout de connaissances spécialisées dans le réseau JDM.

A partir de l'index brut, nous mettons dans le réseau les termes sélectionnés, puis nous repérons les termes composés. Endogène au réseau, un moteur d'inférence permet de découvrir de nouvelles relations.

4.2 Augmentation d'index par propagation à travers le réseau JDM

La tâche d'indexation, afin d'être utile aux praticiens, doit tenir compte des requêtes qu'ils effectuent. Plusieurs auteurs, notamment [Hersh *et al.*, 2001] et [Huang *et al.*, 2003] ont réalisé une indexation automatique des comptes rendus radiologiques en se fondant sur le métathésaurus UMLS. Pour améliorer leurs résultats (en particulier

en ce qui concerne la précision) ils ont utilisé une sous-section de la terminologie UMLS. Toujours dans l'optique d'augmenter la précision, [Hersh *et al.*, 2001] ont délibérément écarté certaines parties des comptes rendus, en particulier la section *indications* ; ils ont ainsi obtenu un index ne contenant que les termes strictement médicaux. Or en pratique, dans leur requêtes, les utilisateurs d'un système de recherche dédié à la radiologie ont besoin de rechercher non seulement des termes médicaux précis (*perforation digestive, glioblastome, astrocytome, syndrome de Guillain-Barré*) mais aussi des termes composés ou des périphrases de sens général (*accident de ski, femmes jeunes, coups de couteau, armes blanches, coup de sabot, encorner, bouteille de bière, etc*). Nous appelons *termes médicaux précis* les termes qui dans le réseau utilisé pour ce travail sont liés par la relation *domaine* à la *médecine*. Les termes généraux, quant à eux, ne sont pas directement liés par cette relation à la *médecine* mais peuvent l'être à un voisinage de un ou plus.

A notre connaissance, l'indexation automatique de comptes rendus radiologiques a jusqu'à maintenant essentiellement porté sur les termes strictement médicaux, sans tenir compte des informations d'ordre général (chute, accident domestique, arme blanche,..). Dans cette partie, nous décrivons comment, à partir des informations sémantiques, il est possible de définir une augmentation des index bruts construits pour chaque compte rendu afin d'améliorer le rappel de la recherche documentaire. En effet, les médecins radiologues peuvent exprimer leurs requêtes en utilisant des génériques (par exemple, tumeur bénigne du cerveau, tumeur du cerveau, tumeur bénigne, tumeur), des conséquences, des circonstances, etc, sans que ces termes ne soient pour autant explicitement présents dans les comptes rendus.

4.2.1 Indexation standard des comptes rendus

Les différentes étapes pour la création de l'index inversé des termes pour un corpus de documents sont :

- l'extraction de paires d'identifiants (termes, document).
- le tri des paires suivant les clés d'identifiants
- le regroupement des paires en établissant pour chaque identifiant de terme, la liste des identifiants de documents dans lesquels le terme apparaît.

Nous utilisons les méthodes classiques de la recherche documentaire pour ne sélectionner que les termes pertinents, c'est-à-dire la fréquence des termes (TF) et la fréquence documentaire (DF) pour calculer l'IDF (*Inverse Document Frequency*). Ce score permet de donner une importance au concept en fonction de sa fréquence dans le document (TF) pondérée par la fréquence d'apparition du concept dans tous le corpus (IDF). Si DF est le nombre de document du corpus contenant le terme, alors un DF égal à 1 correspond au fait que le terme n'apparaît que dans un document et le tfidf sera fort. Quand le DF est proche du nombre de document dans le corpus, le tfidf sera faible.

D1 = Fracture multifragmentaire, stable de l'aile_iliaque gauche sans atteinte du cotyle. (10 mots)
D2 = Fracture de l'épine_iliaque antéro-supérieure gauche (5 mots)
fracture -> [1,[0]],[2,[0]] =>$tf_{11} = \frac{1}{10}$ tf_{12}=$\frac{1}{5}$ • **DF** = 2
multifragmentaire ->[1,[1]] =>$tf_{21} = \frac{1}{10}$ • **DF**= 1
aile_iliaque ->[1,[5]] =>$tf_{31} = \frac{1}{10}$ • **DF** = 1
gauche ->[1,[6]],[2,[5]] =>$tf_{41} = \frac{1}{10}$ tf_{12}=$\frac{1}{5}$ • **DF** = 2
cotyle ->[1,[9]] =>$tf_{51} = \frac{1}{10}$ • **DF** = 1
épine_iliaque[2,[3]] => tf_{32}=$\frac{1}{5}$ • **DF** = 1
antéro-supérieure[2], [4]] => tf_{32}=$\frac{1}{5}$ • **DF** = 1

Exemple d'index inversé avec le numéro du document,la **position** dans le document, **TF** et **DF**

L'ajout d'un nouveau document dans le système nécessite de recalculer les scores TFIDF. Il s'avère que quand le nombre de documents dans le corpus est important, l'ajout d'un nouveau document ne modifie pas énormément les autres scores tfidf. Le recalcul complet peut donc être légèrement différé.

La reconnaissance des termes composés est effectuée en amont par comparaison au contenu JDM. Un tiret bas remplacera l'espace entre chaque élément d'un terme composé (*fracture_du_tibia*) afin de les conserver comme unité autonome lors de l'extraction.

Nous avons optimisé l'algorithme de recherche des multi-mots. En effet, il serait rédhibitoire de parcourir tous les mots multiples du réseau pour savoir si ils sont

présents dans la phrase. Nous proposons une modification de l'algorithme afin de limiter le nombre d'itérations (algorithme 2) (algorithme de recherche de mots composés). Tous les mots multiples sont stockés dans un tableau avec l'identifiant (*id*) pour chaque mot de ces constituants. Ce tableau est ensuite trié par ordre d'identifiant croissant. Ce tableau ne peut contenir plus de 6 éléments (les mots composés de plus de 6 éléments sont rares dans un compte rendu). Au moment de l'analyse de la phrase, nous appliquons la même méthode en ne tenant compte que des mots apparaissant au moins une fois dans un mot composé. Cette méthode nous permet de réduire le nombre d'itérations. Nous allouons les codes d'identifications des mots de référence par ordre inverse d'apparition d'un des mots de la famille dans un mot multiple, ainsi si nous avons les termes *maladie de Basedow*, *maladie de Kaposi*, ils seraient reconnus par leur dernier terme et non par *de* ou *maladie*.

algorithme 2 rechercheDeMotComposés

A_{mot} := termes de la phrase stockés dans un tableau
A_{mp} := mots présent dans un mot multiple stockés dans un tableau
/* s_{tmp} la taille de A_{mp} */
pour j (1...s_{tmp} - 1) **faire**
  Chercher les mots multiples dont l'élément 1 est égal à A_{mp}[j]
  Verifier si tous les autres composants appartiennent bien à A_{mp}[i + 1...s_{tmp}]
fin pour

Pour la désambiguïsation (algorithme 3) des termes, nous avons utilisé une méthode dérivée de l'algorithme de Lesk ([Lesk, 1986]). Nous comparons le contexte dans le texte (comptes rendus) du mot à désambiguïser aux termes qui lui sont liés dans le réseau JDM. Nous calculons l'intersection et nous sélectionnons la valeur la plus élevée. Par rapport à l'algorithme de Lesk, nous ne nous servons pas de la définition du mot à désambiguïser mais des termes qui possèdent des relations avec le mot à désambiguïser.

Malgré le filtrage fréquentiel, nous conservons les termes situés au voisinage de la médecine même pour de faibles valeurs du TFIDF. Si un mot simple ou composé du texte est dans le réseau JDM, et qu'il est lié par la relation domaine au terme *médecine* (voisinage à distance 1) alors il est ajouté à l'index. De même, des termes non médicaux (*accident de moto, prise de drogue, boule de pétanque*) sont également capturés et ajoutés à l'index dès lors qu'ils possèdent une relation avec un terme lui-même lié à *médecine* (voisinage à distance 2) : ainsi *accident de moto* est ajouté

algorithme 3 DesambiguisationDesTermes

```
Entrée : le mot à désambiguiser et la phrase
Sortie : glose correcte
fonction Desambiguisation(mot, phrase)
gloses ← sens le plus fréquent du mot (poids dans le réseau)
max-overleap ← 0
contexte ← ensemble des mots de la phrase
pour chaque sens du mot faire
  signature ← ensemble des termes liés aux mots dans JDM
  overlap ← CalculChevauchement(signature, contexte)
  si  overlap > max-overleap alors
    max-overleap ← overlap
    glose ← sens
  fin si
fin pour
retour glose
```

car lié à *polytraumatisé* par la relation *conséquence* et *polytraumatisé* est lui-même lié à *médecine* par la relation *domaine*. Nous présentons des exemples de comptes rendus avec leur index brut associés (figure 4.4 et 4.5).

En observant ces index bruts, nous constatons que nous n'avons pas réussi à indexer les termes reliés par la conjonction *et*. En effet, nous capturons correctement *fenêtre parties molles* mais pas *fenêtre osseuse* du fait que fenêtre et osseuse se trouvent séparés. Pour cela il suffit de réaliser une analyse syntaxique pour pouvoir rattacher *fenêtre* et *osseuse* ensemble.

Par ailleurs, pour chaque terme de l'index brut, il est intéressant de déterminer le ou les bons raffinements sémantiques (s'il en possède). Par exemple, dans le compte rendu ci-dessus, les termes *fracture, cheville, chute et loge* sont polysémiques. Dans certains cas, la désambiguïsation peut concerner des termes qui possèdent une signification à la fois dans le domaine général et dans le domaine médical. Par exemple dans la phrase *on note une lame d'épanchement provoquée par la lame de l'arme blanche*, le terme *lame* a bien deux significations différentes. Pour les acronymes il est important de les désambiguïser car un même acronyme peut avoir des significations différentes même à l'intérieur du domaine médical. Par exemple *FO* peut signifier *fond d'œil* (ophtalmologie) ou *foramen ovale* (cœur ou cerveau). Les résultats montrent que la détermination des bons raffinements a de l'importance surtout

indications : fracture du tibia droit, chute de ski **technique** : une série de coupes axiales transverses sur l'ensemble de la cheville sans injection de produit de contraste. etude en fenêtres parties molles et osseuses. **Resultats** fractures diaphysaires spiroïdes à trois fragments principaux du 1/3 distal du tibia et de la fibula avec discret déplacement vers l'avant, sans trait de refend articulaire. Fractures de la base de M2 et de M3 non articulaires et non déplacées. Fracture articulaire de la partie interne de la base de M1 non déplacée. Atrophie avec dégénérescence marquée des corps musculaires de l'ensemble des loges.
articulaire • atrophie • cheville • chute • chute de ski • corps musculaire • coupe axiale transverse • dégénérescence • déplacement • diaphysaire • fenêtres parties molles• fibula • fracture • fracture du tibia • fracture articulaire • loge>anatomie • non articulaire • non déplacée • sans injection de produit de contraste • ski • spiroïde • tibia • trait de refend

FIGURE 4.4 – Exemple d'un compte rendu et de son index brut

indications : trauma facial. **technique** : une série de coupes axiales transverses sur l'ensemble de la face sans injection de produit de contraste. étude en fenêtres parties molles et osseuses. **Resultats** Fracas facial complexe type Le Fort III associé à des traits de fractures multiples naso-ethmoïdo-maxillaire et zygoma. Fracture des deux lames papyracées et des parois latérales et supérieures de l'orbite. Fracture de l'ensemble des parois des sinus maxillaires en bilatérales. Pas d'incarcération musculaire. Fractures des ptérygoïdes. Fracture du sinus frontal. Fracture des OPN. Hémosinus frontal ethmoïdal et maxillaire. **Conclusion** : fracas facial complexe type Le Fort III
trauma facial • fracas facial • coupe axiale transverse • Le Fort III • fracture • lame papyracée • ptérygoïde • sinus frontal • OPN • naso-ethmoïdo-maxillaire • face>visage • fenêtre partie molle • frontal • hémosinus • zygoma

FIGURE 4.5 – Exemple de comptes rendus et de leur index brut.

pour l'algorithme de propagation afin d'éviter de propager le signal à partir de mauvais raffinements.

4.2.2 Algorithme d'augmentation par propagation

Avant de réaliser l'augmentation d'index proprement dite, nous ajoutons à l'index brut les variantes orthographiques ainsi que les synonymes stricts. Parfois, la notion de synonymie est assez large même dans le domaine médical ce qui peut poser problème. En effet *kyste* et *abcès* peuvent apparaître comme synonymes alors que médicalement les deux entités ont des conséquences thérapeutiques très différentes. De plus l'aspect radiologique d'un kyste et d'un abcès n'est pas le même. C'est pour cela que dans notre système d'indexation, nous avons introduit la relation *substitut strict*. Ces substituts stricts sont proposés par des des contributeurs via l'outil Diko puis validés par un expert. L'augmentation est un processus visant à rajouter dans l'index des termes pertinents, mais non présents dans le texte. Pour constituer l'index augmenté à partir de l'index brut, nous adoptons une stratégie consistant à propager des signaux sur le réseau JDM à partir des termes de l'index brut. L'idée est d'*allumer* les termes de l'index brut et de récupérer à leur suite les termes du réseau qui s'allument également. Il est évidemment important, vu la nature du réseau JDM, de réaliser un filtrage (nous expliquons ce filtrage ci-dessous) afin de sélectionner les termes les plus pertinents. En effet, le réseau contenant à la fois des connaissances générales et de spécialité, il a été décidé de réaliser un filtrage selon les poids pour éviter d'avoir trop de bruit. En effet, pour certaines relations il n'est pas nécessaire de récupérer tous les termes liés aux mots de l'index brut mais seulement les plus pertinents.

Définitions : **Augmentation d'index** :
Processus ayant pour objectif à rajouter dans l'index brut des termes pertinents, mais non présents dans le document.

D'autres exemples d'index brut et augmentés sont donnés en annexe (Annexe C et D). A chaque cycle, les termes propagent en parallèle leur activation courante vers leur voisin. L'activation totale n'est que la mémoire des décharges reçues par un

accident de ski • **accident de sport d'hiver** • accident de sport • atrophie • cheville • **cheville>anatomie** • chute • **chute>tomber** • corps musculaire • coupe axiale transverse • dégénérescence • **dégénérescence musculaire** • déplacement • fibula • fracture • **fracture articulaire** • **fracture des membres inférieurs** • **fracture multiple** • **fracture diaphysaire** • **fracture du tibia** • **fracture non articulaire** • **fracture non déplacée** • **fracture spiroïde** • **fracture avec déplacement** • **fracture>lésion** • **imagerie médicale** • **jambe** • **lésion** • **lésion osseuse** • loge • **loge>anatomie** • **médecine** • non articulaire • non déplacée • péroné • **radiologie** • ski • **spiroïde** • **sports d'hiver** • tibia • trait de refend • **trauma** • **traumatisme des membres inférieurs** •

FIGURE 4.6 – Index augmenté correspondant à la figure 4.4
Les termes sont triés par ordre alphabétique avec en gras les termes ajoutés. Les thématiques générales du texte sont bien identifiées (médecine, imagerie médicale, radiologie). Les termes polysémiques ont été raffinés avec leur usage correct en contexte.

classification de Le Fort • coupe axiale tranverse • **disjonction craniofaciale** • **ethmoïde** • fracas facial • fracture • **fracture de Le Fort III** • **hématome en lunettes** • **imagerie médicale lame>anatomie** • lame papyracée • Le Fort III • **massif facial** • **médecine** • **scanner>médecine** • trauma facial • **traumatisme facial** • **traumatisme>physique** • **traumatologie** •

FIGURE 4.7 – Index augmenté correspondant à la figure 4.5.

terme sur l'ensemble du processus. Pour les relations à poids négatif ou inhibitrices, l'activation n'est pas ajoutée mais soustraite. Un terme avec une AC (activation courante) négative ne décharge pas. On remarquera que la distribution du signal se fait proportionnellement aux logarithmes des poids (w) (et non pas proportionnellement aux poids eux-mêmes). A l'issue de l'itération, nous obtenons une liste de termes pondérés que nous ordonnons par poids décroissants. Nous retenons (filtrage) les N termes de poids les plus forts tels que la somme de leur poids représente S% du poids total des termes de cette liste. Plus précisément, l'algorithme que nous avons mis au point s'énonce comme suit (algorithme 4).

algorithme 4 PropagationIndex

INITIALISATION : Les termes du réseau M sont associés à un couple de valeurs (AC, AT), *//activation courante et activation totale*
pour les termes T $\in$ *indexbrut*, nous fixons AC=AT=1. *//Les T sont les sources d'activation*
pour tous les autres termes, AC=AT=0.
nous fixons un nombre d'itération NBI
Nous répétons NBI fois l'opération suivante :
pour chaque terme T du réseau ayant des voisins (t_1,...,t_n) via une relation de type r de M vers t_i de pondération w_i **faire**
AC(t_i) = AC(t_i) + AC(M) * $\frac{log(w_i)}{\sum_{k=0}^{n} log(w_k)}$
AT(t_i) = AT(t_i) + AC(t_i) // *on mémorise ce que reçoit t_i dans AT(t_i)*
AT(T) = 1 *//tous les M ont déchargé leur activation, on recharge les T*
fin pour
SORTIE : filtrage de termes activés avec pourcentage de surface S ; nous retournons les termes activés restants.

Nous n'exploitons pas toutes les relations disponibles dans JDM, car certaines, très lexicales, auraient tendance, dans le cadre de notre application, à dégrader la précision. Les relations que nous utilisons sont les suivantes : *idées associées* , *hyperonymes*, *synonymes*, *caractéristiques typiques*, *symptômes, diagnostiques, parties/tout, lieux typiques, causes, conséquences, domaine, et fréquemment associé à.* Dans l'algorithme, tous les types de relations ont un poids identique. Pour la relation d'hyperonymie, nous sélectionnons tous les hyperonymes du terme (du plus spécifique au plus général). Par exemple pour le terme *cavernome* nous prenons les termes suivants : **malformation vasculaire, hémangiome, hémangiome caverneux, maladie vasculaire**,...). Pour les synonymes, nous avons inclus aussi les phénomènes d'apocope, d'élisions pour diminuer les silences. Par exemple pour le

terme *traumatisme facial*, nous avons rajoutés le terme **trauma facial** (apocope) dans la catégorie synonyme.

4.2.3 Évaluation des index augmentés

Nous avons évalué l'algorithme de propagation de manière statistique par une sélection aléatoire de 200 index augmentés (sur 30 000 calculés). Chaque terme de l'index augmenté a été manuellement évalué comme pertinent ou non. L'évaluation manuelle a été réalisée par un expert en imagerie médicale. Un terme (ou multi-terme) est considéré comme pertinent par un spécialiste lorsqu'il est susceptible de faire l'objet de requêtes. Les couples de valeurs du tableau (table 4.3 sont donc :

— le nombre moyen de termes de l'index augmenté qui ne sont pas dans l'index brut (valeur *nouv*).

— le pourcentage moyen de termes pertinents de l'index augmenté (valeur *pert*)

NBI/S	**10%**	**20%**	**30%**	**40%**	**50%**
1	22/82 %	45/80 %	67/78 %	93/53 %	127/38 %
2	31/95 %	55/92 %	83/89 %	211/57 %	439/41 %
3	48/99 %	**90/97** %	139/95 %	356/53 %	755/34 %
4	111/97 %	**223/92** %	335/87 %	747/45 %	1259 /23 %
5	387/96 %	774/87 %	1161/76 %	1671/26 %	2089/15 %

TABLE 4.3 – Résultats de l'algorithme de propagation.
Présentation des valeurs nouv (à gauche de chaque colonne) et pert (à droite) en fonction des paramètres NBI et S. NBI est le nombre d'itérations effectuées dans le réseau lexical. S est la part retenue de la surface sous la courbe des poids cumulés des termes atteints par l'algorithme de propagation.

En pratique, l'évaluation de la valeur *pert* n'a besoin d'être réalisée qu'une seule fois indépendamment des paramètres NBI (nombre d'itérations effectuées dans le réseau lexical) et S (surface sous la courbe des poids cumulés des termes atteints par l'algorithme). En effet il suffit pour un compte rendu de considérer l'ensemble des termes obtenus pour toutes les valeurs possibles des paramètres, puis d'évaluer la pertinence de chaque terme dans l'ordre de leur poids décroissant. Au bout de 5 termes consécutifs non pertinents, nous considérons que tout ce qui suit est également non pertinent. La valeur *nouv*, elle, peut être calculée automatiquement. Pour un même nombre d'itérations, plus la surface retenue est grande, plus le nombre de termes atteints est important (filtrage faible). C'est-à-dire que le rappel est d'autant plus

important, mais en contrepartie la précision a tendance à diminuer (voire s'écroule au delà de 30%), les termes ajoutés à l'index brut ayant tendance à être de moins en moins pertinents. A l'inverse, plus le nombre d'itérations augmente, plus les termes pertinents sont renforcés (ce sont les voisins mutuels des termes de l'index brut). Le réseau lexical contient des boucles (directes et indirectes) qui agissent comme autant d'auto-renforcements. Le temps de calcul croît considérablement à chaque nouvelle itération, le nombre de termes déchargeant leur activation augmentant très fortement.

Pour NBI = 5, la quasi-totalité du réseau est atteinte (si on exclut le filtrage par S), le diamètre étant d'environ 6 (le réseau JDM est de type petit monde). Globalement, la zone qui semble la plus intéressante pour un temps de calcul raisonnable (quelques secondes) correspond à 3 ou 4 itérations pour une surface inférieure à 30% (valeurs en gras dans la table 4.3) . La meilleure précision est obtenue pour une valeur d'itération égale à **3** et pour une surface de **10%**. Pour une valeur d'itération à 5, il faut prendre une faible surface pour garder une valeur de précision satisfaisante. En effet, pour une telle valeur d'itérations, une très grande partie du réseau a été atteint et par conséquent, un filtrage serré (S =10% ou 20%) est donc nécessaire. Quand NBI = 5 et S = 40% ou 50%, la précision est trop faible pour être prise en compte. Dans l'élaboration de nos index augmenté nous ne gardons pas ceux où NBI = 5.

NBI/S	**10%**	**20%**	**30%**	**40%**	**50%**
1	77 %	71%	68 %	47.7%	34.2 %
2	87 %	80 %	70 %	51%	35.4 %
3	90 %	89 %	87 %	49 %	29 %
4	86 %	82 %	79 %	38 %	18 %
5	88 %	78.5 %	65 %	23.4 %	12 %

TABLE 4.4 – Présentation des nouvelles valeurs *pert* **sans les raffinements** en fonction des paramètres NBI et S.
NBI est le nombre d'itérations effectuées dans le réseau lexical. S est la part retenue de la surface sous la courbe des poids cumulés des termes atteints par l'algorithme de propagation.

La totalité des termes ambigus a été correctement désambiguïsée. Comme vu précédemment, des termes peuvent être présents dans le texte avec une signification à la fois dans le sens général et dans le sens médical (exemple avec le terme *lame*). Cela

signifie que l'index augmenté a systématiquement inclus le bon raffinement quand un raffinement était proposé (ce n'est pas forcément le cas pour des valeurs faibles de NBI et de S). Nous avons recalculé les index augmentés en interdisant les accès aux termes raffinés, et nous avons constaté une baisse globale d'environ 10% de la valeur de *pert* quelle que soit la configuration (NBI et S) (table 4.4). Chercher à sélectionner les sens corrects des termes polysémiques peut donc être réalisés conjointement à la sélection de termes pertinents et aurait même tendance à la favoriser.

Enfin, tous les domaines identifiés ont été pertinents. Rajouter les domaines pertinents dans l'index brut avant l'augmentation ne change pas significativement les résultats.

Nous avons également recalculé les index (bruts et augmentés) en n'autorisant l'accès qu'aux termes directement liés au terme *médecine*, c'est-à-dire aux termes décrivant les maladies ainsi qu'aux termes anatomiques, quelle que soit la relation durant la propagation. Cela s'est traduit par une diminution moyenne de 12% de la valeur *pert*. Il semblerait, au vu des résultats obtenus, que l'utilisation d'une base de connaissances non limitée au domaine de spécialité améliore de façon significative la pertinence de l'index produit.

Nous nous intéressons maintenant sur l'apport de chacune des relations sémantiques utilisées, prise séparément. Pour l'évaluation de cette expansion, nous avons choisis plusieurs échelle de précision (précision pour les 10 (*prec 10*) ou 30 (*prec 30*) premiers documents considérés et la précision moyenne). Nous présentons les résultats de la précision moyenne (table 5.2).

relations séparées	prec 10	prec 30	**prec moy**
synonymie	0.35	0.25	0.33
hyperonymie	0.31	0.25	**0.45**
hyponymie	0.24	0.18	0.30
holonymie	0.21	0.16	0.27
méronymie	0.22	0.12	0.27

TABLE 4.5 – Apport des relations sémantiques prises séparément

Les résultats montrent que la relation d'holonymie apporte moins que sa relation inverse la méronymie. Les relations qui apportent vraiment une amélioration de la précision moyenne sont la relation d'hyponymie et l'hyperonymie ainsi que la syno-

atrophie • cheville • **cheville>anatomie** • corps musculaire • coupe axiale transverse • **dégénérescence musculaire** • fibula • fracture • **fracture articulaire** • **fracture des membres inférieurs** • **fracture multiple** • **fracture diaphysaire** • **fracture du tibia** • **fracture non articulaire** • **fracture non déplacée** • **fracture spiroïde** • **fracture avec déplacement** • **fracture>lésion** • **imagerie médicale** • **jambe** • **lésion** • **lésion osseuse** • **loge>anatomie** • **médecine** • non articulaire • non déplacée • péroné • **radiologie** • **spiroïde** • tibia • trait de refend • **trauma** • **traumatisme des membres inférieurs**

FIGURE 4.8 – Index augmenté correspondant à la figure 4.4
Les termes en gras correspondent aux termes ajoutés alors que les autres correspondent aux termes de l'index brut, en n'autorisant **que les termes liés à médecine**.

nymie. La relation hyperonymie (c'est à dire la généralisation) permet d'améliorer la précision globale.

Nous pouvons aussi prendre en compte une combinaison de ces relations.

relations	prec 10	prec 30	**prec moy**
synonymie + hyperonymie+ hyponymie	0.35	0.29	**0.38**
hyperonymie + hyponymie	0.31	0.25	0.36
holonymie + méronymie	0.24	0.15	0.30
hyperonymie + hyponymie + holonymie + méronymie	0.30	0.20	0.34

TABLE 4.6 – Apport des relations sémantiques combinées

Nous constatons une amélioration des résultats chaque fois que l'hyperonymie est utilisée.

En comparant les deux index (avec et sans les termes généraux), nous pouvons constater que des termes non médicaux potentiellement utiles n'apparaissent plus dans l'index augmenté. Les termes généraux comme *accident de ski* ou *chute>tomber* peuvent faire l'objet de requêtes et par conséquence sont définis comme des termes pertinents.

Nous regardons aussi la propagation relation par relation pour obtenir une meilleure compréhension du bruit, du silence ou des erreurs et d'apporter une solution aux problèmes rencontrés. La relation idée associée peut engendrer beaucoup de bruit.

Nous remarquons que l'ensemble du processus présenté ci-dessus fonctionne de façon thématique sur le texte et sémantique sur le réseau lexical. Il n'y a pas d'analyse syntaxique fine, qui impliquerait selon toute vraisemblance une analyse en constituant et en dépendance des comptes rendus. Les cas d'erreur manifestes (23 termes pour les 200 index, soit environ 10 000 termes) peuvent avoir plusieurs causes :

— défaut d'information dans la base de connaissances (20% des cas d'erreur)
— défaut de rôle sémantique, impliquant la nécessité d'une analyse fine (55%)
— chimérisme [4] : deux parties distinctes du compte rendu ont fait émerger un terme non pertinent (25%).

Conclusion

Dans ce chapitre, nous avons démontré l'utilité d'indexer automatiquement des comptes rendus radiologiques, non seulement avec les termes médicaux mais aussi avec des termes du langage courant susceptibles d'être utilisés dans des requêtes d'utilisateurs, notamment les praticiens hospitaliers (*traumatisme abdominal* par *encornage*). Pour augmenter le rappel sans notablement dégrader la précision, nous ajoutons à l'index brut des termes implicites des comptes rendus, en utilisant comme support la base de connaissances qu'offre le réseau lexico-sémantique JeuxDeMots. Cette augmentation d'index est réalisée en amont de la requête et au moment de celle-ci. Du fait de l'existence d'un index augmenté, nous ne réalisons pas d'expansion de requêtes. A notre connaissance, très peu de travaux prennent en compte des éléments non médicaux présents dans le compte rendu, ou encore effectuent de l'inférence implicite afin de trouver des termes pertinents non présents. Les approches classiques d'augmentation du rappel consistent essentiellement à inclure des termes plus généraux (hyperonymes ou synonymes) à partir d'une ontologie médicale. La présence d'informations de sens commun améliore les résultats : l'hypothèse selon laquelle la non-séparation des connaissances (spécialisées et générales) est plus intéressante que l'usage exclusif de celles de spécialité semble se confirmer, ce que tendent à montrer nos résultats (augmentation de 12% de la précision).

4. apparition d'un terme provenant de deux parties distinctes d'un document

Chapitre 5

Extraction de relations sémantiques

Sommaire

5.1 Patrons lexicaux 129
5.1.1 Définition de patron lexical 129
5.1.2 Exemples de patrons lexicaux 130
5.2 Contraintes sur les patrons 134
5.2.1 Patrons sémantiques 134
5.2.2 Algorithme d'identification des relations 136
5.3 Expérimentation et résultats 137
5.3.1 Expérience 137
5.3.2 Résultats 137
5.4 Modèle PMA (Patient, Modalité, Affection) 141
5.4.1 Patient 142
5.4.2 Modalité 143
5.4.3 Affection 144

Dans ce chapitre, nous abordons l'extraction de relations lexico-sémantiques (cf définitions) à partir de comptes rendus de radiologie en français. Nous nous intéressons uniquement à la détection des relations sémantiques entre termes. La tâche d'extraction de relations sémantiques à partir de documents textuels a été abordée pour différents objectifs (questions-réponses [Abacha et Zweigenbaum, 2011], peuplement d'ontologie,...) et avec diverses techniques (déjà mentionné dans le chapitre 2.5) (règles, apprentissage supervisé ou semi-supervisé, etc).

Notre méthode repose sur des patrons lexicaux construit manuellement à partir d'un sous-ensemble de notre corpus de comptes rendus radiologiques. Étant donné que certains patrons lexicaux peuvent être trop généraux (*de, avec, sous, etc*), des contraintes sémantiques sur les relations, sous forme de règles, ont été ajoutées. Par exemple, dans le segment *fracture du tibia*, le système, déterminant *fracture* comme une *lésion* et *tibia* comme *lieu anatomique*, identifie la relation comme étant une relation de lieu : fracture r_lieu tibia, grâce à des contraintes sémantiques, bien que le patron lexical *du* soit général.

Définition : **Relation lexico-sémantique** :
Fonction de nature sémantique, c'est à dire qui associe les signifiés de deux entités lexico-sémantiques. Par exemple la synonymie, hyperonymie, hyponymie

Le chapitre commence par présenter les patrons lexicaux (section 5.1). Nous abordons ensuite les contraintes sur les patrons (section 5.2). Les expériences et les

résultats seront discutés dans la section 5.3. Enfin, nous aborderons l'extraction des variables PMA (section 5.4).

Les travaux d'extraction de relations présentés dans ce chapitre s'appuient sur les articles [Lafourcade et Ramadier, 2016], [Ramadier et Lafourcade, 2016].

5.1 Patrons lexicaux

Afin d'extraire des relations entre des termes à l'intérieur des documents, nous utilisons un ensemble des patrons lexicaux que nous déterminons à l'aide de médecins radiologues. Ci-dessous, nous présentons des exemples de patrons lexicaux ainsi que les types de relation sémantiques extraits.

5.1.1 Définition de patron lexical

Définitions : **Patron lexical** : Un patron lexical est représenté par une expression régulière décrivant un modèle de segment textuel où les termes cibles sont présents à des emplacements spécifiques en relation avec une relation lexicale.

Afin d'extraire les relations entre termes nous utilisons un ensemble de patrons lexicaux. Pour chaque type de relation sémantique, nous construisons des patrons et nous les comparons avec les phrases pour identifier la relation pertinente. Nous avons sélectionné 10 types de relations (table 5.1) à partir des conseils de radiologues. Les relations d'hyperonymie, d'équivalent sémantique et de synonymie n'apparaissent pas dans ce tableau car notre corpus ne contient pas ce type de relations. Le radiologue n'explicite pas cette relation dans ses comptes rendus (il sait qu'une *grippe* est une *maladie infectieuse* ou qu'un *AVC* est un *accident vasculaire cérébral*). De même, les abréviations ou acronymes (*AVC, ATCD, SA, ...*) ne sont pas explicitées par les médecins car ils connaissent leur signification. Ces relations seront rajoutées si nous appliquons cette méthode aux articles médicaux de wikipédia où les acronymes sont explicités de manière claire. Nous avons laissé la relation *traitement* même si elle n'apparaît pas fréquemment dans notre corpus contrairement à d'autres travaux spécifiquement centrés sur cette relation [Lee *et al.*, 2004]. Pour les relations

spécifiques au patient (sexe, âge, catégorie d'âge), nous en discutons dans la partie 5.5 de ce chapitre.

5.1.2 Exemples de patrons lexicaux

Les patrons sont crées manuellement avec l'aide des médecins à partir de notre corpus de comptes rendus radiologiques. Dans un premier temps, le nombre de patrons est limité à 50 patrons différents. Ce nombre correspond aux patrons les plus fréquents et les plus pertinents (selon les radiologues) présents les dans le corpus. Ce nombre sera élargi dans un futur travail en vu d'extraire davantage de relations.

types de relation *R*	**signification de A *R* B**
caractéristique	A a B comme caractéristique (adjectif) typique possible. Exemple : *carcinome hépatocellulaire* carac *hypervasculaire*
localisation	A a B comme lieu typique où peut se trouver le terme/objet en question. Exemple : *lobe caudé* localisation *foie*
cible	A a B comme population affectée par le terme. Exemple : *rougeole* cible *enfant*
holonymie	A a B comme tout. Exemple : *fémur* holo *membre inférieur*
partie de	A a B comme partie typique. Exemple : *fémur* has part *col du fémur*
signe	A a B pour symptôme/signe. Exemple : *grippe* symptôme *fièvre*
cause	B est une cause de A. Exemple : *cirrhose* cause *alcoolisme*
conséquence	B est une conséquence possible de A. Exemple : *accident vasculaire cérébral* conséquence *hémiplégie*
traitement	A a B comme traitement médical adapté. Exemple : *anévrisme cérébral* traitement *embolisation*
accompagnement	A est souvent accompagné par B. Exemple : *luxation* accompagné par *fracture*

TABLE 5.1 – Liste des relations à détecter.
Relations pertinentes selon les radiologues

Nous présentons (figure 5.1) certaines relations sémantiques ciblées.

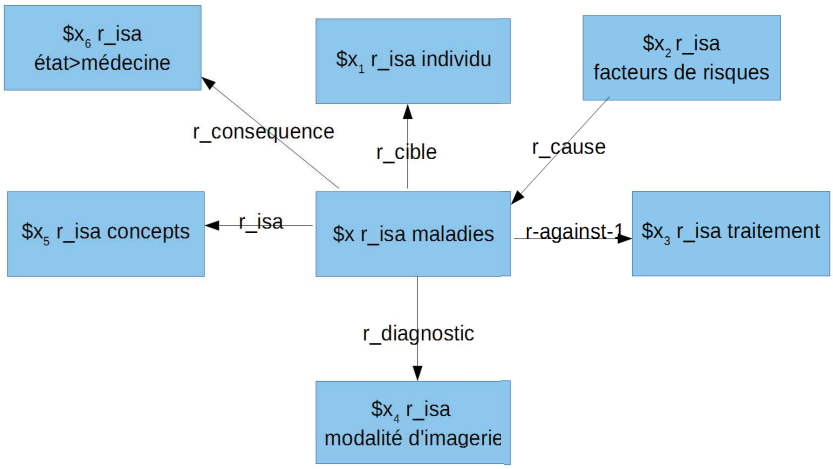


FIGURE 5.1 – Modélisation de certaines relations sémantiques utilisées.

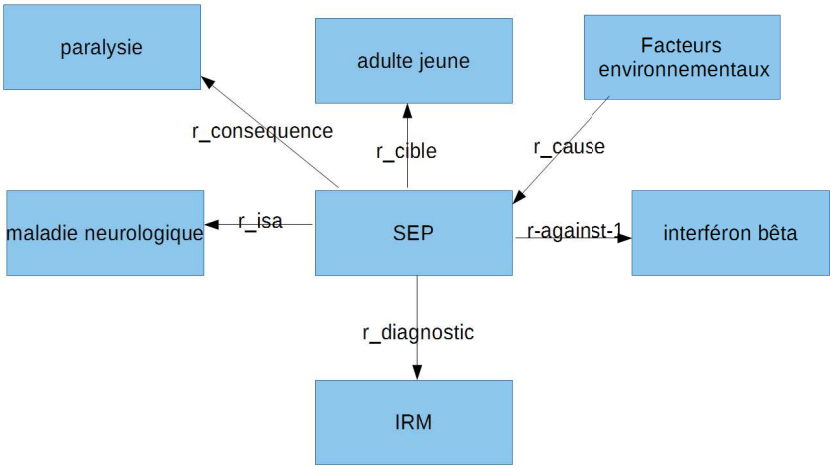


FIGURE 5.2 – Exemple de relations sémantiques utilisées.

Pour les patrons ainsi créés (table 5.2) plusieurs difficultés liées à l'ambiguïté lexicale

sont apparues. Par exemple, pour la *relation de localisation*, nous pouvons distinguer deux types de relations dépendant du patron. La première correspond réellement à la relation r_lieu (glioblastome *au niveau du* cerveau). La deuxième relation est l'holonymie. Un holonyme A d'un terme B est un terme dont le signifié désigne un ensemble comprenant le signifié de B (*ulna* r_holo *membre supérieur*). Nous avons remarqué que pour certains connecteurs (*du* dans *lobe caudé du foie*), les deux relations (*r_lieu et r_holo*) sont correctes (*lobe caudé* r_lieu *foie* et *lobe caudé* r_holo *foie*). Dans le cadre général de la médecine, la relation de localisation qui unit deux termes qui sont des lieux anatomiques (relation d'hyperonymie) pourra être à la fois une relation de lieu et aussi une relation d'holonymie (les deux relations peuvent se confondre). Par contre, dans le cas où un des deux termes est une maladie, un processus pathologique (par exemple, *tumeur du cerveau*, *fracture du col du fémur*), seule la relation de lieu sera correcte.

Nous avons également pris en compte la cooccurrence entre entités (Nom + Adjectif) pour la détermination de la relation sémantique *caractéristique* : *Nom* r_carac *Adjectif*.

types de relation *R*	**exemples de patrons lexicaux pour R**
caractéristique	$x est caractérisé par * $y ; $x d'allure $y ;
localisation	$x au niveau de * $y ; $x se trouve dans * $y ; $x développé aux dépens de * $y ; $x à la jonction de * $y ; $x situé sur le * $y
cible	$x n'affecte que les * $y ; $x chez $y
holonymie	$x fait partie de * $y
partie de	$x a comme partie *$y ; $x se compose de * $y
signe	$x se manisfeste par * $y
cause	$x déclenchant * $y ; $x peut produire * $y
conséquence	$x provoque * $y ; $x menant à $y
traitement	$x traité par * $y
accompagnement *	$x associé à des * $y ; $x s'accompagne d'un $y ; compliqué(e) d'un

TABLE 5.2 – Exemples de patrons lexicaux

L'astérisque * signifie que le patron peut prendre différentes formes. Par exemple, le patron lexical *fait partie* de peut se trouver sous la forme *fait partie de la*, *fait partie du* ou encore *fait partie des*. Nous n'utilisons pas de lemmatiseur avant d'appliquer le patron pour éviter d'avoir une application trop lourde. Cela nous permet d'avoir un système simple et efficace. Les termes composés sont déterminés selon la méthode expliquée dans le paragraphe 4.1.2

- r_lieu (lieu) => lésion *au niveau du* pancréas
- r_lieu (lieu) => lésion méniscale *au niveau du* ménisque latéral
- r_lieu(lieu) => diverticule *situé sur le* versant mésentérique
- r_against-1 (traitement) => CHC *traité par* sorafénib
- r_has_part (partie) => la lésion *se compose de* kystes
- r_target (symptome) => tumeur *se manifeste par une* rectorragie
- r_carac (caractéristique) => masse tissulaire *d'allure* tumorale
- r_accomp (accompagné) => dissection de type A *compliquée d'un* hémopéricarde

Pour certains patrons lexicaux, il a été particulièrement difficile de déterminer précisément le type de relation car le connecteur utilisé est très général (les connecteurs *du* ou *avec* peuvent être associés à plusieurs types de relations). Par exemple, pour le connecteur *du* il peut être lié à une relation de lieu dans l'expression *tumeur du foie* ou bien une relation de type *partie de* (part_of) dans *ventricule du crâne*. De même, le connecteur *avec* peut correspondre à une relation *symptôme* entre *luxation* avec *douleur* et une relation de cible (*target*) dans l'expression *nourrisson* avec *rougeole*. Pour surmonter cette difficulté et réduire le bruit, nous avons ajouté des contraintes d'ordre sémantique sur les termes entourant les patrons lexicaux. Ces contraintes ont été établies grâce aux connaissances contenues dans le réseau JDM.

Voici quelques exemples de connecteurs ambigus qui peuvent être rattachés à plusieurs relations sémantiques :

- luxation *avec* douleur (*r_symptôme*)
- nourrisson *avec* rougeole (*r_cible*)
- cathétériser *avec* une sonde 5f (*r_instrument*)
- sténose *du* tronc coeliaque (*r_lieu*)
- lobe inférieur *du* poumon (r_part_of)

— mort subite *du* nourrisson (*r_ cible*)

5.2 Contraintes sur les patrons

Pour améliorer la précision de l'extraction de relations sémantiques, il est nécessaire de désambiguïser certains connecteurs (*avec, du, de, etc*). Nous expliquons l'ajout de contraintes ainsi que l'algorithme d'extraction de relation à l'aide de patrons sémantiques.

5.2.1 Patrons sémantiques

Le connecteur *avec* peut correspondre à plusieurs relations. Par exemple relation de signe (symptôme) dans l'expression *luxation* avec *douleur* et relation de cible (target) dans *nourrisson* avec *rougeole*). D'autres connecteurs présentent les mêmes similitudes (*de, du, en, etc*). Pour contourner les difficultés liées à la généricité de ces connecteurs, nous ajoutons des contraintes d'ordre sémantique. Ces contraintes sont exprimées à travers une série de règles. Nous en présentons quelques exemples ci-dessous :

— $x **du** $y :
si $x r_isa lieu_anatomique & $y r_isa lieu_anatomique => $x r_holo $y
ligament croisé **du** genou => ligament croisé *r_ holo* genou
lobe caudé du foie => lobe caudé *r_ holo* foie
si $x r_isa maladie & $y r_isa lieu_anatomique => $x r_lieu $y
tumeur **du** foie => tumeur *r_ lieu* foie
adénocarcinome **du** côlon => adénocarcinome *r_ lieu* côlon

— $x **en** $y
si $x r_isa maladie & $y r_isa lieu_anatomique => $x r_lieu $y
fracture **en** C2 => fracture *r_ lieu* C2

— $x **avec** $y :
si $x r_isa maladie & $y r_isa signe_clinique => $x r_sign $y
occlusion avec douleur => occlusion *r_ symtom* douleur

si $x r_isa individu & $y r_isa maladie => $y r_cible $x
patient OH **avec** cirrhose => cirrhose r_cible patient OH

— $x **au niveau du** $y
si $x r_isa maladie & $y r_isa lieu_anatomique => $x r_lieu $y
masse **au niveau de la** tête du pancréas => lésion *r_lieu* tête du pancréas
si $x r_isa maladie & $y r_isa modalité_d'imagerie => $x r_diagnostic $y
SEP **au niveau de** l'IRM => SEP r_diagnostic IRM

— $x **due à** $y
si $x r_isa maladie & $y r_isa microorganisme => $x r_cause $y
tuberculose **due au** bacille de Koch => tuberculose r_cause bacille de Koch

Les règles de contraintes ont pour l'instant été établies manuellement et validées par des experts en imagerie médicale. Pour certains connecteurs et certaines relations il a été impossible de trouver des contraintes générales. Dans le segment textuel suivant : *Hématome extra dural* ***sur*** *blessure par arme à feu* la relation entre *hématome extra dural* et *blessure par arme à feu* est une relation de cause, mais il est très difficile de trouver des contraintes générales. D'autres exemples présentant des difficultés :

— *CHC* **sur** *terrain alcoolique.*
— Le tableau scanographique est celui d'une cholécystite aiguë gangrèneuse, **sur** vésicule biliaire lithiasique dans un contexte de diabète.

D'autre part dans le segment textuel suivant *hypersignal en T2*, il fut difficile de déterminer une relation entre *hypersignal* et *T2* (T2, dans cet exemple, est un contraste en imagerie par résonance magnétique). Dans ce cadre là, *hypersignal* en *T2* est rentré dans la base JDM.

Il serait désirable d'automatiser cette tâche d'extraction. Nous pourrrons utiliser la connaissance entre les termes contenue dans le réseau lexical JDM pour inférer les relations sémantiques portées par des patrons linguistiques récurrent identifiés automatiquement.

5.2.2 Algorithme d'identification des relations

A partir d'un corpus de documents (dans notre cas un corpus de comptes rendus radiologiques), nous appliquons une procédure d'extraction de relations présentée ci-dessous (algoritme 5) :

algorithme 5 extractionDeRelations

Entrée : phrases ou syngtames nominaux
Sortie : ensembles des relations sémantiques extraites
Soit **S** l'ensemble des résultats, qui sera vide à l'initialisation
Trouver des occurences des patrons dans le texte en prenant une fenêtre de mots de taille n
pour (chaque occurrence trouvée, on applique les contraintes aux variables) **faire**
  si (les contraintes sémantiques sont vérifiées) **alors**
    la relation est associée à x et y, donc ajoutée à S
  fin si
fin pour
return S

La valeur de n est la longueur du plus long patron (y compris les deux variables). Les termes composés ont été déterminés en amont (voir algorithme 1 du paragraphe 4.2.1) de l'extraction de relations. L'ensemble résultat S est pondéré, c'est-à-dire que le poids représente le nombre de fois qu'une relation sémantique donnée entre deux termes est trouvée dans le texte.

A partir des syntagmes textuels ci-dessous, nous extrayons :

contusion intra-osseuse de l'os sous-chondral en zone portante du condyle latéral
=> contusion intra-osseuse *r_location* os sous-chondral
=> os sous-chondral *r_has_part* condyle latéral

Hématome pariétal aortique de l'aorte ascendante
=> Hématome pariétal aortique *r_lieu* aorte descendante
Dissection aortique type A avec signes de fissuration
=> Dissection aortique type A *r_accomp* signes de fissuration

5.3 Expérimentation et résultats

Nous présentons les différentes expériences réalisées (avec et sans contraintes) ainsi que les résultats.

5.3.1 Expérience

Pour évaluer les performances de notre système, nous utilisons les mesures classiques, à savoir la précision (P), le rappel (R) (table 5.3 et table 5.4), et la F-mesure. A partir de notre corpus, nous avons extrait 120 000 relations selon la méthode de l'échantillonnage aléatoire simple (*simple random sampling*). Environ 800 de ces relations ont été annotées manuellement par un médecin et un spécialiste en imagerie médicale afin de déterminer le niveau de précision. Afin d'évaluer le rappel, nous identifions les relations dans 300 comptes rendus et appliquons ensuite notre algorithme à des fins de comparaison.

5.3.2 Résultats

Nous présentons les résultats de nos différentes expériences selon les différentes modalités avec et sans contraintes.

sans contraintes sémantiques			
types de relations	**précision**	**rappel**	**F1-mesure**
cause	74%	60%	66%
consequence	70%	62%	63.4%
location	48%	40%	43.6%
traitement	70%	45%	54.7%
partie de	32%	30%	31%
cible	45%	40%	42.4%
caractéristique	60%	58%	50%
lieu	45%	39%	41.7%

TABLE 5.3 – Résultats de l'extraction de relations sémantiques avec patrons linguistiques **sans** contraintes sémantiques

avec contraintes sémantiques			
types de relations	**précisions**	**rappel**	**F1-mesure**
cause	90%	60%	72%
consequence	89%	62%	73%
location	83%	40%	54%
traitement	97%	45%	61.4%
partie de	75%	30%	42.9%
cible	80%	40%	42.4%
caractéristique	88%	58%	70%
lieu	86%	39%	54.6%

TABLE 5.4 – Résultats de l'extraction de relations sémantiques avec patrons linguistiques **avec** contraintes sémantiques

La comparaison de nos résultat avec l'état de l'art (table5.5) doit tenir compte de la spécificité de notre corpus (comptes rendus quotidiens et non des articles scientifiques) et du type de relation sémantique ainsi que des méthodes de construction des patrons lexicaux. Dans d'autres travaux, les patrons sont construit de manière semi-automatique [Abacha et Zweigenbaum, 2011] ou de manière automatique [Embarek et Ferret, 2008]. Nous comparons nos résultats pour la relation traitement (utilisé fréquemment dans l'état de l'art) en soulignant le fait que cette relation est peu présente dans notre corpus. En effet, dans les comptes rendus, il y a souvent énumération des traitements et non une relation maladie-traitement. Par ailleurs, dans les comptes rendus , le style est souvent dégradé aussi bien au niveau lexical que dans la construction de segments textuels. Par exemple, le système n'extrait pas la relation r_againt-1 (traitement) dans le segment textuel suivant : *IDM stenté*. Le système a reconnu *stenté* comme une caractéristique (ce qui est correct) mais pas comme un traitement. En effet, la phrase pourrait être formulée de la façon suivante : *IDM traité par stent*, où dans ce cas le patron *traité par* est beaucoup plus explicite et permet de reconnaître le relation *traitement*. Pour cela , il faudrait différents traitements linguistiques (racinisation, par exemple). De plus, dans cet exemple le verbe *stenter* est une création lexicale. La précision est élevée mais la faible valeur du rappel de notre système s'explique par le fait que souvent les relations ne sont pas formulées explicitement comme expliqué ci-dessus.

comparaison			
système	**Abacha** [1]	**Embarrek** [2]	**IMAIOS**
précision	95.55%	96%	97%
rappel	32.45%	49%	45%

TABLE 5.5 – Comparaisons des résulats pour la relation *traitement*

Les différentes relations trouvées vont être à la fois indexées et aussi servir à alimenter la base de connaissance (figure 5.3).

micro-calcul au sein du canal cystique
micro-calcul ***r_lieu*** *canal cystique*

L'exemple ci-dessus est mis dans l'index car il est susceptible de faire l'objet de requêtes. De plus, nous vérifions si cette relation est déjà présente dans le réseau JDM et dans le cas contraire nous le proposons à un validateur (procédure définie dans le paragraphe 3.1.2).

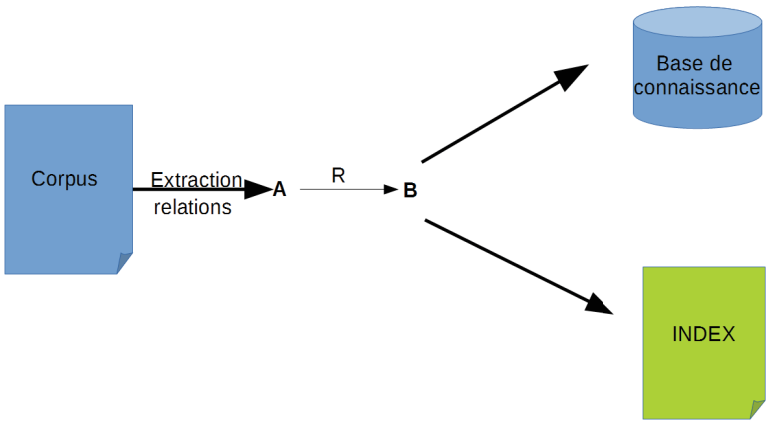


FIGURE 5.3 – Devenir des relations extraites.
Les relations sont à la fois mises dans l'index et proposées pour validation afin de les ajouter dans la base JDM.

Globalement, la mesure de la précision s'améliore de façon non négligeable quand nous ajoutons des contraintes sur les relations sémantiques. C'est surtout vrai pour la relation de lieu. En effet, la précision, sans contraintes sémantiques, est relativement faible, cela est dû au fait que beaucoup de connecteurs sont très généraux (*de, du*). Ces résultats s'expliquent par le fait que l'ajout de contraintes apporte une meilleure caractérisation de la relation (ce qui entraîne une amélioration de la précision) alors que le nombre de relations extraites ne varie pas. Par conséquent, le rappel reste constant. Peut être qu'il serait possible d'améliorer la précision en jouant sur le nombre de contraintes (en augmentant leur nombre) ou leur qualité.

Nous avons aussi appliqué notre méthode à d'autres types de corpus. Pour un corpus de 45 000 recettes de cuisine, nous avons extrait 245 000 relations avec une précision de 95%. L'évaluation a été réalisée de façon manuelle sur un échantillon de 755 relations. De plus, nous avons extrait 789 000 relations à partir de pages wikipedia avec une précision de 92% qui ont été évaluées à partir d'un échantillon de 1250 relations. Par rapport au corpus médical, nous avons rajouté les relations d'hyperonymie et de synonymie. L'extraction de la relation d'hyperonymie sur wikipedia a une précision de 94%. Les résultats sont présentés dans le tableau (table 5.6).

corpus	**précision**	**rappel**
médical (radiologie)	84.8%	50%
recette de cuisine	95%	49%
wikipedia (général)	92%	42%

TABLE 5.6 – Résultat de l'extraction sur différents corpus

Quelques précautions sont à prendre dans l'interprétation de ces résultats. Le nombre et le type de relations ne sont pas identiques entre le corpus médical et les deux autres corpus. En effet, comme mentionné précédemment, les radiologues n'explicitent pas les relations de types hyperonymie ou synonymie. De plus les types de relations ne sont pas les mêmes. Par exemple, la relation traitement (r_against-1) n'est pas utilisée dans les corpus de recettes de cuisine et de wikipedia. Le type de relation extraite dépend fortement de la nature du corpus. En effet, dans le corpus de recette, les relations *partie de* (*r_has_part*) (c'est-à-dire les ingrédients), instrument (*r_instr*) (ustensiles utilisés pour la recette), etc seront prépondérantes dans ce corpus. En ce qui concerne le corpus de l'encyclopédie en ligne Wikipedia, tous

les types de relations sont susceptibles d'être extraites. Le rappel de 42% bien que faible n'est pas véritablement un problème dans la mesure où l'on ne cherche pas à indexer les articles wikipedia, mais en extraire la connaissance la plus précise.

5.4 Modèle PMA (Patient, Modalité, Affection)

Dans la première partie de ce chapitre, nous avons étudié l'extraction de relations à l'intérieur de phrases. Dans cette partie, nous allons extraire des informations dont les relations ne sont pas situées aux mêmes endroits dans le document. Par exemple , la relation *cible* entre la maladie et le patient (*hépatocarcinome r_ cible patient alcoolique*) ne se situe pas dans la même phrase ni dans le même paragraphe. De même certains examens d'imagerie médicales ont des liens importants avec des pathologies (*sclérose en plaques r_ diagnostic IRM*) et dans ce cas-là aussi la relation ne se situe ni au niveau de la phrase ni au niveau du paragraphe. Dans un but de découverte de nouvelles relations ou de nouvelles connaissances, il est important et nécessaire d'extraire ces nouvelles relations. Dans cette perspective, nous proposons le modèle *PMA*. Dans ce modèle, nous extrayons des informations spécifiques liées aux patients (*sexe, âge, indications*), à la modalité d'imagerie médicale (*IRM, scanner, échographie, artériographie,etc*), et à l'affection qui correspond à la pathologie décelée par le praticien (*appendicite, rupture des ligaments croisés, ...*). Ces informations peuvent servir à trouver des corrélations entre, par exemple, le genre et la pathologie ou bien entre une cause et une maladie. Par exemple, si nous avons une relation entre *accident de ski* et *rupture des ligaments croisés* qui revient un certains nombre de fois (détermination d'un seuil) nous pouvons en déduire une certaine corrélation entre les deux entités. Cela pourra nous permettre, si nous disposons d'un corpus conséquent (un million voire plus de comptes rendus), de découvrir de nouvelles corrélations ou connaissances. Cela sera utile à la découverte de nouvelles connaissances et à une meilleure prise en charge de patients.

Femme : utérus, vagin, ovaire, trompe, annexe>anatomie, salpingite Homme : testicules, scrotum, pénis, prostate

FIGURE 5.4 – Exemple de mots clés pour découvrir le genre (homme ou femme) lorsque celui-ci n'est pas explicite.

Anamnèse Defense fig avec sepsis biologique sur notion d un situs inversus n'aurait qu'un **ovaire** a gauche et qu'une **trompe** gauche (coelioscopie il y a 7 ans) appendicite ? Résultats Malrotation intestinale avec répartition gauche du colon et droite du grêle, témoignant d'un mésentère commun complet. En fosse iliaque gauche, présence de multiples collections liquidiennes avec infiltration de la graisse périphérique et épanchements périphériques. L'ensemble étend situé à gauche de l'**utérus**, et à proximité du caecum. L'**annexe** gauche semble hypertrophié. C'est aspect est très en faveur d'une **salpingite** aiguë gauche. Diagnostic Pyosalpinx. Mésentère commun complet.

FIGURE 5.5 – Exemple de comptes rendus ou le genre du patient est déduit par les mots en gras.

5.4.1 Patient

Pour cet item, nous extrayons les mots clés liés à l'âge, au sexe et l'indication (indication de l'examen médical, le pourquoi ?). Pour l'extraction de ces données liées à des dates, nous utilisons le filtrage par motif (*Pattern matching*) [Cai *et al.*, 2016] qui est une technique utilisée dans la fouille de textes ainsi que dans différentes tâches de TALN. Le filtrage par motif peut utiliser des expressions régulières ou des séquences de caractères (listes de mots clés) qui permettent de chercher des motifs spécifiques. Les expressions régulières peuvent servir à détecter les dates en particulier les dates de naissances et les dates d'examen pour déterminer l'âge du patient. De plus, il arrive que le genre (féminin et masculin) d'un patient n'est pas indiqué explicitement mais peut être déduit par certains mots clés (utérus, testicules, seins, etc). Dans le réseau JDM, ces termes sont reliés soit à femme soit à homme par les relations sémantiques *r_target (cible)* ou *r_holo (fait partie de)*.

Le compte rendu (figure 5.5) est un compte rendu brut sans correction orthographique. Les fautes sont d'origine (par exemple, *a sans accent*). Il est important de

noter que la pathologie de l'indication peut être différente de celle de la conclusion. Par exemple, l'indication peut indiquer *appendicite* alors que la conclusion décrit une *pyosalpinx et mésentère commun complet*. Cette distinction sera présente dans l'index par l'inclusion devant l'affection de l'indication *anamnèse* et *conclusion*. Pour l'évaluation de notre méthode, nous le comparons avec un système médical spécialisé [Fiszman *et al.*, 2000]. Nous comparons leurs résultats avec un système d'extraction (SymText [Koehler, 1998]). Leurs résultats servirons de baseline pour évaluer notre méthode sur la variable affection. En effet, Nous n'avons pas utilisé de gold standard pour les autres variables car elles n'étaient pas extraites dans leur système.

méthode	**précisions**	**rappel**
notre méthode	85%	82%

TABLE 5.7 – Résultats de l'extraction de la variable *Patient*

5.4.2 Modalité

Dans cette partie, nous nous intéressons à la modalité utilisée pour une pathologie donnée (figure 5.6). Par exemple, pour une lésion des *ligaments croisés du genou*, la modalité IRM revient dans 95% des cas. Nous pouvons conclure que l'IRM est sûrement la modalité de choix pour une *rupture des croisés*.

Si P est la taille du texte et T la taille de la liste, alors la complexité $\Theta(P * T)$. Cet algorithme fonctionne si la taille des textes est relativement petite, ce qui est le cas avec nos documents.

méthode	**précisions**	**rappel**
notre méthode	90%	85%

TABLE 5.8 – Résultats de notre méthode pour la variable *modalité*

Les résultats montrent une meilleure précision que pour la variable *Patients* car les termes liés aux différentes modalités sont plus spécifiques. Par exemple pour la modalité scanner, nous avons les termes *UH, hyperdensité, triphasique, injection d'iode, phase portale, artérielle, etc* qui sont liés aux termes scanner ou tomodensitométrie.

5.4.3 Affection

L'affection correspond à un processus pathologique présent dans la conclusion du rapport. Nous réalisons cette extraction par une correspondance entre le texte de la conclusion et le réseau JDM. Si une correspondance est trouvé pour une maladie, alors nous ajoutons dans l'index à côté de la maladie *Conclusion*.

méthode	**précisions**	**rappel**
notre méthode	96%	88%
baseline	78%	95%

TABLE 5.9 – Résultats de notre méthode pour la variable affection

Les résultats obtenus (table 5.9) pour cette variable sont meilleurs que pour les autres variables du fait qu'il y a moins de variation syntaxique ou lexicales. Souvent dans la conclusion, le nom de la maladie est indiquée avec un adverbe qui précise le degré de confiance. La meilleure précision s'explique par le fait que la variable Affection est extraite , dans notre système, dans la section *conclusion* alors que dans le système *SymText*, elle est extraite dans la section indication où il existe une plus grande variabilité .

La modalité n'est pas directement citée dans le texte (*scanner*) mais elle est déduite grâce à la base de connaissance. En effet *UH (unité Hounsfield)* est relié dans le réseau à *scanner*.

UH r_ equiv unité Hounsfield et unité Hounsfield r_ associated tomodensitométrie.

Une autre difficulté fut de déterminer la pathologie pour laquelle l'examen est requis. Dans l'exemple ci-dessus, le patient a plusieurs pathologie connue mais l'indication ne concerne que l'abcès cérébral. Dans la partie indication du document, il est indiqué l'histoire médicale du patient et par conséquent, il est souvent indiqué les pathologies présentes ou passées du patient. Si dans la partie anamnèse, plusieurs maladies sont évoqués, nous distinguons les pathologies connues et l'indication réelle de l'examen en regardant le contexte du terme exprimant une maladie. L'indication apparaît souvent précédé du terme "*recherche de*" ou bien d'un point d'interrogation

Anamnèse
Crise convulsive tonicoclonique chez un patient non épileptique connu suivi pour septicémie sur abcès du psoas et spondylodiscite OH chronique / encéphalopathie hépatique connu Recherche abcès cérébral.
Résultats
Volumineuse lésion (43 x 30 x 28 mm) frontale gauche, **hypodense** liquidienne (12**UH**), se rehaussant en périphérie avec important oedème **hypodense** péri-lésionnel. L'aspect est compatible avec le diagnostic suspecté d'abcès dans le contexte. Effet de masse sur le système ventriculaire avec début d'engagement sous-falcoriel et écrasement du ventricule latéral gauche. Absence d'engagement temporal. Pas d'argument en faveur d'une hémorragie intra-crânienne récente. Pas de collection péri-encéphalique. Pas d'anomalie à l'étage sous-tentoriel. Principaux sinus veineux perméables, sans image de thrombus. Pas d'image anévrysmale évidente sur le segment intra-crânien des TSA et le polygone de Willis. Pas de comblement des sinus de la face, des cavités tympaniques et des mastoïdes.
Au total :
Lésion liquidienne frontale gauche avec oedème péri-lésionnel marqué et effet de masse ventriculaire qui est compatible avec l'hypothèse d'un abcès à pyogène dans le contexte.
Conclusion
Abcès cérébral

FIGURE 5.6 – Compte rendu original. Les termes en gras permettent de déduire la modalité (scanner dans cet exemple)

Patient r_isa OH chronique,
encéphalopathie hépatique r_target **P**atient
abcès du psoas r_target **P**atient
abcès cérébral r_target **P**atient (indication)
r_target Patients (indication)
Modalité *r_isa* scanner cérébral
Affection *r_isa* abcès cérébral (conclusion)

FIGURE 5.7 – Extraction des motifs *PMA*

(?). Un autre cas est celui où dans la partie indication n'apparaît pas de noms de maladie (voir annexe B) mais seulement de symptômes. Nous mettons, dans ce cas, ces indications dans la variable *Patients*.

Conclusion

Nous avons présenté dans ce chapitre une approche basée sur des patrons sémantiques pour l'extraction de relations sémantiques. Cette approche se fonde d'une part sur des patrons lexicaux auxquels nous avons rajouté des contraintes sémantiques. Nous avons montré que cet ajout de contraintes entraine une amélioration de la précision sans avoir recours à un étiqueteur morpho-syntaxique ou un analyseur syntaxique. Une piste future est d'améliorer la couverture de notre système par la découverte automatique de patrons lexicaux [Meng et Morioka, 2015]. Dans un deuxièmement temps, nous avons extrait des relations non plus à l'intérieur de phrases mais dans le document. Nous avons appelé ce système PMA, qui permet de relier des informations liées à la variable *Patients* à l'*Affection* ou à la *Modalité*. Ce système a pour l'instant comme but principal d'augmenter la base de connaissances. Nous envisageons principalement trois perspectives à ce travail. Tout d'abord, il faudra améliorer l'extraction de relations sémantique y compris le modèle PMA (obtenir une meilleure précision) en proposant une méthode semi-automatique. Ensuite, nous pourrons élargir le corpus à d'autres spécialités médicales voire à d'autres domaines de spécialité (droit, finance). Enfin, nous envisageons d'ajouter des règles en vue de développer un système d'aide au diagnostic basé sur le raisonnement.

Conclusion générale

Synthèse des travaux

Les travaux présentés ici se situent dans le contexte de l'informatisation des données médicales en particulier dans le domaine de l'imagerie et plus concrètement dans le cadre de l'indexation de documents médicaux (radiologiques). La base de connaissances (*JDM*), utilisée pour cette tâche, contient à la fois des connaissances de sens commun et spécialisées. Cette ressource sémantique contient un ensemble de termes et de relations caractérisant des liens sémantiques entre ces termes. Le but est d'exploiter la structure de cette base (relations sémantiques) pour une meilleure représentation de l'information (analyse des comptes rendus) et ainsi améliorer la correspondance entre le besoin de l'utilisateur et l'information. Dans cette perspective, nous avons présenté principalement **trois contributions**. Après avoir développé à l'intérieur de la base de connaissances générales (le réseau lexico-sémantique JeuxDeMots) une base de spécialité (domaine de l'imagerie médicale), nous avons ajouté des annotations aux relations entre termes pour ajouter des méta-informations servant à analyser de façon plus précise les comptes rendus. L'ajout des informations qualitatives, quantitatives, fréquentielles, etc peut permettre de réordonner les documents retournés après une requête suivant différents facteurs (selon qu'une caractéristique soit fréquente ou rare). Ces annotations permettent envisager de simplifier lexicalement les comptes rendus afin que ces derniers soient compréhensibles par des non-experts (par exemple, les patients). Dans ce cadre là, il faudrait aussi envisager une simplification syntaxique pour rendre les segments textuels ou les phrases plus

simples.

La deuxième apport a été de pouvoir construire des index augmentés à partir des index bruts grâce à un algorithme de propagation à travers le réseau lexico-sémantique. Cette propagation a été réalisée non seulement sur les relations de synonymie, hyperonymie mais également d'autres relations (caractéristiques, cause, conséquence, etc). Dans cette approche de l'indexation, nous avons pris en compte non seulement les termes médicaux (maladies, anatomie, symptômes) mais aussi des termes non médicaux présents dans les documents (surtout dans la partie *indications* des comptes rendus). En effet, les utilisateurs (c'est-à-dire, les radiologues ou les internes en radiologie) peuvent être amené à formuler des requêtes avec des termes du langage courant. Les différentes évaluations de cette augmentation d'index, ont montré une amélioration des résultats (augmentation d'environ 12% de la précision). Il semblerait que l'utilisation d'une base de connaissances non limitée au domaine de spécialité améliore la pertinence de l'index produit par rapport à celui produit qu'avec. La présence d'informations de sens commun améliore les résultats : l'hypothèse selon laquelle la **non séparation des connaissances** (spécialisées et générales) est plus intéressante que l'usage exclusif de celles de spécialité semble se confirmer au vu des résultats. La partie PMA (extraction de mots clés) poursuit comme but de découvrir dans les comptes rendus de nouvelles connaissances de type *cible-pathologie* et permet aussi d'alimenter le réseau lexical.

La troisième contribution réalisée dans le cadre de ces travaux, concerne l'extraction de relations sémantiques. Notre méthode d'extraction est basée sur l'utilisation de patrons lexicaux avec contraintes sémantiques qui peuvent être vérifiées grâce au réseau lexical JDM. Les résultats expérimentaux montrent que l'ajout de ces contraintes améliore de façon significative la précision, sans avoir recours à un analyseur syntaxique ou à un étiquetage morphosyntaxique. Cette méthode montre aussi des résultats intéressants sur d'autres corpus permettant de valider notre méthode non seulement pour un corpus de spécialité (médecine, cuisine) mais aussi pour des corpus plus généraux comme avec l'encyclopédie en ligne Wikipedia.

Enfin, un prototype de moteur de recherche a été réalisé dans le cadre de ces travaux (Annexe E). Ce moteur de recherche permet aux praticiens d'écrire une requête avec des termes non présents dans les comptes rendus. En collaboration avec la société IMAIOS, un des objectifs de cette recherche sémantique est de la combiner avec

la recherche de contenu par l'image (CBIR) pour améliorer la recherche d'image médicale dans un cadre clinique.

Perspectives

Le réseau est toujours en construction et nous complétons avec des spécialités médicales non radiologiques en vue d'élargir l'application de nos travaux à d'autres comptes rendus (rapport d'hospitalisation, compte rendus opératoire, rapport anatomopathologie, ...). La couverture des annotations est elle aussi en perpétuelle évolution (de nouvelles valeurs annotations sont rajoutées à la demande de praticiens).

Afin d'améliorer la qualité de réseau, nous pensons élargir l'extraction de relations sémantiques à différents textes médicaux comme par exemple *Orphanet*, *Dictionnaire médical de l'Académie de Médecine*. Nous envisageons de lier les termes médicaux aux ontologies telle que l'UMLS et Radlex.

Une piste future de notre travail est une meilleure couverture de notre système par la découverte/détection de patrons linguistiques de façon automatique. Nous souhaitons implémenter une méthode automatique ou semi-automatique de détection de contraintes. Cela permettra de créer de nouvelles contraintes afin d'améliorer la précision de notre système. Dans cette configuration, le rappel diminuera car les contraintes seront moins contrôlées.

Un autre objectif est aussi de découvrir dans les comptes rendus de nouvelles connaissances permettant d'alimenter le réseau lexical. Nous envisageons également de déduire à partir du corpus des règles d'inférence et de faire ainsi un raisonnement authentique, c'est-à-dire de proposer par déduction et induction de nouvelles informations médicales, voire des diagnostics.

Bibliographie

[Abacha et Zweigenbaum, 2011] Abacha, A. B. et Zweigenbaum, P. (2011). Automatic extraction of semantic relations between medical entities : a rule based approach. *Journal of biomedical semantics*, 2(5):1–11.

[Abeillé et Blache, 1997] Abeillé, A. et Blache, P. (1997). Etat de l'art : La syntaxe : Etat de l'art. *TAL. Traitement automatique des langues*, 38(2):69–90.

[Alvarez *et al.*, 2004] Alvarez, C., Langlais, P. et Nie, J.-Y. (2004). Word pairs in language modeling for information retrieval. *In Coupling approaches, coupling media and coupling languages for information retrieval*, pages 686–705. Le centre de hautes études internationales d'informatique documentaire.

[Aronson, 2001] Aronson, A. R. (2001). Effective mapping of biomedical text to the umls metathesaurus : the metamap program. *In Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

[Aronson *et al.*, 2004] Aronson, A. R., Mork, J. G., Gay, C. W., Humphrey, S. M. et Rogers, W. J. (2004). The nlm indexing initiative's medical text indexer. *Medinfo*, 11(Pt 1):268–72.

[Auger et Barrière, 2008] Auger, A. et Barrière, C. (2008). Pattern-based approaches to semantic relation extraction : A state-of-the-art. *Terminology*, 14(1):1–19.

[Aussenac-Gilles *et al.*, 2008] Aussenac-Gilles, N., Despres, S. et Szulman, S. (2008). The terminae method and platform for ontology engineering from texts. *Bridging the Gap between Text and Knowledge-Selected Contributions to Ontology Learning and Population from Text*, pages 199–223.

[Avillach *et al.*, 2007] Avillach, P., Joubert, M. et Fieschi, M. (2007). A model for indexing medical documents combining statistical and symbolic knowledge.

In AMIA Annual Symposium Proceedings, volume 2007, pages 31–35. American Medical Informatics Association.

[Baker *et al.*, 1998] Baker, C. F., Fillmore, C. J. et Lowe, J. B. (1998). The berkeley framenet project. *In Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

[Barrows Jr *et al.*, 2000] Barrows Jr, R. C., Busuioc, M. et Friedman, C. (2000). Limited parsing of notational text visit notes : ad-hoc vs. nlp approaches. *In Proceedings of the AMIA Symposium*, pages 51–55. American Medical Informatics Association.

[Bases, 1993] Bases, M. (1993). Automatic knowledge acquisition from medline. *Meth. Inform. Med*, 32(2):120–30.

[Baziz *et al.*, 2005] Baziz, M., Boughanem, M., Pasi, G. et Prade, H. (2005). A fuzzy set approach to concept-based information retrieval. *In EUSFLAT Conf.*, pages 1287–1292.

[Belew, 2000] Belew, R. K. (2000). *Finding out about : a cognitive perspective on search engine technology and the WWW*, volume 1. Cambridge University Press.

[Berry *et al.*, 1995] Berry, M. W., Dumais, S. T. et O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *SIAM review*, 37(4):573–595.

[Bhardwaj *et al.*, 2010] Bhardwaj, V., Passonneau, R. J., Salleb-Aouissi, A. et Ide, N. (2010). Anveshan : a framework for analysis of multiple annotators' labeling behavior. *In Proceedings of the Fourth Linguistic Annotation Workshop*, pages 47–55. Association for Computational Linguistics.

[Biemann, 2005] Biemann, C. (2005). Semantic indexing with typed terms using rapid annotation. *In Proceedings of the TKE-05-Workshop on Methods and Applications of Semantic Indexing, Copenhagen*, page 54.

[Bodenreider, 2004] Bodenreider, O. (2004). The unified medical language system (umls) : integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270.

[Bodenreider *et al.*, 2002] Bodenreider, O., Mitchell, J. A. et McCray, A. T. (2002). Evaluation of the umls as a terminology and knowledge resource for biomedical informatics. *In Proceedings of the AMIA Symposium*, pages 61–65. American Medical Informatics Association.

[Bordogna et Pasi, 2000] Bordogna, G. et Pasi, G. (2000). Modeling vagueness in information retrieval. *In Lectures on information retrieval*, pages 207–241. Springer.

[Bouaud *et al.*, 1996] Bouaud, J., Bachimont, B. et Zweigenbaum, P. (1996). Traitement de la m etonymie bas e sur un mod ele du domaine et sur une recherche heuristique de chemin dans des graphes. *Actes de TALN*, 96.

[Boubekeur, 2008] Boubekeur, F. (2008). *Contribution à la définition de modèles de recherche d'information flexibles basés sur les CP-Nets*. Thèse de doctorat, Université Paul Sabatier-Toulouse III.

[Boughanem *et al.*, 1999] Boughanem, M., Chrisment, C. et Soulé-Dupuy, C. (1999). Query modification based on relevance back-propagation in an ad hoc environment. *Information processing & management*, 35(2):121–139.

[Boughanem *et al.*, 2005] Boughanem, M., Loiseau, Y. et Prade, H. (2005). Rank-ordering documents according to their relevance in information retrieval using refinements of ordered-weighted aggregations. *In International Workshop on Adaptive Multimedia Retrieval*, pages 44–54. Springer.

[Bourigault *et al.*, 1996] Bourigault, D., Gonzalez-Mullier, I. et Gros, C. (1996). Lexter, a natural language processing tool for terminology extraction. *In Proceedings of the 7th EURALEX International Congress*, pages 771–779.

[Brachman, 1983] Brachman, R. J. (1983). What is-a is and isn't : An analysis of taxonomic links in semantic networks. *Computer ;(United States)*, 10.

[Brown *et al.*, 1992] Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D. et Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

[Budovec *et al.*, 2014] Budovec, J. J., Lam, C. A. et Kahn Jr, C. E. (2014). Informatics in radiology : Radiology gamuts ontology : Differential diagnosis for the semantic web. *Radiographics*, 34(1):254–264.

[Burgun et Bodenreider, 2001] Burgun, A. et Bodenreider, O. (2001). Comparing terms, concepts and semantic classes in wordnet and the unified medical language system. *In Proceedings of the NAACL'2001 Workshop,"WordNet and Other Lexical Resources : Applications, Extensions and Customizations*, pages 77–82.

[Cai *et al.*, 2016] Cai, T., Giannopoulos, A. A., Yu, S., Kelil, T., Ripley, B., Kumamaru, K. K., Rybicki, F. J. et Mitsouras, D. (2016). Natural language processing technologies in radiology research and clinical applications. *RadioGraphics*, 36(1):176–191.

[Cao *et al.*, 2008] Cao, G., Nie, J.-Y., Gao, J. et Robertson, S. (2008). Selecting good expansion terms for pseudo-relevance feedback. *In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 243–250. ACM.

[Charlet *et al.*, 2004] Charlet, J., Bachimont, B. et Troncy, R. (2004). Ontologies pour le web sémantique. *Revue I3, page 31p.*

[Chklovski et Pantel, 2004] Chklovski, T. et Pantel, P. (2004). Verbocean : Mining the web for fine-grained semantic verb relations. *In EMNLP*, volume 4, pages 33–40.

[Christensen et Grimsmo, 2008] Christensen, T. et Grimsmo, A. (2008). Instant availability of patient records, but diminished availability of patient information : a multi-method study of gp's use of electronic patient records. *BMC medical informatics and decision making*, 8(1):12.

[Cimino *et al.*, 1994] Cimino, J. J., Clayton, P. D., Hripcsak, G. et Johnson, S. B. (1994). Knowledge-based approaches to the maintenance of a large controlled medical terminology. *Journal of the American Medical Informatics Association*, 1(1):35.

[Claveau et Zweigenbaum, 2005] Claveau, V. et Zweigenbaum, P. (2005). Traduction de termes biomédicaux par inférence de transducteurs. *In Actes de la conférence Traitement automatique des langues naturelles, TALN'05.*

[Clinchant et Gaussier, 2010] Clinchant, S. et Gaussier, E. (2010). Information-based models for ad hoc ir. *In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 234–241. ACM.

[Crestani, 1994] Crestani, F. (1994). Comparing neural and probabilistic relevance feedback in an interactive information retrieval system. *In Neural Networks, 1994. IEEE World Congress on Computational Intelligence., 1994 IEEE International Conference on*, volume 5, pages 3426–3430. IEEE.

[Crestani, 1997] Crestani, F. (1997). Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, 11(6):453–482.

[Crestani et van Rijsbergen, 1998] Crestani, F. et van Rijsbergen, C. J. (1998). A study of probability kinematics in information retrieval. *ACM Transactions on Information Systems (TOIS)*, 16(3):225–255.

[Croft et Lafferty, 2013] Croft, B. et Lafferty, J. (2013). *Language modeling for information retrieval*, volume 13. Springer Science & Business Media.

[Dameron *et al.*, 2005] Dameron, O., Rubin, D. L. et Musen, M. A. (2005). Challenges in converting frame-based ontology into owl : the foundational model of anatomy case-study. *In AMIA*. Citeseer.

[Darmoni *et al.*, 2003] Darmoni, S. J., Jarrousse, E., Zweigenbaum, P., Le Beux, P., Namer, F., Baud, R., Joubert, M., Vallée, H., Côté, R. A., Buemi, A. *et al.* (2003). Vumef : extending the french involvement in the umls metathesaurus. *In AMIA Annual Symposium Proceedings*, volume 2003, page 824. American Medical Informatics Association.

[DeClaris *et al.*, 1994] DeClaris, N., Harman, D., Faloutsos, C., Dumais, S. et Oard, D. (1994). Information filtering and retrieval : Overview, issues and directions. *In Engineering in Medicine and Biology Society, 1994. Engineering Advances : New Opportunities for Biomedical Engineers. Proceedings of the 16th Annual International Conference of the IEEE*, pages A42–A49. IEEE.

[Deerwester *et al.*, 1990] Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W. et Harshman, R. A. (1990). Indexing by latent semantic analysis. *JAsIs*, 41(6):391–407.

[Doerr *et al.*, 2007] Doerr, M., Ore, C.-E. et Stead, S. (2007). The cidoc conceptual reference model : a new standard for knowledge sharing. *In Tutorials, posters, panels and industrial contributions at the 26th international conference on Conceptual modeling-Volume 83*, pages 51–56. Australian Computer Society, Inc.

[Dong et Dong, 2006] Dong, Z. et Dong, Q. (2006). *HowNet and the Computation of Meaning*. World Scientific.

[Donnelly, 2006] Donnelly, K. (2006). Snomed-ct : The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279.

[Dubois et Prade, 2012] Dubois, D. et Prade, H. (2012). *Possibility theory : an approach to computerized processing of uncertainty*. Springer Science, Business Media.

[Edinger *et al.*, 2012] Edinger, T., Cohen, A. M., Bedrick, S., Ambert, K. et Hersh, W. (2012). Barriers to retrieving patient information from electronic health record data : failure analysis from the trec medical records track. *In AMIA Annual Symposium Proceedings*, volume 2012, page 180. American Medical Informatics Association.

[Efthimiadis, 1996] Efthimiadis, E. N. (1996). Query expansion. *Annual review of information science and technology*, 31:121–187.

[Eiben *et al.*, 2012] Eiben, C. B., Siegel, J. B., Bale, J. B., Cooper, S., Khatib, F., Shen, B. W., Players, F., Stoddard, B. L., Popovic, Z. et Baker, D. (2012). Increased diels-alderase activity through backbone remodeling guided by foldit players. *Nature biotechnology*, 30(2):190–192.

[Embarek et Ferret, 2008] Embarek, M. et Ferret, O. (2008). Learning patterns for building resources about semantic relations in the medical domain. *In LREC*.

[Fader *et al.*, 2011] Fader, A., Soderland, S. et Etzioni, O. (2011). Identifying relations for open information extraction. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545. Association for Computational Linguistics.

[Fellbaum, 2005] Fellbaum, C. (2005). Wordnet and wordnets. *Encyclopedia of Language and Linguistics*.

[Fillmore, 1982] Fillmore, C. (1982). Frame semantics. *Linguistics in the morning calm*, pages 111–137.

[Fiorini *et al.*, 2014] Fiorini, N., Ranwez, S., Ranwez, V. et Montmain, J. (2014). Indexation conceptuelle par propagation. *In COnférence en Recherche d'Information et Applications*.

[Fiszman *et al.*, 2000] Fiszman, M., Chapman, W. W., Aronsky, D., Evans, R. S. et Haug, P. J. (2000). Automatic detection of acute bacterial pneumonia from chest x-ray reports. *Journal of the American Medical Informatics Association*, 7(6):593–604.

[Fonseca *et al.*, 2005] Fonseca, B. M., Golgher, P., Pôssas, B., Ribeiro-Neto, B. et Ziviani, N. (2005). Concept-based interactive query expansion. *In Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 696–703. ACM.

[Fort *et al.*, 2011] Fort, K., Adda, G. et Cohen, K. B. (2011). Amazon mechanical turk : Gold mine or coal mine ? *Computational Linguistics*, 37(2):413–420.

[Fox, 1992] Fox, M. S. (1992). The tove project towards a common-sense model of the enterprise. *In Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, pages 25–34. Springer.

[Gangemi *et al.*, 2002] Gangemi, A., Guarino, N., Masolo, C., Oltramari, A. et Schneider, L. (2002). Sweetening ontologies with dolce. *In Knowledge engineering and knowledge management : Ontologies and the semantic Web*, pages 166–181. Springer.

[Getman et Karasiuk, 2014] Getman, A. P. et Karasiuk, V. V. (2014). A crowdsourcing approach to building a legal ontology from text. *Artificial Intelligence and Law*, 22(3):313–335.

[Gillick et Liu, 2010] Gillick, D. et Liu, Y. (2010). Non-expert evaluation of summarization systems is risky. *In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 148–151. Association for Computational Linguistics.

[Girju *et al.*, 2003] Girju, R., Badulescu, A. et Moldovan, D. (2003). Learning semantic constraints for the automatic discovery of part-whole relations. *In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 1–8. Association for Computational Linguistics.

[Golbreich *et al.*, 2005] Golbreich, C., Zhang, S. et Bodenreider, O. (2005). Migrating the fma from protégé to owl. *In Proc. of the 8th International Protégé Conference*.

[Gonzalo *et al.*, 1998] Gonzalo, J., Verdejo, F., Chugur, I. et Cigarran, J. (1998). Indexing with wordnet synsets can improve text retrieval. *arXiv preprint cmp-lg/9808002*.

[Good et Su, 2013] Good, B. M. et Su, A. I. (2013). Crowdsourcing for bioinformatics. *Bioinformatics*, page btt333.

[Grabar, 2004] Grabar, N. (2004). *Terminologie médicale et morphologie : acquisition de ressources morphologiques et leur utilisation pour le traitement de la variation terminologique*. Thèse de doctorat, Paris 6.

[Grabar et Hamon, 2016] Grabar, N. et Hamon, T. (2016). A large rated lexicon with french medical words. *In* Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J. et Piperidis, S., éditeurs : *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

[Grossman et Frieder, 2012] Grossman, D. A. et Frieder, O. (2012). *Information retrieval : Algorithms and heuristics*, volume 15. Springer Science & Business Media.

[Gruber, 1995] Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing ? *International journal of human-computer studies*, 43(5):907–928.

[Guarino, 1997] Guarino, N. (1997). Semantic matching : Formal ontological distinctions for information organization, extraction, and integration. *In Information Extraction A Multidisciplinary Approach to an Emerging Information Technology*, pages 139–170. Springer.

[Guarino *et al.*, 1999] Guarino, N., Masolo, C. et Vetere, G. (1999). Ontoseek : Content-based access to the web. *Intelligent Systems and Their Applications, IEEE*, 14(3):70–80.

[Habert et Jacquemin, 1993] Habert, B. et Jacquemin, C. (1993). Noms composés, termes, dénominations complexes : problématiques linguistiques et traitements automatiques. *TAL. Traitement automatique des langues*, 34(2):119–138.

[Harman, 1993] Harman, D. K. (1993). The first text retrieval conference (trec-1) rockville, md, usa, 4–6 november, 1992. *Information Processing & Management*, 29(4):411–414.

[Harter, 1975] Harter, S. P. (1975). A probabilistic approach to automatic keyword indexing. part ii. an algorithm for probabilistic indexing. *Journal of the american society for information science*, 26(5):280–289.

[Hassell *et al.*, 2006] Hassell, J., Aleman-Meza, B. et Arpinar, I. B. (2006). Ontology-driven automatic entity disambiguation in unstructured text. *In International Semantic Web Conference*, pages 44–57. Springer.

[Hearst, 1992] Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. *In Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics.

[Hersh *et al.*, 2001] Hersh, W., Mailhot, M., Arnott-Smith, C. et Lowe, H. (2001). Selective automated indexing of findings and diagnoses in radiology reports. *Journal of biomedical informatics*, 34(4):262–273.

[Hiemstra et Robertson, 2001] Hiemstra, D. et Robertson, S. E. (2001). Relevance feedback for best match term weighting algorithms in information retrieval.

[Hirst et St-Onge, 1998] Hirst, G. et St-Onge, D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet : An electronic lexical database*, 305:305–332.

[Huang *et al.*, 2003] Huang, Y., Lowe, H. J. et Hersh, W. R. (2003). A pilot study of contextual umls indexing to improve the precision of concept-based representation in xml-structured clinical radiology reports. *Journal of the American Medical Informatics Association*, 10(6):580–587.

[Jacquemin, 1997] Jacquemin, C. (1997). Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus. *Mémoire d'Habilitation à Diriger des Recherches en informatique fondamentale, Université de Nantes*.

[Jacquemin *et al.*, 2002] Jacquemin, C., Daille, B., Royauté, J. et Polanco, X. (2002). In vitro evaluation of a program for machine-aided indexing. *Information processing & management*, 38(6):765–792.

[Järvelin *et al.*, 2001] Järvelin, K., Kekäläinen, J. et Niemi, T. (2001). Expansiontool : Concept-based query expansion and construction. *Information Retrieval*, 4(3-4):231–255.

[Jelinek, 1976] Jelinek, F. (1976). Speech recognition by statistical methods. *Proceedings of the IEEE*, 64:532–556.

[Jeon et Manmatha, 2004] Jeon, J. et Manmatha, R. (2004). Using maximum entropy for automatic image annotation. *In Image and Video Retrieval*, pages 24–32. Springer.

[Jones, 1987] Jones, K. S. (1987). Issues in user modelling for expert systems. *In Artificial intelligence and its applications*, pages 183–195. John Wiley & Sons, Inc.

[Katz *et al.*, 1998] Katz, B., Uzuner, O. et Yuret, D. (1998). Word sense disambiguation for information retrieval. Citeseer.

[Khan, 2000] Khan, L. R. (2000). *Ontology-based information selection*. Thèse de doctorat, University of Southern California.

[Kim *et al.*, 2001] Kim, W., Aronson, A. R. et Wilbur, W. J. (2001). Automatic mesh term assignment and quality assessment. *In Proceedings of the AMIA Symposium*, page 319. American Medical Informatics Association.

[Koehler, 1998] Koehler, S. B. (1998). *SymText : a natural language understanding system for encoding free text medical data*. The University of Utah.

[Krause *et al.*, 2010] Krause, M., Takhtamysheva, A., Wittstock, M. et Malaka, R. (2010). Frontiers of a paradigm : exploring human computation with digital games. *In Proceedings of the acm sigkdd workshop on human computation*, pages 22–25. ACM.

[Krovetz, 1997] Krovetz, R. (1997). Homonymy and polysemy in information retrieval. *In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 72–79. Association for Computational Linguistics.

[Kwok, 1989] Kwok, K. (1989). A neural network for probabilistic information retrieval. *In ACM SIGIR Forum*, volume 23, pages 21–30. ACM.

[Lafferty et Zhai, 2001] Lafferty, J. et Zhai, C. (2001). Document language models, query models, and risk minimization for information retrieval. *In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 111–119. ACM.

[Lafourcade, 2007] Lafourcade, M. (2007). Making people play for lexical acquisition with the jeuxdemots prototype. *In SNLP'07 : 7th international symposium on natural language processing*, page 7.

[Lafourcade, 2011] Lafourcade, M. (2011). *Lexique et analyse sémantique de textes-structures, acquisitions, calculs, et jeux de mots*. Thèse de doctorat, Université Montpellier II-Sciences et Techniques du Languedoc.

[Lafourcade et Joubert, 2008] Lafourcade, M. et Joubert, A. (2008). Jeuxdemots : un prototype ludique pour l'émergence de relations entre termes. *In JADT'08 : Journées internationales d'Analyse statistiques des Données Textuelles*, pages 657–666.

[Lafourcade et Joubert, 2009] Lafourcade, M. et Joubert, A. (2009). Similitude entre les sens d'usage d'un terme dans un réseau lexical. *Traitement Automatique des Langues*, 50(1):179–200.

[Lafourcade *et al.*, 2015] Lafourcade, M., Joubert, A. et Le Brun, N. (2015). *Games with a Purpose (GWAPS)*. John Wiley & Sons.

[Lafourcade et Ramadier, 2016] Lafourcade, M. et Ramadier, L. (2016). Semantic relationextraction with semantic patterns : Experiment on radiology report. *In LREC 2016 Conference on Language Resources and Evaluation*, volume 10.

[Lancaster *et al.*, 1991] Lancaster, F. W., Lancaster, F. W., Lancaster, F. W. et Lancaster, F. W. (1991). *Indexing and abstracting in theory and practice*. Library Association London.

[Langlotz, 2006] Langlotz, C. P. (2006). Radlex : A new method for indexing online educational materials 1. *Radiographics*, 26(6):1595–1597.

[Lapata et Keller, 2005] Lapata, M. et Keller, F. (2005). Web-based models for natural language processing. *ACM Transactions on Speech and Language Processing (TSLP)*, 2(1):3.

[Laporte, 2000] Laporte, E. (2000). Mots et niveau lexical. *Ingénierie des langues*, pages 25–49.

[Lauer, 1995] Lauer, M. (1995). Corpus statistics meet the noun compound : some empirical results. *In Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 47–54. Association for Computational Linguistics.

[Lauser *et al.*, 2006] Lauser, B., Sini, M., Liang, A., Keizer, J. et Katz, S. (2006). From agrovoc to the agricultural ontology service/concept server. an owl model for creating ontologies in the agricultural domain. *In Dublin Core Conference Proceedings*. Dublin Core DCMI.

[Lazaridis *et al.*, 2013] Lazaridis, M., Axenopoulos, A., Rafailidis, D. et Daras, P. (2013). Multimedia search and retrieval using multimodal annotation propagation and indexing techniques. *Signal Processing : Image Communication*, 28(4):351–367.

[Le Duff *et al.*, 2000] Le Duff, F., Burgun, A., Cleret, M., Pouliquen, B., Barac'h, V. et Le Beux, P. (2000). Knowledge acquisition to qualify unified medical language system interconceptual relationships. *In Proceedings of the AMIA Symposium*, page 482. American Medical Informatics Association.

[Leacock *et al.*, 1998] Leacock, C., Miller, G. A. et Chodorow, M. (1998). Using corpus statistics and wordnet relations for sense identification. *Computational Linguistics*, 24(1):147–165.

[Lee *et al.*, 2004] Lee, C.-H., Khoo, C. et Na, J.-C. (2004). Automatic identification of treatment relations for medical ontology learning : An exploratory study. *ADVANCES IN KNOWLEDGE ORGANIZATION*, 9:245–250.

[Lee *et al.*, 1999] Lee, J. H., Cho, H. Y. et Park, H. R. (1999). n-gram-based indexing for korean text retrieval. *Information Processing & Management*, 35(4): 427–441.

[Lenoir *et al.*, 1981] Lenoir, P., Roger, M. J., Frangeul, C. et Chalés, G. (1981). Réalisation, développement et maintenance de la base de données adm (creation, development and maintenance of the data-base of a computer-assisted diagnostic system [adm]). *Informatics for Health and Social Care*, 6(1):51–56.

[Lesk, 1986] Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries : how to tell a pine cone from an ice cream cone. *In Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26. ACM.

[Li et Lu, 2016] Li, H. et Lu, Z. (2016). Deep learning for information retrieval.

[Lieberman *et al.*, 2007] Lieberman, H., Smith, D. et Teeters, A. (2007). Common consensus : a web-based game for collecting commonsense goals. *In ACM Workshop on Common Sense for Intelligent Interfaces*.

[Liu et Singh, 2004] Liu, H. et Singh, P. (2004). Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

[Lowe et Barnett, 1994] Lowe, H. J. et Barnett, G. O. (1994). Understanding and using the medical subject headings (mesh) vocabulary to perform literature searches. *Jama*, 271(14):1103–1108.

[Mahesh *et al.*, 1996] Mahesh, K., Helmreich, S. et Wilson, L. (1996). *Ontology development for machine translation : Ideology and methodology*. Citeseer.

[Manning *et al.*, 2008] Manning, C. D., Raghavan, P., Schütze, H. *et al.* (2008). *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge.

[Marchiori, 1998] Marchiori, M. (1998). The limits of web metadata, and beyond. *Computer Networks and ISDN Systems*, 30(1):1–9.

[Maron et Kuhns, 1960] Maron, M. E. et Kuhns, J. L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the ACM (JACM)*, 7(3):216–244.

[Mauldin, 1991] Mauldin, M. L. (1991). Retrieval performance in ferret a conceptual information retrieval system. *In Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 347–355. ACM.

[McCray *et al.*, 2001] McCray, A. T., Burgun, A. et Bodenreider, O. (2001). Aggregating umls semantic types for reducing conceptual complexity. *Studies in health technology and informatics*, 84(0 1):216.

[McInnes et Stevenson, 2014] McInnes, B. T. et Stevenson, M. (2014). Determining the difficulty of word sense disambiguation. *Journal of biomedical informatics*, 47:83–90.

[Medin, 1989] Medin, D. L. (1989). Concepts and conceptual structure. *American psychologist*, 44(12):1469.

[Melchuk et Gentilhomme, 2000] Melchuk, I. A. et Gentilhomme, Y. (2000). *Cours de morphologie générale :(théorique et descriptive). 5. Sixième partie : modèles morphologiques. Septième partie : principes dela description morphologique*, volume 5. PUM.

[Meng et Morioka, 2015] Meng, F. et Morioka, C. (2015). Automating the generation of lexical patterns for processing free text in clinical documents. *Journal of the American Medical Informatics Association*, 22(5):980–986.

[Mihalcea, 2004] Mihalcea, R. (2004). Co-training and self-training for word sense disambiguation. *In Proceedings of the Conference on Computational Natural Language Learning (CoNLL-2004)*.

[Mihalcea et Moldovan, 2000] Mihalcea, R. et Moldovan, D. (2000). Semantic indexing using wordnet senses. *In Proceedings of the ACL-2000 workshop on Recent advances in natural language processing and information retrieval : held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 11*, pages 35–45. Association for Computational Linguistics.

[Millar *et al.*, 2006] Millar, E., Shen, D., Liu, J. et Nicholas, C. (2006). Performance and scalability of a large-scale n-gram based information retrieval system. *Journal of digital information*, 1(5).

[Miller, 1995] Miller, G. A. (1995). Wordnet : a lexical database for english. *Communications of the ACM*, 38(11):39–41.

[Miller *et al.*, 1990] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. et Miller, K. J. (1990). Introduction to wordnet : An on-line lexical database*. *International journal of lexicography*, 3(4):235–244.

[Mitra *et al.*, 1997] Mitra, M., Buckley, C., Singhal, A., Cardie, C. *et al.* (1997). An analysis of statistical and syntactic phrases. *In RIAO*, volume 97, pages 200–214.

[Moffat et Zobel, 2008] Moffat, A. et Zobel, J. (2008). Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)*, 27(1):2.

[Mortensen, 2013] Mortensen, J. M. (2013). Crowdsourcing ontology verification. *In The Semantic Web–ISWC 2013*, pages 448–455. Springer.

[Mortensen *et al.*, 2015] Mortensen, J. M., Minty, E. P., Januszyk, M., Sweeney, T. E., Rector, A. L., Noy, N. F. et Musen, M. A. (2015). Using the wisdom of the crowds to find critical errors in biomedical ontologies : a study of snomed ct. *Journal of the American Medical Informatics Association*, 22(3):640–648.

[Mothe, 1994] Mothe, J. (1994). *Modèle connexionniste pour la recherche d'information. Expansion dirigée de requêtes et apprentissage*. Thèse de doctorat.

[Mozer, 1984] Mozer, M. C. (1984). Inductive information retrieval using parallel distributed computation. Rapport technique, DTIC Document.

[Namer, 2005] Namer, F. (2005). Morphosemantique pour l'appariement de termes dans le vocabulaire medical. approche multilingue. *Traitement Automatique des Langues*, 46(2):157.

[Navigli et Ponzetto, 2010] Navigli, R. et Ponzetto, S. P. (2010). Babelnet : Building a very large multilingual semantic network. *In Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225. Association for Computational Linguistics.

[Nazar *et al.*, 2012] Nazar, R., Vivaldi, J. et Wanner, L. (2012). Co-occurrence graphs applied to taxonomy extraction in scientific and technical corpora. *Procesamiento del lenguaje natural*, 49:67–74.

[Névéol *et al.*, 2014] Névéol, A., Grosjean, J., Darmoni, S. J. et Zweigenbaum, P. (2014). Language resources for french in the biomedical domain. *In LREC*, pages 2146–2151.

[Névéol *et al.*, 2006] Névéol, A., Rogozan, A. et Darmoni, S. (2006). Automatic indexing of online health resources for a french quality controlled gateway. *Information processing & management*, 42(3):695–709.

[Niles et Pease, 2001] Niles, I. et Pease, A. (2001). Towards a standard upper ontology. *In Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001*, pages 2–9. ACM.

[Nottelmann et Fuhr, 2003] Nottelmann, H. et Fuhr, N. (2003). From retrieval status values to probabilities of relevance for advanced ir applications. *Information retrieval*, 6(3-4):363–388.

[Paice, 1984] Paice, C. D. (1984). Soft evaluation of boolean search queries in information retrieval systems. *Information Technology : Research and Development*, 3(1):33–41.

[Papadimitriou *et al.*, 1998] Papadimitriou, C. H., Tamaki, H., Raghavan, P. et Vempala, S. (1998). Latent semantic indexing : A probabilistic analysis. *In Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 159–168. ACM.

[Pastorello Jr *et al.*, 2008] Pastorello Jr, G. Z., Daltio, J. et Medeiros, C. B. (2008). Multimedia semantic annotation propagation. *In Multimedia, 2008. ISM 2008. Tenth IEEE International Symposium on*, pages 509–514. IEEE.

[Pease *et al.*, 2002] Pease, A., Niles, I. et Li, J. (2002). The suggested upper merged ontology : A large ontology for the semantic web and its applications. *In Working notes of the AAAI-2002 workshop on ontologies and the semantic web*, volume 28.

[Pereira *et al.*, 2009] Pereira, S., Massari, P., Buemi, A., Dahamna, B., Serrot, E., Joubert, M. et Darmoni, S. J. (2009). F-mti : outil d'indexation muititerminologique : application à l'indexation automatique de la snomed international. *Risques, Technologies de l'Information pour les Pratiques Médicales*, pages 57–68.

[Pereira *et al.*, 2008] Pereira, S., Névéol, A., Kerdelhué, G., Serrot, E., Joubert, M. et Darmoni, S. J. (2008). Using multi-terminology indexing for the

assignment of mesh descriptors to health resources in a french online catalogue. *In AMIA Annual Symposium Proceedings*, volume 2008, page 586. American Medical Informatics Association.

[Philipp et Völker, 2005] Philipp, C. et Völker, J. (2005). Text2onto-a framework for ontology learning and data-driven change discovery. *In Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems-NLDB*, volume 5, pages 15–17.

[Ponte et Croft, 1998] Ponte, J. M. et Croft, W. B. (1998). A language modeling approach to information retrieval. *In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281. ACM.

[Pouliquen, 2002] Pouliquen, B. (2002). *Indexation de textes médicaux par extraction de concepts, et ses utilisations*. Thèse de doctorat, Université Rennes 1.

[Prié, 2000] Prié, Y. (2000). Sur la piste de l'indexation conceptuelle de documents. *Document numérique*, 4(1-2):11–35.

[Qiu et Frei, 1993] Qiu, Y. et Frei, H.-P. (1993). Concept based query expansion. *In Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 160–169. ACM.

[Quillan, 1963] Quillan, R. (1963). *A notation for representing conceptual information : an application to semantics and mechanical English paraphrasing*. Systems Development Corp.

[Ramadier et Lafourcade, 2015] Ramadier, L. et Lafourcade, M. (2015). Augmentation d'index par propagation sur un réseau lexical application aux comptes rendus de radiologie. *In TALN : Traitement Automatique des Langues Naturelles*.

[Ramadier et Lafourcade, 2016] Ramadier, L. et Lafourcade, M. (2016). Patrons sémantiques pour l'extraction de relations entre termes-application aux comptes rendus radiologiques. *In TALN 2016*.

[Ramadier *et al.*, 2014a] Ramadier, L., Zarrouk, M., Lafourcade, M. et Micheau, A. (2014a). Annotations et inférences de relations dans un réseau lexico-sémantique : application à la radiologie. *In TALN : Traitement Automatique des Langues Naturelles*.

[Ramadier *et al.*, 2014b] Ramadier, L., Zarrouk, M., Lafourcade, M. et Micheau, A. (2014b). Spreading relation annotations in a lexical semantic network applied to radiology. *In International Conference on Intelligent Text Processing and Computational Linguistics*, pages 40–51. Springer.

[Ren *et al.*, 1999] Ren, F., Fan, L. et Nie, J.-Y. (1999). Saak approach : How to acquire knowledge in an actual application system. *In IASTED International Conference on Artificial Intelligence and Soft Computing, Honolulu*, pages 136–140.

[Rich et Knight, 1991] Rich, E. et Knight, K. (1991). Artificial intelligence. *McGraw-Hill, New*.

[Rindflesch *et al.*, 2000] Rindflesch, T. C., Bean, C. A. et Sneiderman, C. A. (2000). Argument identification for arterial branching predications asserted in cardiac catheterization reports. *In Proceedings of the AMIA Symposium*, page 704. American Medical Informatics Association.

[Rink *et al.*, 2011] Rink, B., Harabagiu, S. et Roberts, K. (2011). Automatic extraction of relations between medical concepts in clinical texts. *Journal of the American Medical Informatics Association*, 18(5):594–600.

[Roberts *et al.*, 2008] Roberts, A., Gaizauskas, R. et Hepple, M. (2008). Extracting clinical relationships from patient narratives. *In Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 10–18. Association for Computational Linguistics.

[Robertson, 2004] Robertson, S. (2004). Understanding inverse document frequency : on theoretical arguments for idf. *Journal of documentation*, 60(5):503–520.

[Robertson, 1977] Robertson, S. E. (1977). The probabilistic character of relevance. *Information Processing & Management*, 13(4):247–251.

[Robertson et Walker, 1994] Robertson, S. E. et Walker, S. (1994). Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. *In Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 232–241. Springer-Verlag New York, Inc.

[Robertson et Walker, 1997] Robertson, S. E. et Walker, S. (1997). On relevance weights with little relevance information. *In ACM SIGIR Forum*, volume 31, pages 16–24. ACM.

[Rosse et Mejino Jr, 2008] Rosse, C. et Mejino Jr, J. L. (2008). The foundational model of anatomy ontology. *In Anatomy Ontologies for Bioinformatics*, pages 59–117. Springer.

[Rosse *et al.*, 2003] Rosse, C., Mejino Jr, J. L. *et al.* (2003). A reference ontology for biomedical informatics : the foundational model of anatomy. *Journal of biomedical informatics*, 36(6):478–500.

[Ruch *et al.*, 1999] Ruch, P., Wagner, J., Bouillon, P., Baud, R. H., Rassinoux, A.-M. et Scherrer, J.-R. (1999). Medtag : tag-like semantics for medical document indexing. *In Proceedings of the AMIA Symposium*, page 137. American Medical Informatics Association.

[Ruppenhofer *et al.*, 2012] Ruppenhofer, J., Ellsworth, M., Petruck, M. R., Johnson, C. R. et Scheffczyk, J. (2012). Framenet ii : Extended theory and practice (2006). *URL http ://framenet. icsi. berkeley. edu/book/book. pdf.*

[Sager *et al.*, 1987] Sager, N., Friedman, C. et Lyman, M. S. (1987). Medical language processing : computer management of narrative data.

[Sagot et Fišer, 2008] Sagot, B. et Fišer, D. (2008). Building a free french wordnet from multilingual resources. *In OntoLex*.

[Sagot *et al.*, 2011] Sagot, B., Fort, K., Adda, G., Mariani, J. et Lang, B. (2011). Un turc mécanique pour les ressources linguistiques : critique de la myriadisation du travail parcellisé. *In TALN'2011-Traitement Automatique des Langues Naturelles*.

[Sajous *et al.*, 2013] Sajous, F., Navarro, E., Gaume, B., Prévot, L. et Chudy, Y. (2013). Semi-automatic enrichment of crowdsourced synonymy networks : the wisigoth system applied to wiktionary. *Language Resources and Evaluation*, 47(1):63–96.

[Salton, 1971] Salton, G. (1971). The smart retrieval system—experiments in automatic document processing.

[Salton, 1989] Salton, G. (1989). Automatic text processing : The transformation, analysis, and retrieval of. *Reading : Addison-Wesley*.

[Salton *et al.*, 1983] Salton, G., Fox, E. A. et Wu, H. (1983). Extended boolean information retrieval. *Communications of the ACM*, 26(11):1022–1036.

[Salton et McGill, 1986] Salton, G. et McGill, M. J. (1986). Introduction to modern information retrieval.

[Sanderson, 1994] Sanderson, M. (1994). Word sense disambiguation and information retrieval. *In Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 142–151. Springer-Verlag New York, Inc.

[Sarasua *et al.*, 2012] Sarasua, C., Simperl, E. et Noy, N. F. (2012). Crowdmap : Crowdsourcing ontology alignment with microtasks. *In International Semantic Web Conference*, pages 525–541. Springer.

[Shen *et al.*, 2014] Shen, Y., He, X., Gao, J., Deng, L. et Mesnil, G. (2014). A latent semantic model with convolutional-pooling structure for information retrieval. *In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 101–110. ACM.

[Shore *et al.*, 2012] Shore, M. W., Rubin, D. L. et Kahn Jr, C. E. (2012). Integration of imaging signs into radlex. *Journal of digital imaging*, 25(1):50–55.

[Silberztein, 1999] Silberztein, M. (1999). Text indexation with intex. *Computers and the Humanities*, 33(3):265–280.

[Simmons, 1963] Simmons, R. F. (1963). *Synthetic language behavior*. System Development Corporation.

[Simpson *et al.*, 2014] Simpson, M. S., Voorhees, E. et Hersh, W. (2014). Overview of the trec 2014 clinical decision support track. *In Proc. 23rd Text Retrieval Conference (TREC 2014). National Institute of Standards and Technology (NIST)*.

[Singhal, 2001] Singhal, A. (2001). Modern information retrieval : A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43.

[Siorpaes et Hepp, 2008] Siorpaes, K. et Hepp, M. (2008). Games with a purpose for the semantic web. *IEEE Intelligent Systems*, (3):50–60.

[Snow *et al.*, 2006] Snow, R., Jurafsky, D. et Ng, A. Y. (2006). Semantic taxonomy induction from heterogenous evidence. *In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 801–808. Association for Computational Linguistics.

[Soergel, 1994] Soergel, D. (1994). Indexing and retrieval performance : The logical evidence. *Journal of the American Society for Information Science*, 45(8):589.

[Song *et al.*, 2015] Song, M., Kim, W. C., Lee, D., Heo, G. E. et Kang, K. Y. (2015). Pkde4j : Entity and relation extraction for public knowledge discovery. *Journal of biomedical informatics*, 57:320–332.

[Sowa, 1983] Sowa, J. F. (1983). Conceptual structures : information processing in mind and machine.

[Stairmand et Black, 1996] Stairmand, M. et Black, B. (1996). Conceptual and contextual indexing of documents using wordnet-derived lexical chains. *In Proceedings of 18th BCS-IRSG Annual Colloquium on Information Retrieval Research*.

[Stapley et Benoit, 2000] Stapley, B. J. et Benoit, G. (2000). Biobibliometrics : information retrieval and visualization from co-occurrences of gene names in medline abstracts. *In Pac Symp Biocomput*, volume 5, pages 529–540.

[Stein *et al.*, 1997] Stein, A., Gulla, J. A., Müller, A. et Thiel, U. (1997). Conversational interaction for semantic access to multimedia information. *In Intelligent Multimedia Information Retrieval*, pages 399–421. MIT Press.

[Suchanek *et al.*, 2006] Suchanek, F. M., Ifrim, G. et Weikum, G. (2006). Combining linguistic and statistical analysis to extract relations from web documents. *In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 712–717. ACM.

[Tang *et al.*, 2005] Tang, L., Zhang, Y. et Fu, X. (2005). The statistic properties of chinese semantic network in hownet. *In 2005 International Conference on Natural Language Processing and Knowledge Engineering*, pages 58–61. IEEE.

[Thaler *et al.*, 2011] Thaler, S., Siorpaes, K., Simperl, E. et Hofer, C. (2011). A survey on games for knowledge acquisition. *Rapport technique, STI*, page 26.

[Uzuner *et al.*, 2011] Uzuner, Ö., South, B. R., Shen, S. et DuVall, S. L. (2011). 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

[Venhuizen *et al.*, 2013] Venhuizen, N., Basile, V., Evang, K. et Bos, J. (2013). Gamification for word sense labeling. *In Proc. 10th International Conference on Computational Semantics (IWCS-2013)*, pages 397–403.

[Von Ahn et Dabbish, 2008] Von Ahn, L. et Dabbish, L. (2008). Designing games with a purpose. *Communications of the ACM*, 51(8):58–67.

[Voorhees, 1993] Voorhees, E. M. (1993). Using wordnet to disambiguate word senses for text retrieval. *In Proceedings of the 16th annual international ACM*

SIGIR conference on Research and development in information retrieval, pages 171–180. ACM.

[Voorhees et Harman, 2001] Voorhees, E. M. et Harman, D. (2001). Overview of trec 2001. *In Trec*, pages 1–15.

[Vossen, 1998] Vossen, P. (1998). *A multilingual database with lexical semantic networks*. Springer.

[Wang *et al.*, 2006] Wang, T., Li, Y., Bontcheva, K., Cunningham, H. et Wang, J. (2006). Automatic extraction of hierarchical relations from text. *In European Semantic Web Conference*, pages 215–229. Springer.

[Weske-Heck *et al.*, 2002] Weske-Heck, G., Zaiss, A., Zabel, M., Schulz, S., Giere, W., Schopen, M. et Klar, R. (2002). The german specialist lexicon. *In Proceedings of the AMIA Symposium*, page 884. American Medical Informatics Association.

[Woods, 1997] Woods, W. A. (1997). Conceptual indexing : A better way to organize knowledge.

[Xiao *et al.*, 2005] Xiao, J., Su, J., Zhou, G.-d. et Tan, C. (2005). Protein-protein interaction extraction : a supervised learning approach. *In Proc Symp on Semantic Mining in Biomedicine*, pages 51–59.

[Zadrozny et Kacprzyk, 2005] Zadrozny, S. et Kacprzyk, J. (2005). An extended fuzzy boolean model of information retrieval revisited. *In FUZZ-IEEE*, pages 1020–1025.

[Zaragoza *et al.*, 2004] Zaragoza, H., Craswell, N., Taylor, M. J., Saria, S. et Robertson, S. E. (2004). Microsoft cambridge at trec 13 : Web and hard tracks. *In TREC*, volume 4, pages 1–1.

[Zarrouk *et al.*, 2013] Zarrouk, M., Lafourcade, M. et Joubert, A. (2013). Inference and reconciliation in a crowdsourced lexical-semantic network. *Computación y Sistemas*, 17(2):147–159.

[Zeng-Treitler *et al.*, 2007] Zeng-Treitler, Q., Goryachev, S., Kim, H., Keselman, A. et Rosendale, D. (2007). Making texts in electronic health records comprehensible to consumers : a prototype translator. *In AMIA*, pages 846–50.

[Zhai et Lafferty, 2001] Zhai, C. et Lafferty, J. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. *In Procee-*

dings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pages 334–342. ACM.

[Zhang et Su, 2001] Zhang, H. et Su, Z. (2001). Improving cbir by semantic propagation and cross modality query expansion. *In Proc. of the int. workshop on multimedia content-based indexing and retrieval. Brescia*, pages 19–21.

[Zipf, 1935] Zipf, G. K. (1935). The psycho-biology of language.

[Zobel *et al.*, 1998] Zobel, J., Moffat, A. et Ramamohanarao, K. (1998). Inverted files versus signature files for text indexing. *ACM Transactions on Database Systems (TODS)*, 23(4):453–490.

[Zweigenbaum, 1999] Zweigenbaum, P. (1999). Encoder l'information médicale : des terminologies aux systèmes de représentation des connaissances. *Innovation Stratégique en Information de Santé*, 2:5.

[Zweigenbaum, 2004] Zweigenbaum, P. (2004). L'umls entre langue et ontologie : une approche pragmatique dans le domaine médical. *Revue d'intelligence artificielle*, 18(1):111–137.

[Zweigenbaum *et al.*, 2003] Zweigenbaum, P., Baud, R., Burgun, A., Namer, F., Jarrousse, É., Grabar, N., Ruch, P., Le Duff, F., Thirion, B. et Darmoni, S. (2003). Towards a unified medical lexicon for french. *In Medical Informatics in Europe (MIE)*.

[Zweigenbaum et Grabar, 2000] Zweigenbaum, P. et Grabar, N. (2000). Expériences d'acquisition automatique de connaissances morphologiques par amorçage à partir d'un thesaurus. *In Proceedings of the*, volume 12.

Annexes

Annexe A : Exemples de comptes rendus de radiologie

Compte rendu radiologique numéro 1

patient : @@@
date de naissance : 1940
scanner abdominal
indications : bilan pré-greffe rénale.
technique : coupes axiales sur l'abdomen et le pelvis.
resultats :
l'examen confirme la présence de signe de diverticulite avec la présence de volumineux diverticules individualisés en particulier un diverticule au niveau de la jonction recto-sigmoïdienne et de plus petits diverticules sigmoïdiens associés à une infiltration significative de la graisse péri- sigmoïdienne.
on retrouve l'atrophie rénale connue bilatérale. le greffon rénal est en position gauche.
il est à noter que l'examen est réalisé sans injection de produit de contraste.
@ @ @

Compte rendu radiologique numéro 2

angioscanner thoraco-abdominal
motif de l'examen : hypertension artérielle mal équilibrée dans le cadre d'un syndrome de Myrrhe avec dysplasie artérielle et une hypoplasie diffuse de l'aorte descendante. recherche d'un éventuel rétrécissement localisé en particulier au niveau de l'isthme aortique (à l'échographie, accès limité, l'aorte devient hypoplasique à partir de cette zone sans flux véritable de coarctation aortique (diamètre minimal entre 8 et 9 mm).
technique : acquisition spiralée thoraco-abdominale jusqu'à la bifurcation aortique, après injection de 60 cc de ioméron en intraveineux, sans complication en cours d'examen. reconstructions dans les différents plans. rapport de dose : dlp = 219.54 mgy.cm
resultat
thorax : morphologie habituelle de l'aorte ascendante qui diminue progressivement de calibre après l'émergence du tronc brachiocéphalique droit. au niveau de la région isthmique, présence d'un rétrécissement localisé avec image en "feuillet" inférieure. le rétrécissement concerne seulement les diamètres antéropostérieur et vertical (6 mm), alors que le diamètre transversal (11 mm) est conservé. l'aorte en amont mesure environ 18.8 mm et en aval 18.8 puis 14 mm pour diminuer progressivement de calibre à 10.8 mm. pas de réseau de collatérales : les vaisseaux de la gerbe, les artères intercostales, les artères thoraciques internes sont de calibre normal.
pas d'anomalie du parenchyme pulmonaire. pas d'épanchement pleural ou péricardique. pas de mise en évidence d'adénomégalie médiastinale.
a la jonction thoraco abdominale, nette diminution du calibre de l'aorte sur environ 20 mm de hauteur, qui mesure dans son minimum environ 3.7 mm de diamètre antéro-postérieur. en aval l'aorte est globalement hypoplasique, régulière (7 mm de diamètre). pas de visualisation de l'émergence du tronc coeliaque suspect d'une sténose à ce niveau. l'émergence de l'artère mésentérique supérieure est bien identifiée ainsi que l'émergence des deux artères rénales , fines mais régulières. au niveau abdominal, en tenant compte du temps d'injection, pas d'anomalie hépatique, splénique, pancréatique, rénale ou surrénalienne. pas d'épaississement des parois digestives. pas de mise en évidence d'air ou de liquide libre. pas d'anomalie des structures osseuses.
conclusion
rétrécissement diffus de l'aorte thoraco-abdominale qui débute après l'émergence de l'artère sous-clavière gauche mais nettement plus marqué à la jonction thoraco abdominale. rétrécissement vertical localisé au niveau de l'isthme aortique, avec conservation du diamètre transversal. L'aspect évoque plutôt une plicature bien que l'arc aortique n'apparaisse pas allongé de façon anormale. l'absence de collatérales est en faveur de cette hypothèse.
probable sténose serrée en regard de l'origine du tronc coeliaque.

Compte rendu radiologique numéro 3

SCANNER THORACO ABDOMINO PELVIEN
INDICATIONS : Patient opéré par coelio d'un néo du rectum le 8.08.2010. Repris quelques jours parès pour péritonite stercorale sur lachage de suture (hartman). A J6, fièvre à 39C. Métastaséctomie hépatique foie gauche.
TECHNIQUE : Acquisition hélicoidale thoraco-abdomino-pelvienne après injection de 110cc de xenetix 350. DLP : 1936mGy-cm
RESULTATS :
* Abdomino-pelvien : Ganglions juxta-centimétriques rétropéritonéaux para-aortique gauche.
Métastases hépatiques connues dans le foie droit et stigmates de métastaséctomie dans le foie gauche.
Air intra-vécisal en rapport avec la sonde urinaire. Absence de lésion osseuse d'allure secondaire.
CONCLUSION :
Présence de multiples collections intra-abdominales dont les plus volumineuses sont sous phrénique et au niveau du flanc droit(de 12cm de plus grand axe, avec paroi) avec signe de péritonite radiologique Bulles de pneumopéritoine à proximité de la poche de colostomie , à confronter à la date de reprise chirurgicale. Artère mésentérique inférieure non visualisée.
Epaississement pariétal de l'ensemble du cadre colique Epanchement pleural bilatéral. @ @ @

Annexe B : Exemples de comptes rendus et indexation des variables PMA

cas numéro 1

Douleurs sus ombilicales remontant dans les deux fosses lombaires chez une *patiente de 33 ans*.
Pas de syndrome inflammatoire biologique. Résultats
Formation d'allure kystique aux parois épaissies de l'aire ovarienne droite.
Hémopéritoine *(40 UH)* modéré pelvien.
L'ensemble est en faveur d'un *kyste ovarien* droit hémorragique rompu.
Diagnostic
Kyste hémorragique rompu

Patient *r_isa* patiente
Patient (âge = 33 ans)
Patient *r_symptome* douleurs sus ombilicales
Modalidé *r_isa* scanner (40 UH)
Affection *r_isa* kyste ovarien

cas numéro 2

Anamnèse Déficit hémicorps gauche survenu brutalement. Antécédents d'AVC.
Résultats Plage d'hypodensité fronto-insulaire et lenticulo-caudée droite en rapport avec un accident ischemique sylvien droit récent. Aspect spontanément hyperdense de la portion M1 de l'artère cérébrale droite en rapport avec un thrombus endoluminal, dont l'occlusion est confirmée par l'angioscanner cervico-céphalique. Hypodensité cortico-sous-corticale du gyrus frontal moyen droit en rapport avec un accident ischemique ancien, séquellaire.
Conclusion Récidive d'accident vasculaire cérébral sylvien ischémique en TDM précoce

Patient *r_symptome* déficit hémicorps
Modalidé *r_isa* scanner (hypodensité, hyperdensité)
Affection *r_isa* accident vasculaire cérébral (conclusion)

Annexe C : Exemples d'index augmenté

indications : fracture du tibia droit, chute de ski **technique** : une série de coupes axiales transverses sur l'ensemble de la cheville sans injection de produit de contraste. etude en fenêtres parties molles et osseuses. **Resultats** fractures diaphysaires spiroïdes à trois fragments principaux du 1/3 distal du tibia et de la fibula avec discret déplacement vers l'avant, sans trait de refend articulaire. Fractures de la base de M2 et de M3 non articulaires et non déplacées. Fracture articulaire de la partie interne de la base de M1 non déplacée. Atrophie avec dégénérescence marquée des corps musculaires de l'ensemble des loges.

fracture des membres inférieurs • accident de sport d'hiver • fracture du tibia • fracture diaphysaire • accident de ski • traumatisme des membres inférieurs • fracture multiple • cheville>anatomie •fracture articulaire • chute>tomber • dégénérescence musculaire • fracture non articulaire • fracture non déplacée •fracture>lésion • fracture spiroïde • fracture avec déplacement • imagerie médicale • trauma • jambe • lésion • lésion osseuse • loge>anatomie • radiologie • médecine • spiroïde • sports d'hiver

Annexe D : Exemple de compte rendu avec les index brut et augmenté

Anamnèse Déficit hémicorps gauche survenu brutalement. Antécédents d'AVC.
Résultats Plage d'hypodensité fronto-insulaire et lenticulo-caudée droite en rapport avec un accident ischemique sylvien droit récent. Aspect spontanément hyperdense de la portion M1 de l'artère cérébrale droite en rapport avec un thrombus endoluminal, dont l'occlusion est confirmée par l'angioscanner cervico-céphalique. Hypodensité cortico-sous-corticale du gyrus frontal moyen droit en rapport avec un accident ischemique ancien, séquellaire.
Conclusion Récidive d'accident vasculaire cérébral sylvien ischémique en TDM précoce

Compte rendu initial

déficit • hémicorps • AVC • hypodensité • fronto-insulaire • lenticulo-caudée • accident ischémique • sylvien • hyperdense • artère cérébrale droite • thrombus • endoluminal • occlusion • angioscanner cervico-céphalique • hypodensité • cortico-sous-corticale • gyrus frontal moyen • accident ischemique ancien • séquellaire • récidive • sylvien • ischémique • TDM

Index brut du compte rendu

paralysie • moitié du corps • accident vasculaire cérébral • densité • scanner • gyrus • circonvolution cérébrale • territoire sylvien • hyperdensité • artère cérébrale • artère du cerveau • caillot • angioscanner • produit de contraste • vaisseaux du cou • iode • cerveau>anatomie • lobe frontal • accident ischemique • séquelle • rechute • aire de Broca • aire de Wernicke • tomodensitométrie capsule interne •

Index augmenté (terme non présent dans le comptemais susceptible de faire l'objet de requête)

Annexe E : Exemple du prototype du moteur de recherche

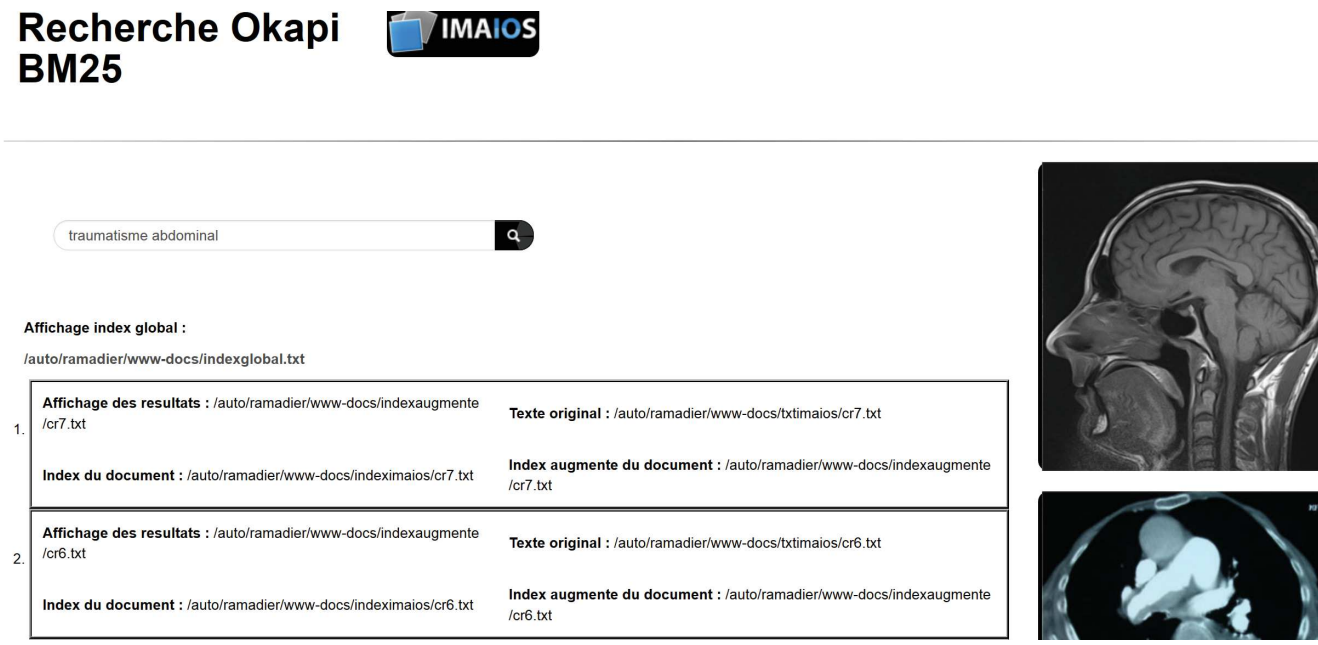


FIGURE 8 – Capture écran du prototype du moteur de recherche Okapi