REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE MOULOUD MAMMERI Tizi-Ouzou

Faculté de Génie Electrique et Informatique

Département d'Informatique

MEMOIRE

En vue de l'obtention du diplôme de Master en informatique

Spécialité : Conduite de projets informatiques

Thème

Proposition et évaluation d'une mesure de similarité sémantique entre concepts MeSH

Proposé et dirigé par :

M^me^ : F. AMIROUCHE

Réalisé par :

M^elle^ : DJELIL Katia

2012/2013

Remerciements

Au terme de ce travail je tiens à adresser mes plus sincères et vifs remerciements et ma profonde gratitude à Mme F.AMIROUCHE pour avoir accepté de m'encadrer, et pour les conseils et orientations tant précieux qu'elle m'a prodigué.

Je tiens à exprimer toute ma reconnaissance à Mlle S.CHERDIOUI et Mlle W. AZZOUG pour les multiples conseils qu'elles m'ont prodigué, et pour leur disponibilité à mon égard.

J'adresse également mes sincères remerciements et mes respects aux membres du jury pour m'avoir fait l'honneur de juger mon travail.

Merci

DEDICACES

A mes parents,

A ma sœur Radia,

A mon frère Idir,

A mes Ami(e)s,

Je dédie ce travail.

Sommaire

Liste des figures

Liste des tableaux

INTRODUCTION GENERALE

Introduction Générale

L'évolution des technologies de l'information a fourni d'importants avantages à la recherche biomédicale, afin de permettre aux scientifiques et aux chercheurs d'appliquer les sciences informatiques à leurs études et expériences. Par conséquent, une quantité importante de données biomédicales a été générée. A titre d'exemple, MEDLINE (*Medical Literature Analysis and Retrieval System Online*) est la base de données bibliographique de premier ordre, développée par la NLM (*US National Library of Medicine*). Elle contient plus de 19 millions de références d'articles en science de la vie, notamment de la biomédecine.

Cette information est utile uniquement si les méthodes pour traiter les données disponibles sont efficaces. Les systèmes de recherche d'information sont alors face à un grand défi pour traiter ces grandes quantités de données. Pour cela l'utilisation de terminologies a été reconnue utile et nécessaire afin de représenter et traiter les données biomédicales. MeSH est un référentiel dans ce domaine, en effet, les concepts qu'il hiérarchise sont utilisés dans les tâches d'indexation manuelles ou automatiques de documents biomédicaux notamment ceux de la base MEDLINE.

Dans leurs tâches de recherche les systèmes de recherche biomédicaux ont besoin de concepts sémantiquement liés. Ceux-ci sont offerts à travers la similarité sémantique qui existe entre les concepts. La similarité sémantique est généralement estimée entre concepts grâce à des métriques dites mesures de similarité sémantiques.

De nombreuses mesures de similarité sémantique ont été proposées dans la littérature, La plupart ont été conçues pour évaluer la similarité sémantique entre mots dans WordNet, une ontologie du domaine général. Ces mesures se basent sur la structure de la ressource utilisée ou sur les propriétés qu'offrent ces ressources. Dans le domaine biomédical, une tentative de mise en place d'une mesure de similarité sémantique pour la comparaison de concepts issus du thésaurus MeSH a été avancée celle-ci se base principalement sur les propriétés de ces concepts.

Notre travail s'inscrit dans ce contexte et vise principalement à proposer et à implémenter une mesure de similarité sémantique entre concepts biomédicaux issus du thésaurus MeSH. Notre démarche pour ce faire consiste principalement à adapter les mesures existantes dédiées à WordNet, pour une utilisation spécifique dans MeSH.

Organisation du mémoire

Dans le but d'atteindre nos objectifs, nous avons structuré notre mémoire en quatre chapitres comme suit :

- Le premier chapitre introduit l'indexation et la recherche d'information dans le domaine biomédical.
- Le deuxième chapitre présente un état de l'art des mesures de similarité sémantique existantes dans le domaine général.
- Le troisième chapitre présente notre contribution en décrivant dans un premier temps notre mesure de similarité sémantique entre concepts biomédicaux, et dans un second temps, un exemple illustratif de l'utilisation de cette mesure.
- Le quatrième chapitre présente l'évaluation expérimentale de notre approche.

Et enfin, nous conclurons en proposant des perspectives permettant d'améliorer notre proposition.

CHAPITRE I

Recherche d’information dans le domaine Biomédical

I. Introduction

Depuis l'apparition de l'informatique, les connaissances stockées sur support numérique n'ont cessé de s'accumuler, et ce, dans de nombreux domaines en particulier le domaine biomédical. Ainsi, la Recherche d'Information (RI) devient davantage cruciale et les Systèmes de Recherche d'Information (SRI) deviennent une aide inestimable. Pour cela, plusieurs approches de recherche ont été proposées, les premières qualifiées d'approches classiques, se basent sur une recherche par mots clés dans lesquelles, les documents et les requêtes sont représentés par des sacs de mots souvent pondérés, et leur pertinence vis-à-vis d'une requête utilisateur est souvent estimée en s'appuyant sur les fréquences d'apparition des mots de la requête dans ces mêmes documents. Cependant, ces approches ne tiennent pas compte des sens des mots, ce qui peut parfois poser problème étant donnée la richesse et l'ambiguïté de la langue dans tous les domaines. Pour pallier ces problèmes d'autres approches se basant sur les sens des mots pour représenter les documents et les requêtes sont apparues à travers l'indexation sémantique, qui a servi notamment dans le domaine biomédical.

Dans ce chapitre, nous allons donner un aperçu des notions de base de la recherche d'information classique, nous allons ensuite décrire l'indexation sémantique, et la recherche d'information dans le domaine biomédical.

I. La Recherche d'information

La recherche d'information (RI) est un ensemble de méthodes et procédures ayant pour objet d'extraire à partir d'un ensemble de documents, les informations voulues exprimées par l'utilisateur à travers une requête [Serres, 02]. Elle est mise en œuvre à travers les Systèmes de Recherche d'Informations (SRI) qui sont des ensembles de programmes informatiques et de procédures qui ont pour but de sélectionner les informations dites pertinentes répondant aux besoins des utilisateurs.

II.1. Le processus de RI classique

Le processus de RI a pour but de mettre en correspondance les représentations des informations contenues dans un fond documentaire d'une part, avec celles des besoins de l'utilisateur d'autre part, ou en d'autre termes, de correspondre au mieux la pertinence système avec la pertinence utilisateur.

Ce processus est composé de trois fonctions principales qui sont :

- **L'indexation :** elle permet de créer une représentation des documents et des requêtes auxquels elle associe des descripteurs, ensuite un poids est attribué à chacun de ces descripteurs, il permet de déterminer son degré de représentativité dans le document ou la requête.
- **L'appariement document-requête** : il permet la comparaison des représentations issues de l'étape d'indexation afin de déterminer leur degré de correspondance (ou similarité), et de sélectionner l'ensemble des documents potentiellement pertinents pour une requête. Il existe deux méthodes d'appariement : l'appariement exact qui permet de récupérer les documents qui correspondent exactement à la requête spécifiée, et l'appariement approché dans lequel un score de pertinence RSV(Q,D) (*Retriviel Status Value*) entre la requête Q indexée et les descripteurs du document D est calculé. Les documents qui correspondent au mieux à la requête sont alors retournés à l'utilisateur.
- **La reformulation de requêtes :** Elle permet de réécrire autrement la requête utilisateur puisqu'il est quasi impossible aujourd'hui de retrouver des informations pertinentes en utilisant seulement la requête initiale de l'utilisateur, et ce à cause du volume croissant des bases documentaires. Cette reformulation consiste à modifier la requête initiale par ajout de termes significatifs et/ou ré-estimation de leur poids. Si les termes rajoutés proviennent des documents de la collection, on parle de réinjection de pertinence si le processus est supervisé, et de pseudo réinjection de pertinence si le processus est automatique. Si par contre, les termes rajoutés ne proviennent pas des documents de la collection, on parle d'expansion de requête.

Les fonctionnalités d'un SRI sont représentées schématiquement par ce que l'on appelle communément le Processus en « U ». Ce processus schématisé dans la figure I.1 explique de manière générale la recherche dans un SRI qui consiste à mettre en correspondance le besoin en information de l'utilisateur exprimé sous forme de requête, et les informations disponibles. Le but étant de lui retourner un ensemble de résultats pertinents. La pertinence du document est alors dépendante du jugement de cet utilisateur car il doit contenir l'information qu'il recherche. Dans ce processus, les informations manipulées se trouvent dans des documents regroupés dans des collections et des requêtes, ces trois éléments sont définis comme suit :

- La collection de documents : elle constitue l'ensemble des informations exploitables et accessibles. Elle est composée d'un ensemble de documents.

- Le document : il constitue l'information élémentaire d'une collection de documents. Il est constitué par un texte, un morceau de texte, une image, une bande de vidéo etc., qui peut être retourné en réponse à une requête (ou besoin en information) d'un utilisateur.

- La requête : elle constitue l'expression du besoin en information de l'utilisateur. Elle représente l'interface entre un SRI et l'utilisateur. Elle est souvent exprimée par un ensemble de mots clés (exemple des systèmes SMART et OKAPI), mais elle peut être également exprimée en langage naturel (exemple du système SMART),…etc.

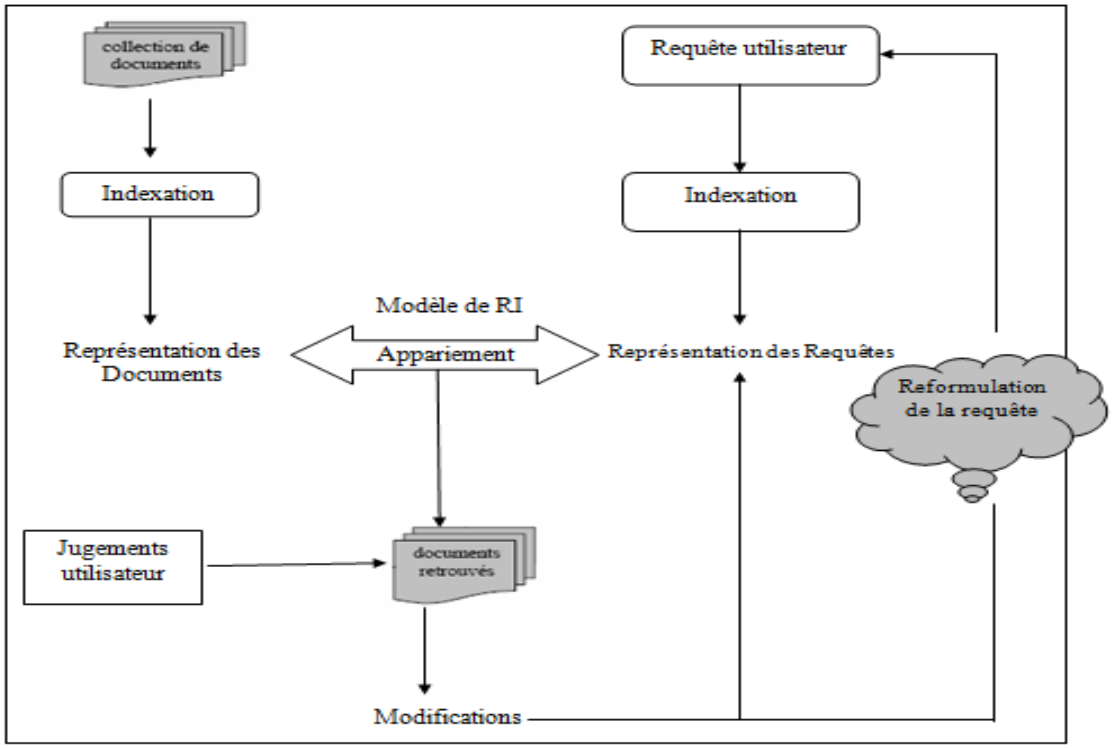

Figure I.1 : Processus en U de la recherche d'information

La majorité des systèmes actuellement disponibles dans la RI classique permettent de déterminer et de sélectionner les documents pertinents pour une requête utilisateur en se basant sur l'existence ou non des mots représentant la requête dans les documents. Cette représentation ne prend pas en considération le sens des mots qu'elle utilise, alors qu'un document peut être pertinent même s'il ne partage pas les mêmes mots avec la requête. De là est née la nécessité de prendre en considération le sens des mots comme facteur de choix des éléments représentatifs des documents et des requêtes, ceci est mis en œuvre à travers l'indexation sémantique.

L'indexation sémantique offre une possibilité de traiter avec l'ambiguïté de la langue naturelle. En effet, elle permet de représenter les documents et les requêtes par des entités

véhiculant des sens appelées concepts. Pour ce faire, elle se base principalement sur un processus de désambiguïsation des sens des mots. Ce processus s'appuie dans la plupart des cas sur des ressources terminologiques externes [Vorhees, 93] [Baziz , 05] [Amirouche et al, 11]. Il associe généralement un score de désambiguïsation au différents sens possibles d'un mot fournis par ces ressources. Ces scores sont attribués dans la plupart des cas à base de proximité (similarité) sémantique entre les concepts [Baziz , 05] [Amirouche et al, 11]. Cette notion est l'objet de notre étude, nous la détaillerons dans le prochain chapitre.

La RI est appliquée dans différents domaines. Dans ce qui suit nous allons nous intéresser à son utilisation dans le domaine biomédical.

II.2. Indexation de la littérature biomédicale

Avec l'accroissement régulier des publications dans le domaine biomédical, la littérature biomédicale se diversifie, ce qui a créé la nécessité d'un traitement automatique de celle-ci à de fins diverses comme la RI, la classification, les analyses statistiques, etc.

Une étape primordiale requise dans cette automatisation est l'indexation. Le but est de créer une représentation à base de mots clés ou concepts permettant de faciliter la récupération dans un ensemble de documents.

Pour mettre en œuvre cette indexation, plusieurs terminologies normalisées sont utilisées. En effet, la standardisation du langage biomédical a donné naissance à des ressources terminologiques à l'exemple du thésaurus MeSH (Medical Subject Headings) pour la représentation des concepts biomédicaux, de l'ontologie GO (Gene Ontology) pour la représentation des gènes et protéines, SNOMED (Systematized NOmenclature of MEDicine) pour la représentation des actes médicaux, et UMLS (Unified Medical Language System) qui regroupe plus de 150 de ces terminologies. L'indexation des documents biomédicaux s'appuie donc sur ces terminologies diverses comme principales sources d'identification des descripteurs.

Dans ce qui suit, nous allons présenter l'information biomédicale, les principales ressources terminologiques biomédicales, et brièvement la RI dans le domaine biomédical.

II.2.1. L'information biomédicale

L'information biomédicale est composée essentiellement des bases de données bibliographiques qui font référence à des revues scientifiques et des comptes rendus des conférences du milieu biomédical. A titre d'exemple, MEDLINE (MEDical Literature

Analysis and Retrieval System on LINE) qui est une banque documentaire produite par la NLM (National Library of Medecine) qui couvre les domaines biomédicaux tels que la biologie, la biochimie, la médecine clinique, la santé publique, la pharmacologie, l'économie liée à la santé, la toxicologie, l'odontologie, la psychiatrie et la médecine vétérinaire. MEDLINE est accessible en ligne via le portail PubMed[1].

II.2.2. Ressources terminologiques biomédicales

L'information biomédicale est aussi regroupée dans différents ressources terminologiques tel que les terminologies, les thésaurus, les classifications, les nomenclatures et les ontologies.

Terminologie : une terminologie est un ensemble de termes rigoureusement définis qui sont spécifiques à un domaine. Elle implique la normalisation des termes d'un domaine afin de pouvoir les organiser les uns par rapport aux autres.

Thésaurus : un thésaurus est un langage documentaire fondé sur une structuration hiérarchisée, alphabétique au premier niveau puis thématique. Les termes normalisés étant reliés à des termes plus précis. Ils y sont organisés de manière conceptuelle et reliés entre eux par des relations sémantiques.

Classification : une classification est la répartition systématique en classes, en catégorie d'êtres, de choses, d'objets ou de notions ayant des caractères communs notamment afin d'en faciliter l'étude.

Nomenclature : une nomenclature est une terminologie qui vise à recenser tous les termes d'un domaine et qui fournit un éventail plus varié et plus précis de concepts.

Ontologie : une ontologie est une description formelle d'un domaine à travers ses concepts et les relations qui existent entre eux.

Dans ce qui suit nous présenteront deux principales ressources biomédicales à savoir le thésaurus MeSH et le système de langage médical unifié UMLS.

II.2.2.1. Le thésaurus MeSH

Le MeSH est un thésaurus de référence dans le domaine biomédical développé par la NLM aux Etats-Unis. Il s'agit d'un vocabulaire normalisé, qui permet d'exprimer une notion donnée d'une manière indépendante du langage courant. Ce vocabulaire est utlisé pour

[1] http://www.ncbi.nlm.nih.gov/pubmed

indexer, classer et rechercher des documents, notamment ceux de la base MEDLINE à travers son interface PubMed. Il est constitué de seize domaines indépendants (figure I.2) structurés selon seize arboréscences pouvant contenir jusqu'à douze niveaux hiéararchiques qui partent des termes générique vers les termes spécifiques.

1. + Anatomy [A]
2. + Organisms [B]
3. + Diseases [C]
4. + Chemicals and Drugs [D]
5. + Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]
6. + Psychiatry and Psychology [F]
7. + Phenomena and Processes [G]
8. + Disciplines and Occupations [H]
9. + Anthropology, Education, Sociology and Social Phenomena [I]
10. + Technology, Industry, Agriculture [J]
11. + Humanities [K]
12. + Information Science [L]
13. + Named Groups [M]
14. + Health Care [N]
15. + Publication Characteristics [V]
16. + Geographicals [Z]

Figure I.2 : Les seize domaines de MeSH[2]

Chacun des domaines (arborescences) de MeSH contient un ensemble de descripteurs. Un descripteur est composé d'un ou plusieurs concepts et porte le nom de l'un de ces concepts dit concept préféré, les autres concepts sont dits non préférés, une relation d'associativité peut exister entre un concept préféré et un ou plusieurs concepts non préférés. Les descripteurs sont quand à eux reliés par l'une des relations hiérarchiques suivantes :

- **Méronymie / holonymie :** Un concept X est *méronyme* d'un concept Y si X « est une partie constituante de» (part of) ou « membre de» (member of) Y. Y est alors dit *holonyme* de X.
- **Hyponymie / hyperonymie :** Un concept est hyperonyme si son sens inclut d'autres concepts, qui sont ses hyponymes. Un concept X est donc, un *hyponym*e d'un concept Y si X « est un type de » (is-a) Y. Y est alors dit *hyperonyme* de X.

La relation hiérarchique dans MeSH est exprimée par un ensemble de codes alphanumériques attribués aux descipteurs. La figure I.3 montre un extrait de l'arborescence A « Anatomie » de MeSH.

[2] www.nlm.nih.gov/egi/mesh/2013

Anatomy [A]

Body Regions [A01]

Extremities [A01.378]

Upper Extrimity [A01.378.800]

Hand [A01.378.800.667]

Fingers [A01.378.800.667.430]

Thumb [A01.378.800.667.430.705]

Head [A01.456]

Face [A01.456.505]

Eye [A01.456.505.420]

Sense Organs [A09]

Eye [A09.371]

Figure I.3 : Extrait de l'arborescence A (domaine « Anatomie ») de MeSH[3]

Dans cet extrait, le concept préféré ***fingers*** (doigts) est une partie constituante du concept préféré ***hand*** (Main), et le concept préféré ***thumb*** (pouce) est un hyponyme (est-un) du concept préféré **fingers** (doigt).

Dans Mesh, Certains descripteurs ont plusieurs localisations, au sein de la même catégorie ou de catégories différentes, ces descripteurs sont alors dits poly-hiérarchiques**,** par exemple le descripteur ***« pain »*** appartient à plusieurs endroits de l'arborescence C (Maladies) avec les codes C10.597.617, C23.888.646, et C23.888.592.612, mais aussi à l'arborescence F (Psychiatrie et psychologie) avec le code F02.830.816.444.

Un concept est composé d'un ou plusieurs termes et porte le nom de l'un d'entre eux qui est le terme préféré, les autres sont dits non préférés. Les termes composant un concept sont reliés par une relation de synonymie.

II.2.2.2. Le système de langage médical unifié UMLS

Développé par la NLM, UMLS est un système multilingue qui contient plus de neuf millions de termes synonymes regroupés en concepts, reliés entre eux par des relations. Ce système réunit trois bases de connaissances à savoir, le méta thésaurus, le réseau sémantique et le SPECIALIST Lexicon.

[3] www.nlm.nih.gov/egi/mesh/2013

- Le méta thésaurus UMLS constitue une base unifiée des concepts issus de plus de 150 terminologies médicales (dont MeSH, CIM-10, SNOMED). Il regroupe les différents termes synonymes (issus des différentes terminologies qui le composent) sous un même concept. Les relations entre ces concepts son celles des terminologies de base.
- Le réseau sémantique a pour objectif de fournir une catégorisation cohérente de tous les concepts représentés dans le méta thésaurus UMLS, et de fournir un ensemble de relations utiles entre ces catégories. En effet, il comporte 135 catégories de concepts nommées « types sémantiques » qui fournissent une catégorisation cohérente de tous les concepts du méta thésaurus organisés sous la forme de réseau (par exemple Disease, Syndrome, Clinical Drug,…), et un ensemble de 54 relations sémantiques qui relient les types sémantiques entre eux, elles peuvent être hiérarchiques (is-a), ou non hiérarchiques. Chaque concept du méta thésaurus est catégorisé par au moins un type sémantique du réseau sémantique, et ce, indépendamment de sa position hiérarchique dans le vocabulaire dont il est issu.
- Le SPECIALIST Lexicon contient les informations syntaxiques, morphologiques et orthographiques nécessaires au traitement automatique de la langue anglaise. Chacune de ses entrées possède une forme de base (lemme), une catégorie syntaxique, un identifiant unique et éventuellement des variantes orthographiques. Il est également utilisé pour des tâches de traitement automatique de la langue.

II.2.3. Indexation contrôlée et RI dans les documents biomédicaux

Comme l'indexation dans les différents domaines, l'indexation en RI biomédicale vise à apporter des facilités d'accès à la littérature biomédicale en affectant à chaque document une liste de termes désignant des concepts issus d'une ou plusieurs terminologies biomédicales [Névéol et al., 06], cette indexation est alors dite contrôlée. En effet, l'indexation contrôlée dans le domaine biomédical est un processus basé sur une terminologie (indexation mono-terminologique), ou sur plusieurs terminologies (indexation multi-terminologique) consistant à construire une représentation d'un document en choisissant ses descripteurs dans un langage documentaire préalablement défini. Elle est basée généralement sur une méthode d'*extraction de concepts* qui peut être réalisée *manuellement* par des documentalistes ou experts du domaine ayant des connaissances approfondies des terminologies et des années d'expérience, ou *automatiquement* par des approches linguistiques ou statistiques [Fu et al., 02].

L'efficacité des systèmes de recherche est influencée par le degré de chevauchement des termes entre les requêtes des utilisateurs et les documents pertinents. Quand un utilisateur cherche l'information dans une collection de documents, il peut formuler la requête en utilisant d'autres expressions pour mentionner la même information dans le document. Cela cause un problème d'incompatibilité des termes qui donne des résultats de recherche pauvres. Dans le domaine biomédical, les documents contiennent de nombreuses expressions différentes ou des variantes de termes pour un même concept, comme la synonymie ('*cancer*', '*tumor*' sont des synonymes du concept '*neoplasm*'), les abréviations (AMP est synonyme de *Adenosime Monophosphate*), ou encore les variations lexicales tel que la différence dans le cas d'inflexion singulier-pluriel. Pour remédier à ce problème, plusieurs travaux se sont intéressés à l'expansion des documents.

L'expansion de documents est réalisée lors de la phase d'indexation et vise à accroître le degré de chevauchement des mots entre la requête des utilisateurs et les documents observés. Elle peut contribuer à renforcer la sémantique du document en élargissant le contenu du document avec les termes les plus informatifs.

Un schéma global de la RI dans le domaine biomédical est présenté dans la figure suivante :

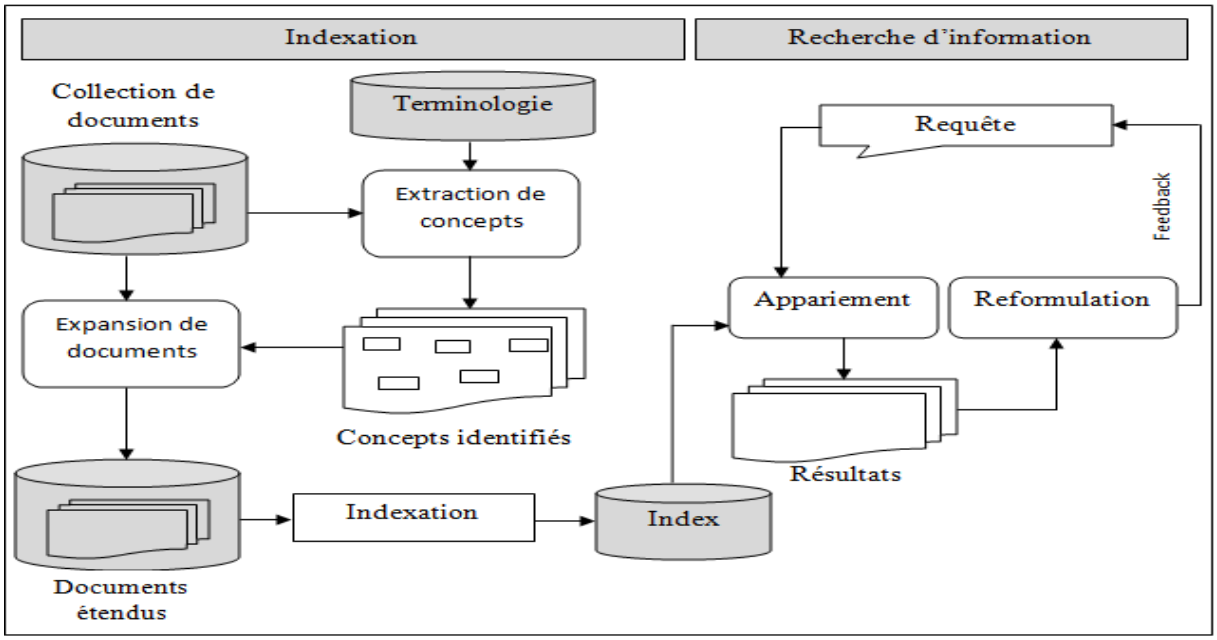

Figure I.4: Architecture générale du SRI dans le domaine biomédical

II.2.4. Evaluation des SRI biomédicaux

Il existe à ce jour deux campagnes d'évaluation en RI proposant des tâches dédiées à l'évaluation de la RI biomédicale : CLEF et TREC. Dans ce qui suit nous nous intéressons à la compagne TREC plus particulièrement à TREC Genomics pour la RI de la littérature biomédicale et au protocole d'évaluation TREC.

- **TREC Genomics :** La piste TREC Genomics, qui a duré de 2003 à 2007, est devenue une des pistes de recherche importantes dans le domaine biomédical, notamment les documents dans la base bibliographique de MEDLINE. Les documents de MEDLINE ont été extraits pour développer des collections tests dédiées à l'évaluation des approches de RI dans TREC Genomics. Au début de TREC Genomics en 2003, une sous-collection de MEDLINE qui couvre les données entre 2002 et 2003 a été extraite pour construire une collection d'évaluation à partir de 525,938 enregistrements de MEDLINE. Chaque enregistrement (appelé MEDLINE record) contient plusieurs champs importants pour les expérimentations en RI comme : .PMID (*PubMed Unique IDentifier*), .TI (*TItle*), .AB (*ABstract*), MH (*MeSH Headings*). Les requêtes ont été construites en se basant sur les besoins d'informations des scientifiques dans le domaine biomédical.

- **Protocole d'évaluation TREC :** Différentes mesures d'évaluation peuvent être utilisées dans un cadre d'évaluation par exemple :

- **La précision à *X*** premiers documents (dénotée *P@X*), est donc la proportion de documents pertinents par rapport aux *X* premiers documents renvoyés par le SRI. La précision à *X* mesure la satisfaction de l'utilisateur concernant les *X* premiers documents pertinents.

- **La MAP** (Mean Average Precision) correspond à la précision moyenne calculée sur l'ensemble des documents pertinents retournés. La MAP mesure la capacité du modèle d'appariement ou d'un SRI à pouvoir sélectionner les documents pertinents, en réponse à un ensemble de requêtes.

II. Conclusion

Dans ce chapitre nous avons introduit dans un premier temps la recherche d'information et son évolution à travers le processus d'indexation sémantique basé sur les ressources externes dont l'étape primordiale est la désambiguïsation, nous nous sommes ensuite intéressés à la recherche d'information dans le domaine biomédical à travers l'utilisation des ressources terminologiques externes, nous avons présenté deux de ces terminologies à savoir, le thésaurus MeSH qui sert à l'indexation de la littérature biomédicale et UMLS dont le principal objectif est de collecter et unifier les différentes ressources biomédicales. Et enfin nous avons présenté la compagne d'évaluation de SRI biomédicaux TREC.

La similarité sémantique est une notion très impliquée dans le processus de recherche d'information basé sur l'indexation sémantique notamment dans le processus de désambiguïsation dont les scores sont cumulés à partir de mesures de similarité sémantique. Ces mesures feront l'objet de notre prochain chapitre.

CHAPITRE II

Etat de l’art des mesures de similarité sémantique

I. Introduction

En recherche d'information sémantique, la notion de similarité sémantique est primordiale. En effet, les mesures de similarité sémantique jouent un rôle important, en particulier dans le processus de désambiguïsation des termes. L'objectif des mesures de similarité est d'estimer la ressemblance entre les concepts (auxquels les termes des requêtes et documents sont rattachés). Un concept se réfère à un sens particulier d'un terme donné. Dans ce contexte, plusieurs approches pour l'évaluation de la similarité sémantique entre concepts appartenant à une ontologie ou différentes ontologies (dans le cas d'alignement d'ontologies) ont été proposées dans la littérature. Ces approches se basent sur différents aspects, par exemple la structure hiérarchique (arborescente) de l'ontologie, le contenu informatif des différents concepts intégrant des mesures statistiques, ou sur les propriétés des concepts comparés. L'objectif de ce chapitre est de présenter les différentes mesures de similarité entre concepts. Pour cela, nous définissons la notion de similarité sémantique, puis nous donnons un état de l'art des différentes approches de calcul de similarité sémantique entre concepts.

I. La similarité sémantique

La similarité sémantique est une notion définie entre deux concepts soit au sein d'une même structure hiérarchique (ontologie), soit dans le cas d'alignement d'ontologies (les deux concepts à comparer appartenant respectivement à deux hiérarchies conceptuelles distinctes). La mesure de similarité sémantique est généralement définie par la fonction inverse de la distance sémantique. En effet, plus deux concepts sont similaires, moins ils sont distants.

- Une distance « dist » respecte les trois propriétés suivantes :
  - Nullité de la distance d'un concept avec lui-même : $dist(a,a) = 0$.
  - Symétrie : $dist(a,b) = dist(b,a)$.
  - Inégalité triangulaire: $dist(a,b) \leq dist(a,c) + dist(c,b)$.

  Où a, b et c sont des concepts quelconques.
- Une mesure de similarité est une fonction « sim » : $S^2 \rightarrow [0,1]$, avec S l'ensemble des concepts.

Les opinions divergent quant aux propriétés que devrait respecter une mesure de similarité. Tversky [Tversky, 77] a montré que les mesures de similarité conformes à la perception humaine ne satisfont pas toujours les propriétés mathématiques d'une mesure,

néanmoins il est pratiquement admis qu'une mesure de similarité doit être réflexive et symétrique :

- sim(x,x) = 1 : réflexivité
- sim(x,y) = sim(y,x) : symétrie.

Dans la littérature, différents approches de calcul des mesures de similarité sémantique ont été proposées. Ces approches ont été classées par A. Hliaoutakis [Hliaoutakis, 05] en quatre principales catégories qui constituent les approches basées sur le comptage d'arcs, les approches basées sur le contenu informatif, les approches basées sur les propriétés, et les approches hybrides. Ces approches peuvent sont appliquées pour évaluer la similarité entre deux concepts appartenant à une même ontologie ou pour comparer des concepts de deux ontologie différents pour certaines. Dans la section qui suit nous allons donner un état de l'art de ces différentes approches de calcul de similarité sémantique.

II. Etat de l'art

Plusieurs approches d'évaluation de la similarité entre concepts on été proposées, la plupart d'entre elles pour comparer des concepts appartenant au réseau lexical WordNet[1] qui répertorie, classifie et met en relation le contenu sémantique et lexical de la langue anglaise en couvrant la majorité des noms, verbes, adjectifs et adverbes de la langue anglaise, qu'il structure en un réseau de nœuds (constitués par des ensembles de termes synonymes appelés ***synsets***) et de liens (représentent les relations entre les synsets qui peuvent être des relations lexicales ou sémantiques) et/ou d'autres terminologies.

Cependant, l'approche de Rada et al. [Rada et al, 89] que nous allons présenter dans la section qui suit a été évaluée sur des concepts appartenant à l'ontologie du domaine biomédical MeSH et l'approche de Hliaoutakis qui sera présentée dans les approches hybrides a été mise en œuvre afin de comparer des concepts issus du thésaurus MeSH et des concepts de MeSH et Wordnet. Dans ce qui suit définir chacune des quatre catégories d'approches et citer les différents travaux réalisés dans leur contexte.

[1] http://wordnet.princeton.edu/

III.1. Approches basées sur le comptage d'arcs

L'idée principale de ces approches est de calculer la similarité sémantique entre concepts en se basant sur la structure hiérarchique de l'ontologie. En effet, dans ces approches la similarité est évaluée entre concepts selon le chemin qui les relie ou leurs positions dans la hiérarchie. Nous allons dans ce qui suit détailler ces approches.

III.1.1. Approche de Rada et al [Rada et al, 89]

Rada et al ont été, à notre connaissance, les premiers à suggérer que la similarité dans un réseau sémantique peut être calculée en se fondant sur les liens taxonomiques « is-a ».
Plus généralement, le calcul de la similarité entre concepts peut être fondé sur les liens hiérarchiques de spécialisation/généralisation. Dans cette approche, la distance entre deux concepts est calculée en comptant le nombre d'arcs (liens) séparant les concepts de la taxonomie. La définition de la similarité donnée par ces auteurs pour cette approche est ainsi établie : *« soient A et B deux concepts représentés par les nœuds a et b dans une hiérarchie « is-a », la distance conceptuelle entre A et B : dist(a,b), est représentée par le minimum du nombre d'arcs séparant les nœuds a et b »*. Formellement, cela voudrait dire que la distance entre les deux concepts est représentée par la longueur du plus court chemin séparant les nœuds qui les représentent dans la hiérarchie.
Les auteurs ont alors défini la similarité sémantique entre les concepts a et b comme étant l'inverse de la distance qui les sépare comme suit :

$$Sim(a,b) = \frac{1}{dist(a,b)}$$

III.1.2. Approche de Wu et Palmer [Wu et al., 94]

Wu et Palmer ont développé leur approche dans le cadre du traitement des langues plus exactement dans le contexte de la traduction automatique des verbes entre l'anglais et le mandarin chinois. Pour éviter les problèmes d'ambiguïté, leur mesure s'applique à un domaine conceptuel qui correspond à un point de vue donné pour lequel un mot a un seul sens. Les auteurs se sont basés sur l'ontologie WordNet pour calculer la similarité entre deux concepts associés à deux verbes différents.

L'idée principale de cette approche se base sur le petit généralisant commun aux deux concepts (Least Common Subsumer ou LCS) à comparer. C'est-à-dire le généralisant commun qui subsume les deux concepts et qui est le plus éloigné de la racine. Cette mesure prend en compte à la fois la profondeur du concept généralisant commun et les distances qui

séparent les concepts à comparer de leur généralisant commun. A titre d'illustration, soient les concepts de la figure II.1, la similarité entre c_1 et c_2 est donné par :

$$Sim_{W\&P}(c_1, c_2) = \frac{2 * N}{N_1 + N_2 + (2 * N)}$$

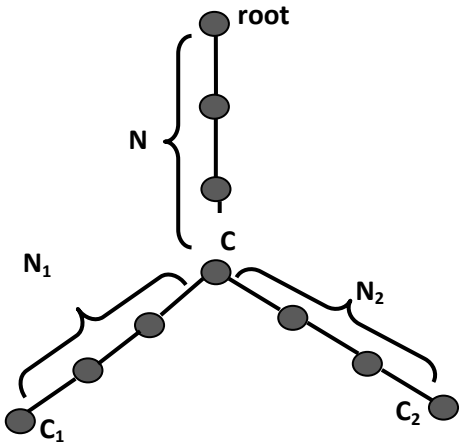

Figure II.1: Calcul de similarité en employant l'approche de Wu et Palmer

Formellement, cette mesure a été transformée par ses auteurs comme suit:

$$Sim_{W\&P}(c_1, c_2) = \frac{2 * profondeur(LCS)}{profondeur_{LCS}(c_1) + profondeur_{LCS}(c_2)}$$

Où :

- LCS est le petit généralisant commun de c_1 et c_2.
- Profondeur(LCS) est le nombre d'arcs minimal qui séparent LCS de la racine.
- « $profondeur_{LCS}(c_i)$ » est le nombre d'arcs qui séparent c_i de la racine en passant par LCS.

La profondeur du LCS est la profondeur globale qui permet de normaliser le calcul par rapport à la position des concepts dans la hiérarchie. Deux concepts identiques ont une similarité maximale de 1. Plus les concepts sont éloignés, plus la mesure décroît. Elle atteint la valeur nulle pour deux concepts qui ont la racine de la hiérarchie comme petit généralisant commun.

III.1.3. Approche de Leacock et Chodorow [Leacock et al, 98]

La mesure de Leacock et chodorow est une mesure basée sur le chemin qui relie les nœuds entre eux. Elle dépend de la longueur du plus court chemin entre concepts dans une hiérarchie « is-a ». Le plus court chemin est celui qui comprend le plus petit nombre de nœuds intermédiaires.

La similarité entre les noms est alors inversement proportionnelle à la profondeur maximale de la hiérarchie WordNet notée « D » qui représente la taille du plus long chemin de la feuille au nœud racine dans la hiérarchie. Cette mesure est alors définie comme suit :

$$\boldsymbol{Sim(c_1, c_2) = -log \frac{(length(c_1, c_2))}{2 * D}}$$

Où length (c1, c2) représente le plus court chemin entre deux nœuds, et D la profondeur maximale dans la taxonomie (elle est égale à 16 pour WordNet 1.7). Dans l'exemple suivant figure II.2, le plus court chemin entre *credit card* et *medium of exchange* est celui qui passe par le nœud *credit*. La mesure est alors calculée comme suit :

Sim (credit card, medium of exchange) = - log (1 / (2*16))

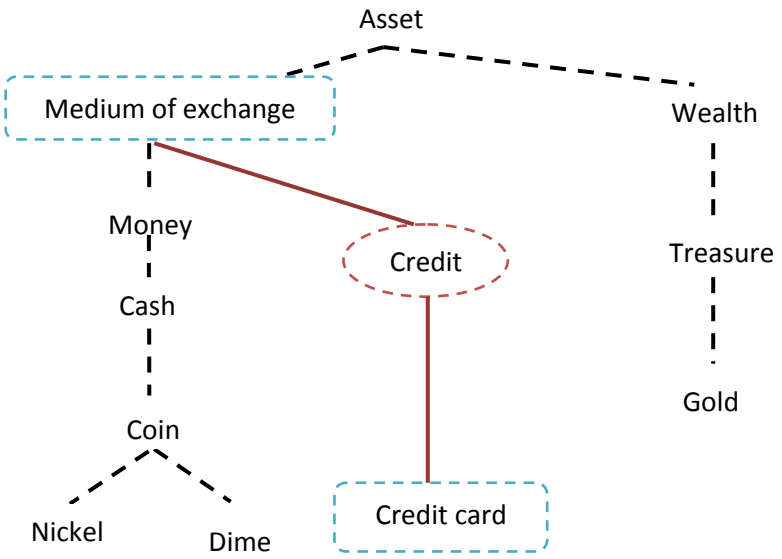

Figure II.2: Exemple de taxonomie WordNet

III.1.4. Approche de Hirst et St-Onge [Hirst et al., 98]

Hirst et St-Onge calculent la proximité sémantique qui est une notion plus large que la similarité sémantique. En effet, dans leur approche, les auteurs prennent en compte toutes les relations qui existent dans WordNet. Le cadre de leur approche est l'identification de chaînes

lexicales dans le but de détecter les malapropismes[2]. Ils s'intéressent donc aux mots, et pas aux concepts, et c'est bien entre ceux-là qu'ils attribuent une valeur de proximité.
Pour cela, les auteurs font un classement de toutes les relations de WordNet en tant qu'horizontales, ascendantes et descendantes comme suit :

- **Les relations ascendantes** lient des concepts plus spécifiques à des concepts plus généraux. Il s'agit des relations d'hyperonymie et meronymie. La relation « is-a » en est un exemple.
- **Les relations descendantes** au contraire, relient des concepts plus généraux à des concepts plus spécifiques. Il s'agit des relations d'hyponymie et holonymie.
- **Les relations horizontales** représentent toutes les autres, c'est-à-dire celles qui maintiennent le degré de spécificité. Par exemple, la relation *antonymie* en est une.

En plus de ce classement des types de relations, ils définissent un degré de relation : *très-fort (extra-strong), fort (strong), moyen-fort (medium-strong)* et *faible*. Ceux-ci sont définis comme suit :

- **La relation très-forte :** Elle se tient entre un mot et lui-même. Les mots qui ont ce degré de relation ont une similarité constante de 3*T. (T étant une constante).
- **La relation forte :** Elle se tient entre deux mots dans WordNet. Une relation forte est considérée entre deux mots s'ils vérifient l'une des trois conditions établies préalablement par les auteurs est satisfaite. Les mots qui ont ce degré de relation ont une similarité constante de 2*T. (T étant une constante). Les trois conditions sont décrites comme suit :
  - La première condition est que les mots appartiennent au même synset. Par exemple, les deux mots « *person* » et « *human* » sont fortement similaires, car ils appartiennent tous les deux à un même synset dans WordNet même s'ils existent séparément dans d'autres synsets comme montré dans la figure suivante :

[2] Termes se rapprochant ou étant identiques phonétiquement à un terme correct, mais n'ayant pas le sens voulu par le contexte.

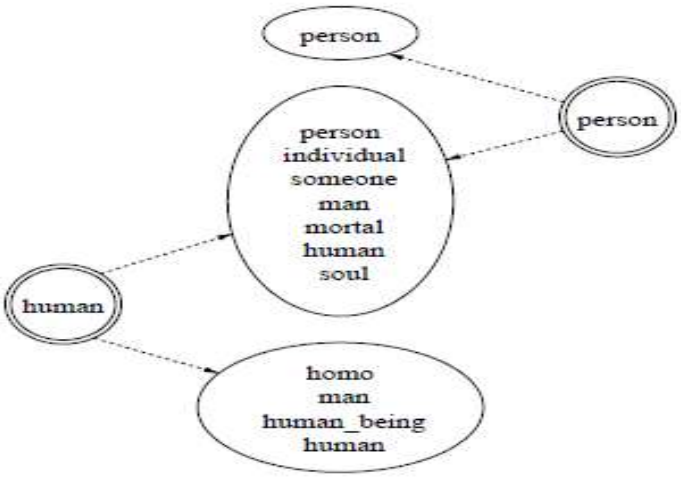

Figure II.3 : synsets contenant les concepts « *person* » et « *human* » dans WordNet

- La seconde condition est que les deux mots se trouvent dans deux synsets différents, ces deux synsets doivent être reliés par une relation horizontale. Par exemple, Les deux mots de la figure suivante « *precursor* » et « *successor* » sont fortement liés car ils appartiennent à deux synsets différents et deux de ces synsets sont reliés par une relation horizontale comme montré dans la figure suivante :

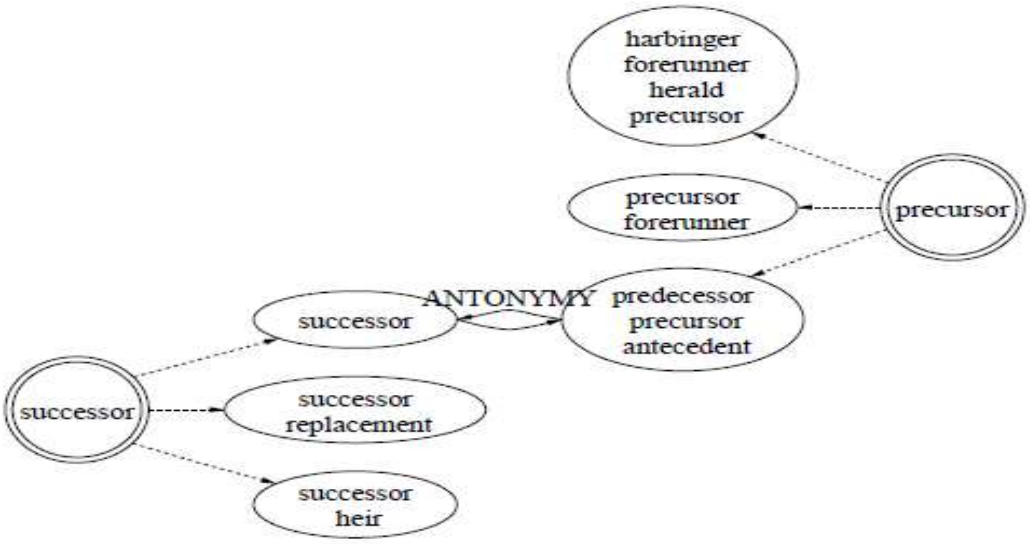

Figure II.4: synsets contenant les concepts « *precursor*» et « *successor* » dans WordNet

- La troisième condition est que si l'un des deux mots est un mot composé, et que l'autre est une partie de celui-ci et qu'il existe n'importe quelle relation entre les deux synsets qui les contiennent. Par exemple, dans la figure qui suit, les deux mots « *school* » et « *private_school* » sont fortement liés car le mot « *private_school* » est composé du mot « *school* ».

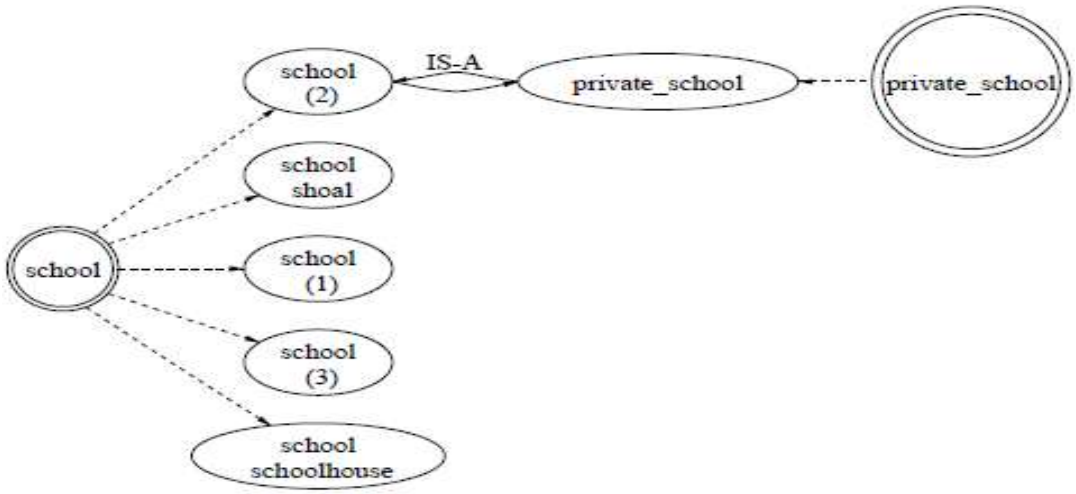

Figure II.5 : synsets contenant les concepts « *private_school* » et « *school* » dans WordNet

- **La relation Moyen–fort** : les auteurs proposent une série de chemins « acceptables » entre deux synsets, où les chemins sont composés de deux à cinq liens et les relations dénotant les liens d'un chemin sont différentes (ascendantes, descendantes ou horizontales) comme illustré dans la figure suivante:

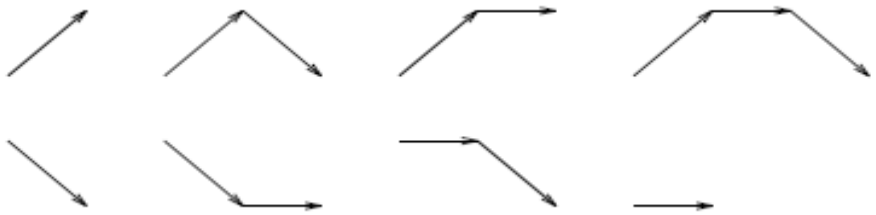

Figure II.6: Série des chemins acceptables proposée par Hirst et St-Onge

Les auteurs ont supposé que si les chemins reliant les deux synsets contenant les deux mots étaient formés par des arcs suivant l'orientation proposée alors les deux mots sont moyennement similaires. L'idée est que deux mots sont proches sémantiquement si leurs synsets sont connectés par un chemin qui n'est pas très long et qui ne change pas souvent de direction.

A la différence des deux autres relations, un poids est attribué à celle-ci. Ce poids sert à calculer la proximité entre les mots et représente le poids du plus court chemin qui mène du synset du mot à un autre. Il est calculé comme suit :

$$weight = C - pathlength - (k * number\ of\ changes\ of\ directions)$$

Où :

- « number of change of directions » représente le nombre de changements de directions dans le chemin reliant les deux concepts à comparer.
- « pathlength » représente la longueur du chemin reliant les deux concepts.
- C et k représentent des constantes.

- **La relation faible :** Elle est attribuée pour tous les autres cas, elle signifie selon les auteurs que les deux mots sont trop éloignés. Les mots qui ont ce degré de relation ont donc une similarité nulle.

III.2. Approches basées sur le calcul du contenu informatif

Les approches vues précédemment se basent sur la structure hiérarchique des concepts dans l'ontologie afin d'évaluer la similarité entre eux. Dans cette section, des mesures qui intègrent les statistiques sur les corpus comme paramètre pour l'évaluation des similarités sémantiques sont présentées. En effet, la connaissance rapportée par l'analyse de corpus est utilisée pour compléter les informations déjà présentes dans les ontologies ou taxonomies. Dans la littérature, les approches basées sur le contenu de l'information sont également nommées approches basées sur le corpus ou approches basées sur la théorie de l'information. Ces approches emploient la notion du contenu de l'information (IC), qui peut être considérée comme une mesure de quantification de la quantité d'informations exprimée par un concept. Ces approches utilisent des corpus pour calculer les valeurs de IC nécessaires en associant des probabilités à chaque concept dans la taxonomie. Ces probabilités sont basées sur les occurrences de mots dans les corpus.

III.2.1. Calcul du contenu de l'information (IC)

La notion du contenu de l'information (ou contenu informatif) « IC » a été introduite pour la première fois par Resnik [Resnik, 99] qui a suggéré que le contenu informatif d'un concept traduit la pertinence de celui-ci dans un corpus en tenant compte de la fréquence d'apparition des mots auxquels il se réfère ainsi que la fréquence d'apparition des concepts qu'il généralise. Plus précisément le contenu informatif se calcule par la formule suivante :

$$CI(c) = -\log(p(c))$$

Où $p\ (\ c\ )$ est la probabilité de retrouver qu'un mot du corpus soit une instance du concept c (un des mots référés par le concept c ou par un de ses descendants).

Dans les expérimentations de Resnik [Resnik, 99], ces probabilités sont calculées par la formule suivante :

$$p(c) = \frac{frequence(c)}{N}$$

Où :

- « N » est le nombre total de concepts et
- "$frequence(c) = \sum_{w \in instance(c)} count(w)$" avec instance (c) qui représente l'ensemble des termes possibles pour le concept c mais également pour l'ensemble de ses descendants dans la hiérarchie.

Plus un concept est général, plus son contenu informatif est faible. A l'inverse, plus le concept est spécifique plus son contenu informatif est important.

Une deuxième approche est proposée par Seco et al [Seco et al, 04] qui pensent que l'utilisation d'un corpus est un défaut de l'approche de Resnik [Resnik, 99]. En effet, selon les auteurs, WordNet seul suffit pour trouver le contenu informationnel des nœuds. Leur thèse est qu'il est possible de retirer de la structure de WordNet un sens du nombre d'hyponymes qu'a un concept : plus il dispose de descendants, moins il contient d'information (les concepts généraux sont moins spécifiques que les concepts se retrouvant aux niveaux bas dans la hiérarchie). Pareillement, les feuilles de la taxonomie ont une valeur informative maximale, car elles sont les plus spécialisées. Leur calcul du contenu informationnel de chaque nœud est présenté dans la formule suivante :

$$IC(c) = \frac{\log\left(\frac{hypo(c)+1}{max_{wn}}\right)}{\log\left(\frac{1}{max_{wn}}\right)} = 1 - \frac{\log(hypo(c)+1)}{\log(max_{wn})}$$

Avec :

- hypo(c) qui indique le nombre d'hyponymes dont dispose le concept c,
- max_{wn} est une constante qui indique le nombre de concepts de la taxonomie.

Le dénominateur, qui représente le concept avec la plus forte valeur informative, permet d'assurer que les valeurs sont incluses dans l'intervalle [0,1].

De plus, cette formule permet d'assurer que le contenu informationnel ainsi défini croit de façon monotone depuis la racine, qui a une valeur de CI égale à 0, jusqu'aux feuilles qui ont une valeur du contenu informationnel égale à 1 comme voulu.

III.2.2. Approches de calcul de similarité utilisant le contenu informatif IC

Plusieurs auteurs ont mis en place des approches basées sur le contenu informatif des concepts dans une hiérarchie afin d'évaluer la similarité entre eux. L'intuition derrière l'utilisation de la notion du contenu informatif est que la similarité entre deux concepts est la portion d'information qu'ils ont en commun qui, dans le cadre d'une ontologie, peut être déterminée par le plus petit généralisant commun qui les subsume (*LCS*). Cette intuition est indirectement appliquée par les mesures présentées dans la section précédente qui calculent la similarité avec le nombre d'arcs qui séparent deux concepts.

III.2.2.1. Approche de Resnik [Resnik,99]

Resnik définit la similarité sémantique entre deux concepts par la quantité d'information qu'ils partagent. Cette information partagée est évaluée numériquement par le contenu informatif du plus petit généralisant commun aux deux concepts comparés (LCS) comme suit :

$$Sim(c_1, c_2) = CI(LCS(c_1, c_2))$$

Ainsi, si deux concepts sont très éloignés et ont comme racine le LCS, leur similarité est nulle.

III.2.2.2. Approche de Jiang et Conrath [Jiang et al.,97]

L'approche de Jiang et Conrath palie aux limites de la mesure de Resnik présentée précédemment en combinant le contenu informatif du LCS aux contenus informatifs des deux concepts comparés. En effet, les auteurs proposent une distance qui est non seulement dépendante du chemin parcouru, mais aussi, de la quantité d'information véhiculée par les nœuds traversés par ce chemin. Le calcul de la longueur du chemin entre les deux concepts s'appuie sur la recherche du plus petit généralisant commun subsumant les deux concepts LCS(c_1,c_2). Le plus court chemin entre c_1 et c_2 sp(c_1,c_2) est alors l'unique chemin passant par LCS(c_1,c_2). Ensuite pour chaque arête de la hiérarchie reliant deux concepts x et y, les auteurs proposent de définir un poids TC(x,y) ϵ [0,1]. La distance de Jiang et Conrath a été alors définie comme suit :

$$dist(c_1, c_2) = \sum_{c \in SP(c_1,c_2) \setminus LCS(c_1,c_2)} [IC(c) - IC(parent(c)] * TC(c, parent(c))$$

Selon [Budanitsky et al, 06], cette mesure de distance est probablement la plus utilisée actuellement et la plus efficace pour déterminer la proximité sémantique entre deux concepts. Comme la plupart des mesures de similarités elle se limite à la relation de subsomption, mais elle offre la possibilité de prendre en compte des valeurs différentes pour chaque arête de la hiérarchie des concepts, les TC(c, parent(c)). Une fonctionnalité qui a rarement été prise en compte. En effet, dans l'article initial, les auteurs ont effectué leur évaluation avec un poids d'arête constant (TC=1). Depuis, les systèmes utilisant la formule de Jiang et Conrath se servent de la version simplifiée sans poids sur les arêtes formulée pas les auteurs comme suit :

$$dist_{JC_{Simple}}(c_1, c_2) = (IC(c_1) * IC(c_2)) - 2 * IC(LCS(c_1, c_2)$$

La similarité entre les deux concepts est alors calculée par l'inverse de la distance qui les sépare.

III.2.2.3. Approche de Lin [Lin, 1993]

Les auteurs de cette approche proposent d'évaluer la similarité entre deux concepts en tenant compte à la fois de leur contenu d'information (IC), et le contenu d'information de leur concept commun le plus spécifique comme suit :

$$Sim_{Lin}(c_1, c_2) = \frac{2 * IC(LCS(c_1, c_2))}{IC(c_1) + IC(c_2)}$$

Ceci peut être vu comme le contenu d'information de l'intersection des deux concepts (multiplié par deux) qui est divisé sur leur somme.

Les valeurs retournées par cette mesure varient entre 1 (concepts complètement similaires) et 0. Dans ce cas, la mesure d'un concept comparé à lui-même aura la valeur 1.

Cette mesure semble combiner les propriétés des mesures présentées ci-dessus, c'est à dire nous offre à la fois des informations sur la taille de l'ontologie et le classement des paires de termes différents.

III.3. Approches basées sur les propriétés des concepts

Les approches qui s'appuient sur les propriétés des concepts s'appuient sur une base qualitative ; elles sont fondées sur la comparaison du nombre de propriétés communes par rapport au nombre total de propriétés. Etablissant leur base sur la théorie ensembliste, ces approches produisent une mesure de similarité qui n'est pas nécessairement une métrique car

les propriétés de symétrie et de transitivité ne sont pas toujours respectées. Dans ce qui suit nous allons introduire une approche proposée par A.Tversky se basant sur les propriétés que peut avoir un concept dans une ontologie tel que les attributs.

III.3.1. Approche de Tversky [Tversky, 77]

L'approche de Tversky prend en compte le nombre de propriétés communes et les différences entre les deux concepts à comparer. En effet, cette approche est basée sur la description des concepts. Selon l'auteur, la similarité entre deux concepts est évaluée par le nombre pondéré de propriétés en commun auquel est retiré le nombre de propriétés spécifiques à chacun des concepts. Ce qui voudrait dire que, plus les concepts ont de propriétés communes et moins de non communes, plus ils sont similaires. La mesure a été alors exprimée par la formule suivante :

$$Sim_{Tversky}(c_1, c_2) = \alpha.comm(c_1, c_2) - \beta.diff(c_1, c_2) - \gamma.diff(c_2, c_1)$$

Où : α, β, γ sont des constantes qui mènent à différents types de similarités.

Dans cette mesure, si $\beta = \gamma = 0$ et $= 1$, la similarité entre c_1 et c_2 correspond à la quantité de propriétés en commun. Si $\alpha = 0$, $\beta > 0$ et $\gamma > 0$, les concepts c_1 et c_2 sont évalués selon ce qui les différencie ce qui en fait une mesure de dissimilarité non de similarité.

Cette mesure est donc dépendante du cardinal des propriétés de chaque concept. Une seconde version de la formule précédente a été donnée par son auteur dans [Tversky, 77] comme suit :

$$Sim_{Tversky}(c_1, c_2) = \frac{|C_1 \cap C_2|}{|C_1 \cap C_2| + \beta|C_1 \backslash C_2| + \gamma|C_2 \backslash C_1|}$$

Où C_1 et C_2 sont les ensembles de descriptions (propriétés) de c_1 et c_2 respectivement. β et γ définissent l'importance relative des propriétés non communes.

III.4. Approches hybrides

Les approches de cette catégorie combinent certaines propriétés des approches présentées ci-dessus. En effet, elles peuvent tenir compte des relations is-a qui séparent les concepts ou les propriétés qu'elles peuvent partager ou non. Dans ce qui suit nous allons présenter deux approches, l'une proposées par Rodriguez [Rodriguez, 00] utilisée pour comparer des classes d'entités entre elles et l'autre proposée par Hliaoutakis [Hliaoutakis, 05] qui est utilisée pour comparer des concepts soit appartenant à une même ontologie ou à des ontologies différentes. Nous allons ensuite présenter une approche mise en œuvre dans le but

d'aligner des concepts appartenant à plusieurs ontologies mais pouvant être adaptée dans le cas d'une seule ontologie.

III.4.1. Approche de Rodriguez [Rodriguez, 00]

L'auteur de cette approche a appliqué la mesure proposée par Tversky [Tversky, 77] vue précédemment à la comparaison de classes d'entités (qui sont des classes conçues pour faciliter l'accès à des données qui ont les mêmes particularités (entités)). Dans cette approche, les auteurs combinent l'approche de Tversky, ainsi que des facteurs qui tiennent compte du contexte, pour produire une mesure de similarité qui associe des entités spatiales appartenant à la même ontologie (modèle Matching Distance) ou à des ontologies différentes (Modèle Triple Matching Distance). Ce modèle tient compte du fait que l'évaluation de la similarité s'effectue dans un domaine de discours, lequel définit un contexte et que, par conséquent, la mesure de similarité entre deux entités de classe est dépendante du contexte.

❖ **Modèle Matching Distance**

Le modèle Matching Distance s'applique à des ontologies où les classes sont organisées taxonomiquement et sont liées par des relations de généralisation (*is-a*) ou des relations d'agrégation (*part-of*). Dans le premier modèle Matching Distance, la similarité entre les entités de classes est évaluée par une somme pondérée des similarités calculées selon les trois catégories de propriétés : les *attributs*, qui caractérisent intrinsèquement les entités, les *parties* qui peuvent être des propriétés intrinsèques ou des éléments ayant une relation *part-of* avec les entités, et les *fonctions* qui caractérisent le comportement et le rôle des entités. Cette mesure de similarité et donnée par la formule suivante :

$$S(c_1, c_2) = \omega_a S_a(c_1, c_2) + \omega_p S_p(c_1, c_2) + \omega_f S_f(c_1, c_2)$$

Les similarités sont pondérées par les poids ω_a pour la similarité des attributs, ω_p pour la similarité des parties et ω_f pour la similarité des fonctions ; ces poids ont pour rôle d'ajuster la similarité en fonction du contexte. Ce modèle offre donc deux possibilités pour l'évaluation du poids de chaque type de propriétés, selon la conception que l'utilisateur peut avoir de la pertinence dans un contexte donné. La similarité selon chaque propriété est donnée par la forme suivante de la similarité du modèle de Tversky:

$$S_{a/p/f}(c_1, c_2) = \frac{|C_1 \cap C_2|}{|C_1 \cap C_2| + \alpha|C_1 \backslash C_2| + (1-\alpha)|C_2 \backslash C_1|}$$

Le modèle Matching Distance intègre la distance entre les concepts dans le graphe de la taxonomie à l'intérieur du paramètre α, lequel est une fonction de la profondeur d_1 du concept c_1 et de la profondeur d_2 du concept c_2 par rapport à la racine de la taxonomie comme suit:

$$\alpha = \begin{cases} \dfrac{d_1}{d_1 + d_2} \; si \; d_1 \leq d_2 \\ 1 - \dfrac{d_1}{d_1 + d_2} \; sinon \end{cases}$$

❖ **Modèle Triple Matching Distance**

Dans l'extension du modèle Matching Distance pour plusieurs ontologies (modèle Triple Matching Distance), les graphes des ontologies sont liés par une superclasse commune afin de pouvoir calculer la distance entre les concepts dans le graphe global. La similarité de voisinage des concepts dans le graphe et la similarité lexicale (similarité entre les noms des concepts) sont ajoutées à la somme pondérée du modèle Matching Distance.

Le voisinage d'un concept est l'ensemble des entités de classe qui se trouvent à un rayon r (défini selon les besoins) du concept dans le graphe. Pour un rayon assez grand, les voisinages peuvent différer de manière significative et la similarité des voisinages sera faible ; à l'opposé, un rayon faible augmente la probabilité d'avoir un voisinage contenant plus d'éléments communs. Le choix du rayon r dépendra aussi de la taille du graphe de l'ontologie.

L'évaluation de la similarité des voisinages V_1 et V_2 de deux concepts prend une forme similaire au modèle de Tversky comme suit:

$$S(c_1, c_2) = \frac{|V_1 \cap V_2|}{|V_1 \cap V_2| + \alpha(c_1, c_2) * \delta(c_1, V_1 \cap V_2) + \big(1 - \alpha(c_1, c_2)\big) * \delta(c_2, V_1 \cap V_2)}$$

L'intersection entre les voisinages est approximée par la similarité maximale des entités de classe entre les voisinages:

$$V_1 \cap V_2 = \left[\sum_{i<n} max_{j<m} \, S(V_{1_i}, V_{2_j})\right] - \vartheta S(c_1, c_2)$$

Où : V_{1_i}est un élément de V_1 de cardinalité n

V_{2_j}est un élément de V_2 de cardinalité m

$$\vartheta = \begin{cases} 1 \; si \; S(c_1, c_2) = \max S\left(c_1, V_{2_j}\right) pour \; j \leq m \\ 0 \; sinon \end{cases}$$

Le modèle Triple Matching Distance englobe donc plusieurs aspects (contexte, propriétés des concepts, distance dans les graphes, similarité des voisinages dans les graphes et similarité lexicale) qui peuvent influencer la similarité, au lieu de se concentrer sur un seul aspect des concepts qui peut être plus ou moins représentatif.

III.4.2. Approche de Hliaoutakis [Hliaoutakis, 05]

En se basant sur le modèle Triple Matching Distance de Rodriguez, Hliaoutakis [Hliaoutakis, 05] a mis en oeuvre une nouvelle approche qui est une adaptation de l'approche précédente pour la comparaison de concepts appartenant au thésaurus MeSH et/ou concepts appartenant l'un à MeSH l'autre à WordNet. En effet, en plus du voisinage, l'auteur s'est basé sur les synsets de WordNet (ou les termes synonymes de MeSH) et les définitions (« glosses » dans WordNet et « Scope Note » dans MeSH) pour définir sa similarité entre concepts.

Selon l'auteur, deux concepts sont semblables si leurs synsets ou leurs ensembles de descripteurs ou les synsets des termes dans leur voisinage (plus spécifiques ou plus généraux) sont semblables. Il a donc procédé à la modification de la mesure précédente comme suit :

- Premièrement, l'auteur a remplacé les caractéristiques prises en compte par les auteurs de l'approche précédente [Rodriguez, 00] définis par la fonction S_u par une fonction S_{gloss} qui prend en compte ce glossaire et qui est calculée comme suit :

$$S_{gloss}(a^p, b^q) = \frac{|A \cap B|}{|A \cup B|}$$

  Avec : A et B des ensembles contenant des termes formant les glossaires (définitions) des concepts a^p et b^q.

- Deuxièmement, l'auteur a proposé de calculer la similarité de voisinage avec un rayon r=1 en utilisant la fonction suivante :

$$S_{neigh}(a^p, b^q) = \max_i \frac{|A_i \cap B_i|}{|A_i \cup B_i|}$$

Avec « i » qui dénote les relations d'hyponymie et d'hyperonymie des concepts composant les ensembles A et B relatifs aux voisins respectivement de a et b dans les ressources p et q respectivement. En d'autres termes l'auteur propose de prendre la similarité maximum entre les deux concepts dans une part du voisinage mais pas dans toute la ressource. Cette similarité étant calculée par la somme des similarités selon les synonymes et les définitions des concepts.

Donc en utilisant ces deux fonctions et en se basant sur la mesure de similarité proposée par Rodriguez, l'auteur a proposé la mesure de similarité suivante pour le thésaurus MeSH :

$$S_{Hliaoutakis}(a^p, b^q) = \omega_{Syn} S_{Syn}(a^p, b^q) + \omega_{gloss} S_{gloss}(a, b) + \omega_{neigh} S_{neigh}(a^p, b^q)$$

III.4.3. Approche de Al-Mubaid et Nguyen [Al-Mubaid et al., 09]

Al-Mubaid et Nguyen ont proposé nouvelle une mesure basée sur la structure de l'ontologie pour mesurer la similarité sémantique entre des concepts appartenant à plusieurs ontologies contenant un ou plusieurs nœuds communs comme dans la figure II.8. Les auteurs ont utilisé MeSH et SNOMED dans le cadre du système de langue médical unifié (UMLS) pour définir leur mesure de similarité entre deux concepts. Dans cette figure des deux fragments des deux ontologies MeSH et SNOMED, le concept (nœuds) commun qui se trouve dans les deux est « Abdominal pain ».

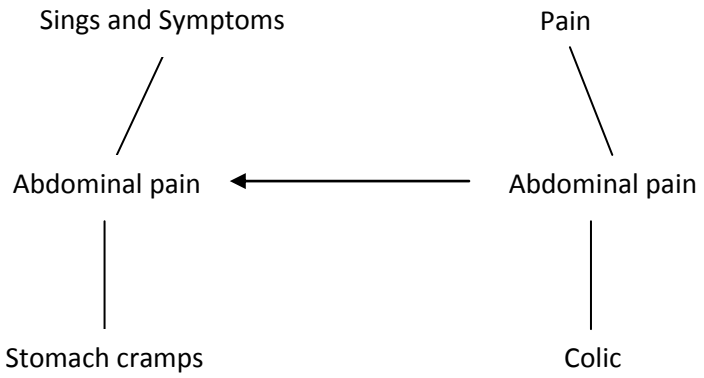

Figure II.7 : fragments des SNOMED (à gauche) et MeSH (à droite)

Les auteurs ont tout d'abord introduit une nouvelle caractéristique qui est la spécificité commune entre deux concepts comme suit :

- **Nouvelle spécificité commune**

Les auteurs définissent la spécificité commune entre deux concepts en utilisant la profondeur de leur plus petit concept subsumant (LCS) par rapport à la profondeur globale de la hiérarchie comme suit :

$$CSpec(c_1, c_2) = D - Depth(LCS(c_1, c_2))$$

Où: D est la profondeur de l'ontologie.

Depth (LCS(c_1,c_2)) est la profondeur du concept commun le plus spécifique entre c_1 et c_2.

Selon Al-Mubaid et Nguyen, la similarité sémantique entre les concepts appartenant à différentes ontologies (ontologies croisées) et ayant quelques concepts en commun, est mesurée en considérant deux types d'ontologies :

- *Ontologie primaire* : celles qui ont le plus de concepts.
- *Ontologie secondaire* : celles qui ont le moins de concepts.

Pour mesurer la similarité entre concepts dans ce cas, la longueur du chemin séparant les concepts à comparer et la profondeur des ontologies sont prises en compte. Les auteurs ont alors soulevé quatre cas possibles comme suit : les concepts comparés appartiennent tous les deux à l'ontologie primaire (cas 1), l'un des concepts appartient à l'ontologie primaire et l'autre à la secondaire (cas2), deux concepts comparés appartiennent à l'ontologie secondaire (cas 3), les concepts appartiennent à de multiples ontologies secondaires (cas 4).

Alors d'après les auteurs, la similarité sémantique entre deux concepts est estimée pour les quatre cas en suivant l'un des processus suivant :

- **Cas 1 :** Dans le cas où les deux concepts sont dans la même ontologie (primaire), les caractéristiques de longueur du chemin et de profondeur sont employées pour obtenir la distance sémantique entre les deux concepts comme suit :

$$SimDist(c_1, c_2) = \log\big((path - 1)^{\alpha} * (CSpec)^{\beta} + k\big)$$

  Où $\alpha>0$ et $\beta>0$ sont des facteurs de contribution des deux caractéristiques respectivement et $k \geq 0$ une constante qui garantit que le résultat sera positif.

- **Cas 2 :** Dans ce cas, l'un des concepts appartient à l'ontologie primaire et l'autre à la secondaire, la similarité sémantique est alors mesurée en suivant les caractéristiques suivantes :
  - La spécificité commune des deux concepts (CSpec).
  - La longueur du chemin entre les deux concepts.
  - La densité des deux types d'ontologies.

Dans le cas où les deux concepts appartiennent à deux ontologies différentes, les LCS entre les deux concepts est calculé en utilisant un nœud qui assemble les nœuds communs aux deux ontologies comme dans la figure suivante :

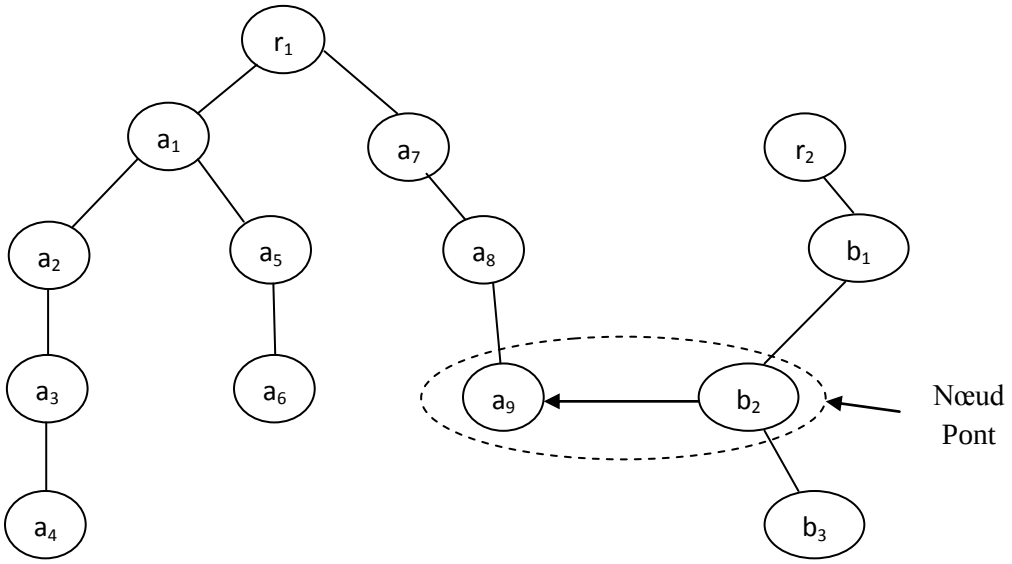

Figure II.8: deux fragments d'ontologies connectés par un nœud pont

Dans cette figure deux fragments d'ontologies sont pris en compte avec la racine de l'ontologie primaire (celle qui contient le plus grand nombre de nœuds et donc de concepts) qui est r_1, et la racine de l'ontologie secondaire (avec le moins nombre de nœuds) avec la racine r_2. L'ontologie secondaire est reliée à l'ontologie primaire en joignant les nœuds communs en un seul nœud appelé « nœud pont ».

Selon les auteurs le plus petit concept père commun (LCS) est représenté par le LCS entre le concept c_1 (appartenant à l'ontologie primaire) et le nœud pont :

$$LCS(c_1, c_2) = LCS(c_1, bridge_n)$$

La longueur du chemin entre les deux concepts est calculée par la somme des longueurs des plus courts chemins reliant le concept c_1 au nœud pont ($bridge_n$) « d_1 », et le concept c2 au nœud pont ($bridge_n$) « d_2 » comme suit :

$$Path(c_1, c_2) = d_1 + d_2 - 1$$

Mais il est à remarquer que les longueurs des chemins sont sur des échelles différentes vu que dans l'ontologie primaire il y a plus de nœuds que dans l'ontologie secondaire et donc, le chemin vers le nœud pont sera probablement plus long que dans l'ontologie secondaire, les auteurs ont donc proposé une autre façon de calculer la longueur du chemin en utilisant les profondeurs des deux ontologies comme suit :

$$PathRate = \frac{2D_1 - 1}{2D_2 - 1}$$

Où D_1 et D_2 représentent les profondeurs des ontologies primaire et secondaire respectivement.

La spécificité commune est elle aussi adaptée à ce cas, elle est calculée en utilisant les profondeurs des ontologies primaire et secondaires D_1 et D_2 comme suit :

$$CSpecRate = \frac{D_1 - 1}{D_2 - 1}$$

Par ailleurs, il peut y avoir plusieurs nœuds ponts qui relient les deux ontologies, il peut y avoir plusieurs longueurs de chemins $Path_i$, dans ce cas les auteurs ont introduit la distance sémantique entre les deux concepts à comparer comme suit :

- Tout d'abord, la spécificité commune aux deux concepts est calculée comme suit :

$$CSpec_i(c_1, c_2) = D_1 - Depth(LCS(C_1, Bridge_i))$$

- Ensuite, la similarité sémantique entre les deux concepts est mesurée avec la formule suivante :

$$SemDist_i(c_1, c_2) = \log\big((path_i - 1)^{\alpha} * (CSpec_i)^{\beta} + k\big)$$

- **Cas 3 :** Dans ce cas, les deux concepts comparés appartiennent à l'ontologie secondaire. Par conséquent selon les auteurs, la distance sémantique entre ces concepts peut être calculée seulement quand le chemin les reliant Path (c_1, c_2) et leur spécificité commune CSpec(c_1,c_2) sont traduits à l'échelle de l'ontologie primaire. Ces deux caractéristiques sont alors données par les formules suivantes :

$$Path(c_1, c_2) = Path(c_1, c_2)_{secondary} * PathRate$$
$$CSpec(c_1, c_2) = CSpec(c_1, c_2)_{secondary} * CSpecRate$$

  Où : Path (c_1, c_2)$_{secondary}$ est le plus court chemin entre c_1 et c_2 dans l'ontologie secondaire.

  CSpec (c_1, c_2)$_{secondary}$ est calculé en se basant sur l'ontologie secondaire en utilisant l'équation ($CSpec(c_1, c_2) = D - Depth(LCS(c_1, c_2))$), La distance sémantique entre les concepts dans l'ontologie secondaire est alors calculée par :

$$SimDist(c_1, c_2) = \log\big((path - 1)^{\alpha} * (CSpec)^{\beta} + k\big)$$

- **Cas 4 :** les concepts appartiennent à deux ontologies secondaires différentes, parmi lesquelles une agit temporairement comme une ontologie primaire, et l'autre agit comme une ontologie secondaire. Pour cela, l'ontologie secondaire qui a le plus grand nombre de concepts est choisie comme ontologie primaire. La mesure de similarité sémantique est alors calculée en utilisant la longueur du chemin entre les deux concepts dans l'ontologie secondaire et leur spécificité commune CSpec en les ramenant à l'échelle de l'ontologie primaire suivant les formules données dans le cas 3.

III. Conclusion

Dans ce chapitre, nous avons introduit la notion de similarité sémantique entre concepts qui est une notion très importante dans le domaine de la recherche d'information. Nous avons ainsi fait un état de l'art des diverses approches qui existent pour le calcul de cette similarité. Dans le chapitre suivant nous nous intéresserons au calcul des similarités sémantiques entre concepts du domaine biomédical, nous exploiterons ainsi la hiérarchie offerte par ce thésaurus et les relations offertes par le réseau sémantique UMLS pour proposer et évaluer une mesure de similarité sémantique entre concepts biomédicaux.

CHAPITRE III

Proposition d’une mesure de similarité sémantique

I. Introduction

Dans le chapitre précédent, nous avons présenté les différents travaux liés aux mesures de similarité sémantique entre concepts d'une ontologie. La plupart des approches existantes ont été conçues pour évaluer la similarité sémantique entre mots dans WordNet, une ontologie du domaine général. Ces mesures se basent sur la structure de la ressource utilisée, sur la notion de contenu informatif ou sur les propriétés qu'offrent ces ressources. Dans le domaine biomédical, une tentative de mise en place d'une mesure de similarité sémantique pour la comparaison de concepts issus du thésaurus MeSH a été avancée celle-ci se base principalement sur les propriétés de ces concepts.

Notre travail s'inscrit dans ce contexte et vise à proposer et à implémenter une mesure de similarité sémantique entre concepts biomédicaux du thésaurus MeSH. Notre démarche pour ce faire consiste principalement à adapter les mesures existantes dédiées à WordNet, pour une utilisation spécifique dans MeSH. Pour que cette adaptation soit possible, il faut que la structure de WordNet et celle de MeSH soient similaires. Or, ce n'est pas le cas. WordNet est en effet une mono-hiérarchie de concepts tandis que MeSH comprend seize hiérarchies de concepts indépendantes. Cette caractéristique implique une première question centrale dans ce problème d'adaptation : comment peut-on ramener la structure de MeSH à une mono-hiérarchie de concepts de la même manière que WordNet ? La seconde question à laquelle nous nous confrontons consiste à proposer une mesure de similarité plus performante que les mesures existantes.

Dans ce chapitre, nous présentons nos propositions pour solutionner ces problèmes. Dans un premier temps, nous présentons nos contributions pour la mono-hiérarchisation de MeSH. Puis, dans un second temps, nous présentons notre nouvelle mesure de similarité sémantique entre concepts biomédicaux de MeSH.

I. Mono-hiérarchisation du thésaurus MeSH

MeSH est un thésaurus multi-hiérarchique. Il est structuré en seize arborescences de concepts indépendantes, composant les seize domaines de MeSH. Les mesures de similarité basées sur la structure s'appuient en général sur le chemin reliant deux concepts. Or dans le contexte d'arborescences indépendantes, un tel chemin n'existe pas. De ce fait, il devient impossible de mesurer la similarité entre les concepts considérés. L'idée est alors de

transformer la structure de MeSH en une arborescence mono-hiérarchique de sorte à ce que, quelques soient les concepts considérés, il est toujours possible de trouver un chemin les reliant. Dans ce contexte nous proposons deux approches de mono-hiérarchisation de MeSH :

- La première se base sur la mono-hiérarchisation de MeSH avec une racine commune,

- La seconde s'appuie sur le réseau sémantique UMLS pour construire une mono-hiérarchie MeSH.

Dans ce qui suit, nous détaillons chacune de nos deux propositions.

II.1. Mono-hiérarchisation avec une racine commune

Afin de restructurer MeSH en une seule arborescence hiérarchique, nous proposons dans un premier temps de relier les seize racines des différentes arborescences de MeSH à une racine commune avec une relation hiérarchique is-a. Ceci est illustré dans la figure suivante :

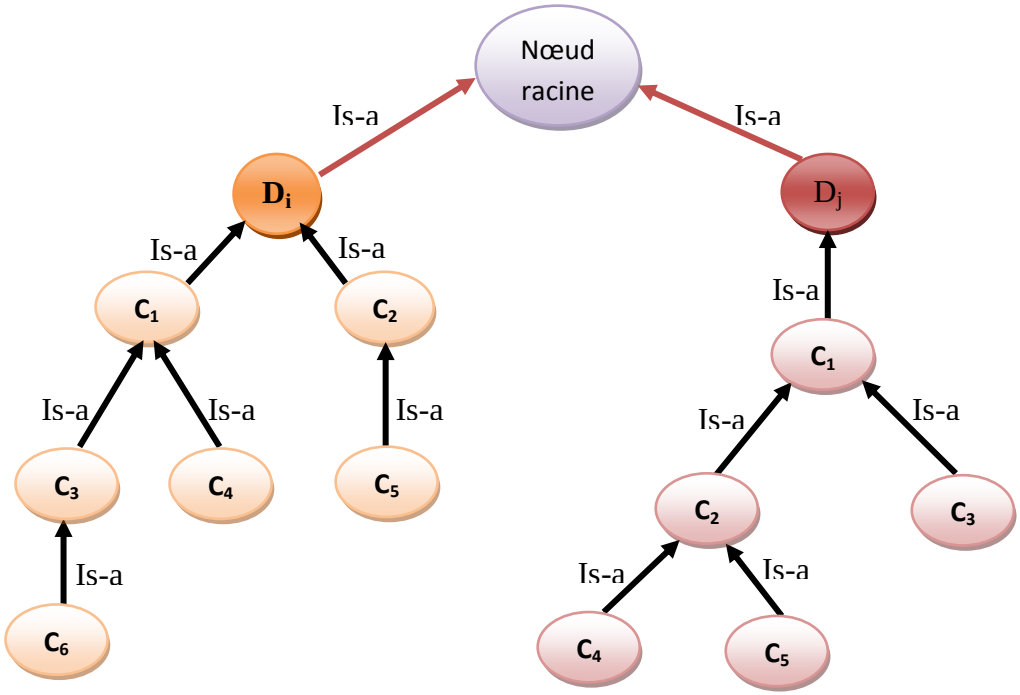

Figure III.1 : Liaison de deux domaines MeSH avec une racine unique

Ainsi, pour relier deux concepts c_i et c_j appartenant respectivement aux domaines D_i et D_j, nous utilisons le chemin (succession d'arcs portant des relations is-a) entre les deux domaines qui passe par la racine commune rajoutée.

II.2. Mono-hiérarchisation basée sur les types sémantiques UMLS

Notre seconde contribution pour restructurer MeSH en une seule arborescence consiste en l'utilisation du système de langage médical unifié UMLS qui compte parmi les terminologies qu'il regroupe le thésaurus MeSH.

Comme nous l'avons vu dans le premier chapitre, UMLS regroupe trois types de bases de connaissances dont le réseau sémantique. En effet, tout en maintenant l'ensemble des relations hiérarchiques ou autres proposées par chacune des sources qu'il regroupe, UMLS organise l'ensemble des concepts au sein d'un réseau sémantique qui lui est propre. Celui-ci consiste en 134 types sémantiques hiérarchisés par des liens is-a.

Dans notre contribution, nous utilisons cette notion avec l'intuition que si les types sémantiques hiérarchisés UMLS regroupent des ensembles de concepts et qu'il en est de même pour les domaines de MeSH, nous pouvons classifier chacun de ces derniers dans le réseau sémantique UMLS en lui associant un type sémantique.

L'idée est donc de mettre en correspondance les domaines de MeSH et les types sémantiques UMLS dans l'objectif de mapper ces domaines dans le réseau sémantique UMLS et ainsi les relier par des relations hiérarchiques. Pour cela, nous partons de l'hypothèse que plus un domaine MeSH partage de concepts avec un type sémantique UMLS, plus, ce type sémantique est apte à le représenter dans le réseau sémantique UMLS.

Notre démarche consiste à associer un score basé sur le nombre de concepts communs partagés entre les domaines MeSH et les types sémantiques UMLS. Le type sémantique TS_i partageant le maximum de concepts avec le domaine D_j est alors sélectionné comme type sémantique adéquat. Ceci peut être formulé comme suit :

$$TS(D_j) = \arg\,\mathrm{Max}(|c_{TS_i}| \cap |C_{D_j}|)$$

Où :

$\left|c_{TS_i}\right| \cap \left|C_{D_j}\right|$ désigne le nombre de concepts communs partagés entre le type sémantique TS_i et le domaine D_j.

La liaison entre les concepts appartenant à différents domaines est alors faite en utilisant les liens is-a intra domaines et les liens is-a inter domaines qui sont des liens inter types sémantiques. Ceci est illustré dans la figure III.2 suivante :

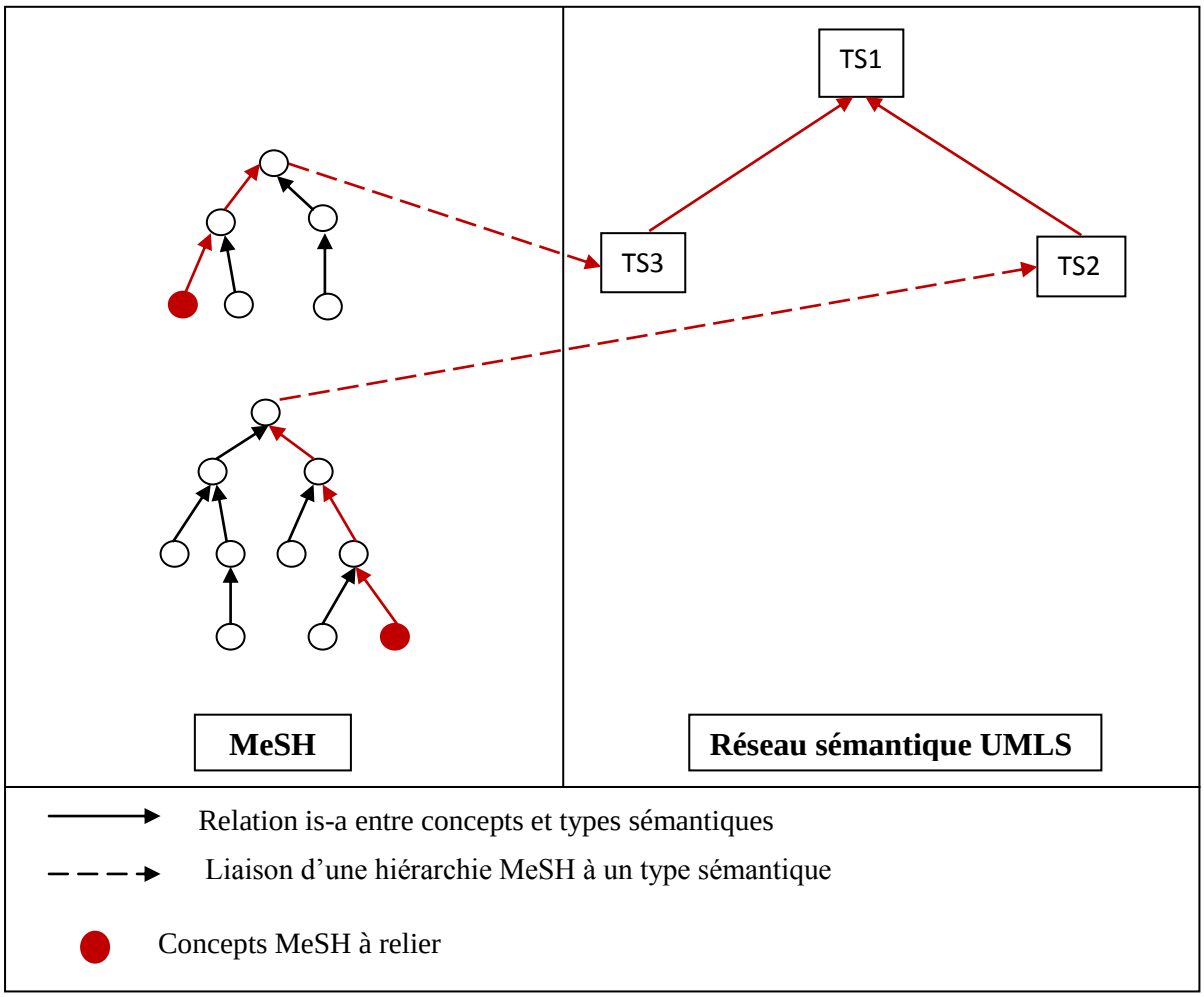

Figure III.2 : Liaison de deux concepts à travers les racines de leurs hiérarchies avec UMLS

II. Une nouvelle mesure de similarité sémantique

Le thésaurus MeSH est structuré d'une manière multi-hiérarchique et chacun de ses concepts est accompagné d'informations permettant de mieux comprendre son sens. Dans ce contexte nous proposons une nouvelle mesure de similarité sémantique entre concepts MeSH basée à la fois sur cette structure avec une mesure structurelle et sur ces informations avec une mesure informationnelle. En effet, nous estimons que la similarité entre deux concepts MeSH est fortement liée à la distance qui les sépare et aux informations qu'ils partagent. Formellement, notre mesure de similarité est représentée par une somme pondérée de la similarité structurelle (sim_{str}) et la similarité informationnelle (sim_{inf}) comme suit :

$$Sim(c_i, c_j) = \propto \left(Sim_{str}(c_i, c_j)\right) + (1-\propto)\left(Sim_{inf}(c_i, c_j)\right)$$

Dans ce qui suit nous présentons l'intuition derrière chacune de ces mesures.

III.1. Similarité structurelle

L'intuition derrière l'attribution de la similarité structurelle est que plus deux concepts sont proches dans leurs positions hiérarchiques, plus ils sont similaires et inversement, plus deux concepts sont éloignés par leurs positions hiérarchiques moins ils sont similaires. La similarité serait alors inversement proportionnelle à la distance qui sépare les concepts. La distance séparant les concepts est calculée par le nombre d'arcs véhiculant la relation is-a séparant les nœuds qui les représentent dans la hiérarchie. Formellement ceci est défini comme suit :

$$Sim_{str}(c_i, c_j) = \left(\frac{1}{dist(c_i, c_j)}\right)$$

Dans le cas de concepts appartenant à des domaines différents, la distance est calculée en utilisant l'une des deux contributions pour la mono-hiérarchisation de MeSH présentées dans la section précédente.

III.2. Similarité informationnelle

Les concepts dans MeSH sont accompagnés d'informations qui mettent en valeur l'idée qu'ils expriment. Parmi ces informations nous avons les définitions des concepts et leurs synonymes.

Notre intuition est que plus deux concepts partagent d'informations communes entre eux, plus ils sont similaires. Dans le même principe, plus le voisinage de deux concepts partage d'informations en commun, plus ces deux concepts sont similaires.

De ce fait, nous utilisons les synonymes des concepts et les mots de leurs définitions, ainsi que les synonymes de leurs voisinages et les mots de leurs définitions pour estimer cette similarité. Nous avons défini le voisinage des deux concepts par leurs ascendants et descendants directs.

Cette similarité est formellement définie comme suit :

$$Sim_{inf}(c_i, c_j) = \left(\frac{|S_i \cap S_j|}{|S_i \cup S_j|}\right) + \left(\frac{|D_i \cap D_j|}{|D_i \cup D_j|}\right) + \left(\frac{|S_{vois_i} \cap S_{vois_j}|}{|S_{vois_i} \cup S_{vois_j}|}\right) + \left(\frac{|D_{vois_i} \cap D_{vois_j}|}{|D_{vois_i} \cup D_{vois_j}|}\right)$$

Où :

- S_i, S_j, D_i, D_j représentent respectivement les ensembles des synonymes et mots constituant les définitions pour les concepts i et j.
- S_{vois_i}, S_{vois_j}, D_{vois_i} , D_{vois_j} respectivement les synonymes et les mots des définitions des voisins de c_i et c_j par la relation is-a.

III. Exemple illustratif

Dans cette section nous présentons un exemple illustratif d'application de notre mesure de similarité sémantique pour comparer deux concepts, nous utiliserons les deux propositions pour l'estimation de la similarité structurelle présentées précédemment.

Soient à comparer les deux concepts « *Vaccines* » et « *Immunity* » appartenant à deux domaines différents. Les différentes parties de notre mesure de similarité sont calculées comme suit :

IV.1. Similarité structurelle

Les localisations des deux concepts « *Vaccines* » et « *Immunity* » dans les arborescences D et G de MeSH respectivement sont « D20.215.894 » et « G12.450 ». Les deux concepts n'appartiennent pas au même domaine, donc dans un premier cas, nous calculons la distance qui les sépare en utilisant une racine unique reliant leurs deux domaines, ensuite nous utilisons les types sémantiques associés à leurs catégories respectives dans UMLS.

- **CAS 1 :** Calcul de la distance avec ajout d'une racine commune

La figure suivante montre les segments d'hiérarchies MeSH où se trouvent les deux concepts à comparer.

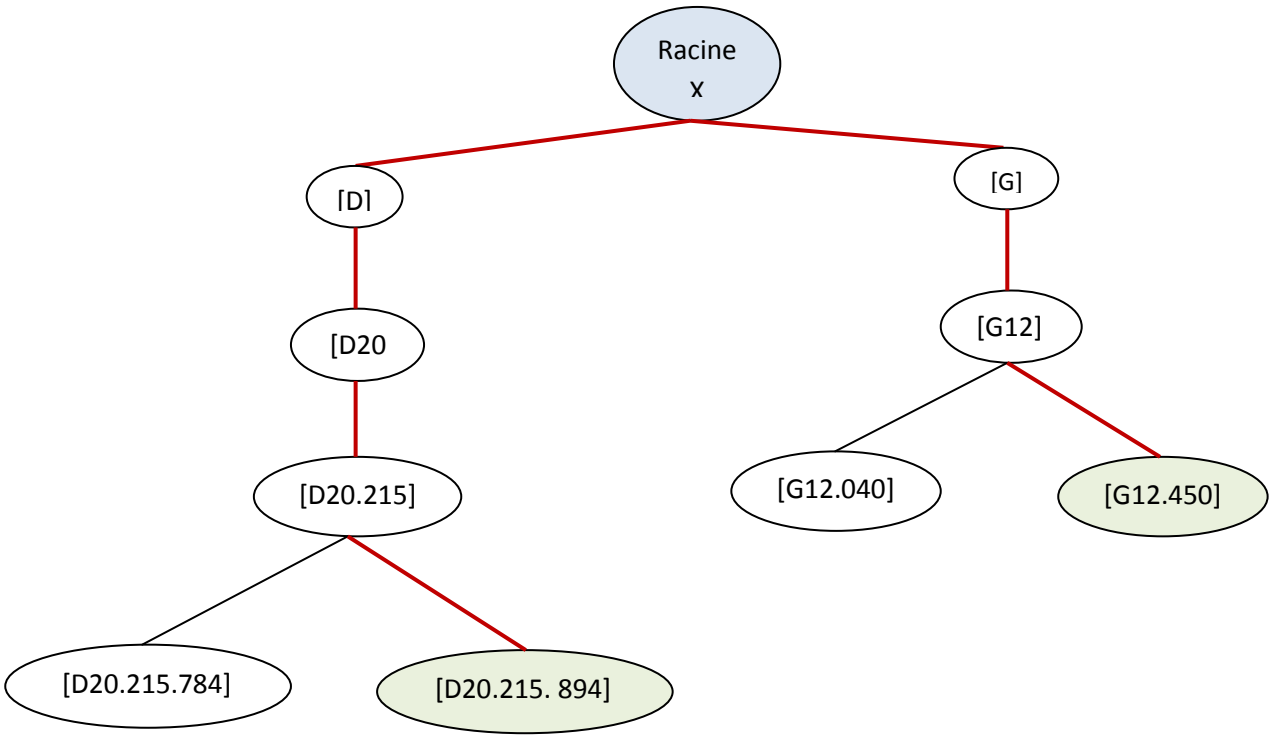

Figure III.3 : Segments des hiérarchies des concepts « *Vaccines* » et « *Immunity* » reliés par une racine

➔ La similarité structurelle est donc calculée par l'inverse de la distance séparant les deux concepts comme suit :

$$Sim_{str}(\textit{Vaccines}, \textit{Immunity}) = 1/7 = 0.143$$

- **CAS 2 :** Avec utilisation d'UMLS

Les deux types sémantiques associés aux deux domaines de MeSH « D » et « G » sont « *Pharmacologic Substance* » et « Natural Phenomenon or Process ». Ces deux types sémantiques sont situés dans l'arborescence des types sémantiques UMLS avec les localisations « A1.4.1.1.1 » et « B2.2 » respectivement. Ceci peut être schématisé comme suit :

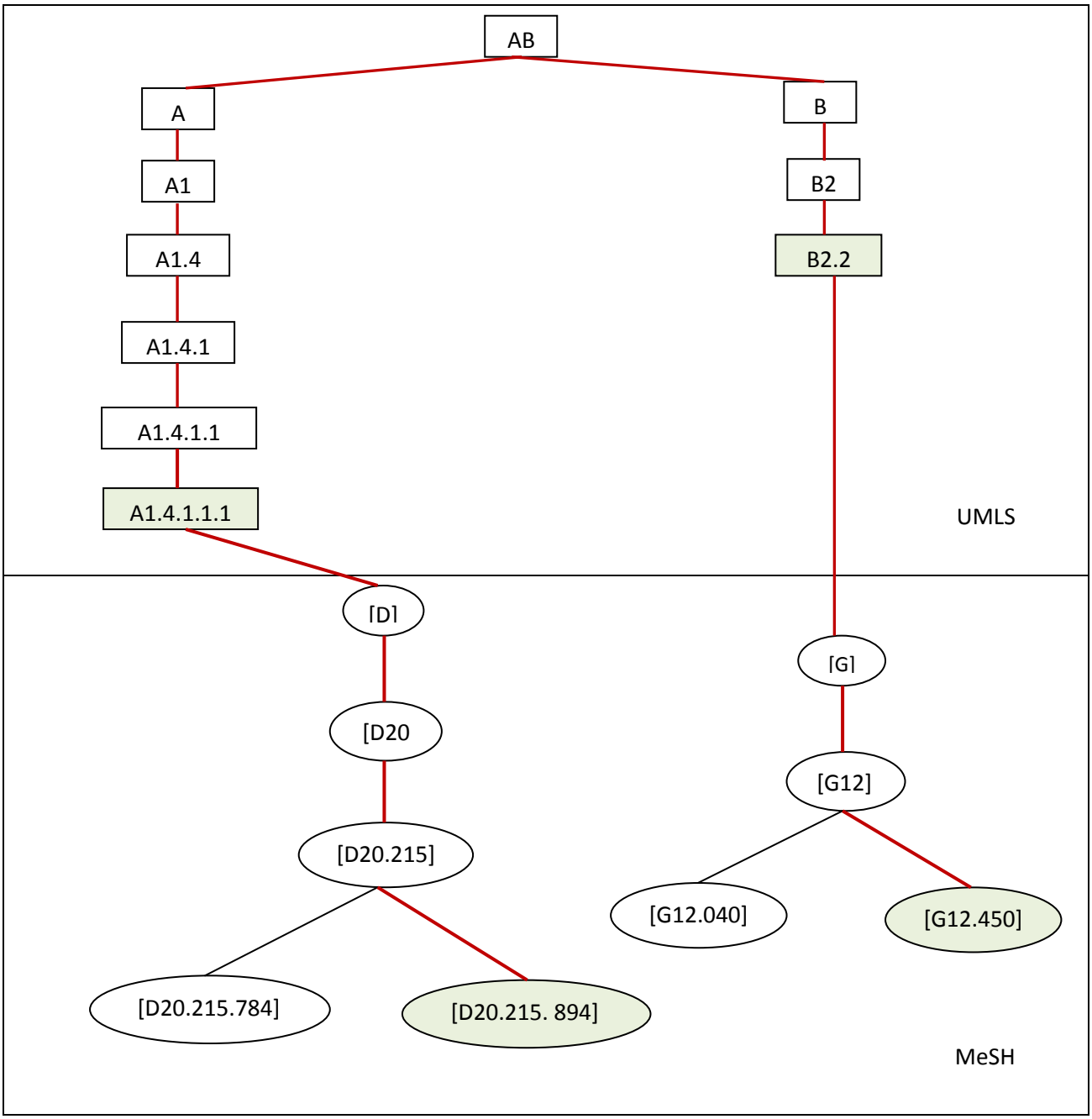

Figure III.4 : Liaison des deux concepts « *Vaccines* » et « *Immunity* » par leurs types sémantiques

➔ La similarité structurelle est donc calculée par l'inverse de la distance séparant les deux concepts comme suit :

$$Sim_{str}(\textit{Vaccines}, \textit{Immunity}) = 1/16 = 0.0625$$

IV.2. Similarité informationnelle

Dans ce qui suit nous calculons la partie informationnelle de notre mesure qui concerne les synonymes, définitions, synonymes du voisinage et définitions du voisinage.

- **Synonymes :** Il n'existe pas de synonymes pour les deux concepts « Vaccines » et « Immunity », la valeur est donc nulle.

- **Définitions :** Les définitions des deux concepts sont les suivantes :

« Vaccines »: Suspensions of killed or attenuated microorganisms (bacteria, viruses, fungi, protozoa, or rickettsiae), antigenic proteins derived from them, or synthetic constructs, administered for the prevention, amelioration, or treatment of infectious and other diseases.
« Immunity »: Nonsusceptibility to the invasive or pathogenic effects of foreign microorganisms or to the toxic effect of antigenic substances.

➔ La similarité selon les définitions est donc:

$$Sim_{def}(Vaccines, Immunity) = 2/28 = 0,0714$$

- **Synonymes du voisinage :**

Concept : « Vaccines »	
Voisin	**Synonymes**
Biological Agents : [D20.215]	« Biologic Products »,« Products, Biological » , « Agents, Biological »,« Products, Biologic » , « Products, Natural »
Alzheimer Vaccines : [D20.215.894.067]	« Vaccines, Alzheimer »
Bacterial Vaccines : [D20.215.894.135]	« Bacterial Vaccine » , « Bacterin » , « Vaccine, Bacterial » , « Vaccines, Bacterial »
Cancer Vaccines : [D20.215.894.200]	/

Tableau III.1 : Synonymes du voisinage du concept « Vaccines »

Concept : « Immunity »	
Voisin	**Synonymes**
Immune System Phenomena: [G12]	« Immune System Concepts» , « Immune System Phenomenon» , « Concept, Immune System» , « Concepts, Immune System» , « Immune System Concept» , « Phenomena, Immune System » , « Phenomenon, Immune System» , « Immune System Process»,

	«Process, Immune System», «Processes, Immune System »
Adaptive Immunity: [G12.450.050]	« Adaptive Immune Response » , « Adoptive Immunity» , « Immunity, Adaptive» , «Immune Response, Adaptive» , «Immunity, Acquired» , « Immunity, Adoptive » , «Response, Adaptive Immune»
Autoimmunity: [G12.450.192]	« Autoimmune Responses» , « Autoimmunities»
Cross Protection: [G12.450.275]	« Protection, Cross »

Tableau III.2 : Synonymes du voisinage du concept « Immunity »

➔ La similarité selon les synonymes du voisinage est donc:

$$Sim_{vois\_Syn}(Vaccines, Immunity) = 0$$

- **Définitios du voisinage :**

Concept : « Vaccines »	
Voisin	**Définitions**
Biological Agents : [D20.215]	Organisms or complex pharmaceutical substances, preparations, or agents of organic origin, usually obtained by biological methods or assay. Biological agents are differentiated from BIOLOGICAL FACTORS in that the latter are compounds with biological or physiological activity made by living organisms. (From Webster's 3d ed).
Alzheimer Vaccines : [D20.215.894.067]	Vaccines or candidate vaccines used to prevent or treat ALZHEIMER DISEASE.
Bacterial Vaccines : [D20.215.894.135]	Suspensions of attenuated or killed bacteria administered for the prevention or treatment of infectious bacterial disease.
Cancer Vaccines : [D20.215.894.200]	Vaccines or candidate vaccines designed to prevent or treat cancer. Vaccines are produced using the patient's own whole tumor cells as the source of antigens, or using tumor-specific antigens, often

	recombinantly produced.
Fungal Vaccines : [D20.215.894.354]	Suspensions of attenuated or killed fungi administered for the prevention or treatment of infectious fungal disease.
Protozoan Vaccines : [D20.215.894.582]	Suspensions of attenuated or killed protozoa administered for the prevention or treatment of infectious protozoan disease.
Toxoids : [D20.215.894.691]	Preparations of pathogenic organisms or their derivatives made nontoxic and intended for active immunologic prophylaxis. They include deactivated toxins. Anatoxin toxoids are distinct from anatoxins that are TROPANES found in CYANOBACTERIA.
Vaccines, Attenuated : [D20.215.894.811]	Live vaccines prepared from microorganisms which have undergone physical adaptation (e.g., by radiation or temperature conditioning) or serial passage in laboratory animal hosts or infected tissue/cell cultures, in order to produce avirulent mutant strains capable of inducing protective immunity.
Vaccines, Combined : [D20.215.894.815]	Two or more vaccines in a single dosage form.
Vaccines, Contraceptive :[D20.215.894.818]	Vaccines or candidate vaccines used to prevent conception.
Vaccines, Inactivated :[D20.215.894.830]	Vaccines in which the infectious microbial nucleic acid components have been destroyed by chemical or physical treatment (e.g., formalin, beta-propiolactone, gamma radiation) without affecting the antigenicity or immunogenicity of the viral coat or bacterial outer membrane proteins.
Vaccines, Synthetic : [D20.215.894.865]	Small synthetic peptides that mimic surface antigens of pathogens and are immunogenic, or vaccines manufactured with the aid of recombinant DNA techniques. The latter vaccines may also be whole viruses whose nucleic acids have been modified.

Viral Vaccines : [D20.215.894.899]	Suspensions of attenuated or killed viruses administered for the prevention or treatment of infectious viral disease.

Tableau III.3 : Définitions du voisinage du concept « *Vaccines* »

Concept : « Immunity »	
Voisin	**Définitions**
Immune System Phenomena: [G12]	The characteristic properties and processes involved in IMMUNITY and an organism's immune response.
Adaptive Immunity: [G12.450.050]	Protection from an infectious disease agent that is mediated by B- and T- LYMPHOCYTES following exposure to specific antigen, and characterized by IMMUNOLOGIC MEMORY. It can result from either previous infection with that agent or vaccination (IMMUNITY, ACTIVE), or transfer of antibody or lymphocytes from an immune donor (IMMUNIZATION, PASSIVE).
Autoimmunity: [G12.450.192]	Process whereby the immune system reacts against the body's own tissues. Autoimmunity may produce or be caused by AUTOIMMUNE DISEASES.
Immunity, Innate : [G12.450.564]	The capacity of a normal organism to remain unaffected by microorganisms and their toxins. It results from the presence of naturally occurring ANTI-INFECTIVE AGENTS, constitutional factors such as BODY TEMPERATURE and immediate acting immune cells such as NATURAL KILLER CELLS.
Immunity, Maternally-Acquired : [G12.450.570]	Resistance to a disease-causing agent induced by the introduction of maternal immunity into the fetus by transplacental transfer or into the neonate through colostrum and milk.
Immunity, Mucosal : [G12.450.573]	Nonsusceptibility to the pathogenic effects of foreign microorganisms or antigenic substances as a result of antibody secretions of the mucous

	membranes. Mucosal epithelia in the gastrointestinal, respiratory, and reproductive tracts produce a form of IgA (IMMUNOGLOBULIN A, SECRETORY) that serves to protect these ports of entry into the body.
Plant Immunity : [G12.450.800]	The inherent or induced capacity of plants to withstand or ward off biological attack by pathogens.
T-Cell Antigen Receptor Specificity: [G12.450.900]	The property of the T-CELL RECEPTOR which enables it to react with some antigens and not others. The specificity is derived from the structure of the receptor's variable region which has the ability to recognize certain antigens in conjunction with the MAJOR HISTOCOMPATIBILITY COMPLEX molecule.

Tableau III.4 : Définitions du voisinage du concept « *Immunity* »

➔ La similarité selon les définitions du voisinage est donc:

$$Sim_{vois\text{-}Def}(Vaccines, Immunity) = 7/150 = 0.0467$$

IV.3 Similarité finale

Après avoir calculé les deux types de similarités structurelle et informationnelle nous pouvons calcule la similarité finale pour les deux cas, en utilisant une racine commune et pour le cas où nous utilisons les types sémantiques UMLS comme suit :

Cas1 : Sim (*Vaccines, Immunity*) = α (0.143) + (1-α) (0.0714+0.0467)

Cas2 : Sim (*Vaccines, Immunity*) = α (0.0625) + (1-α) (0.0714+0.0467)

IV. Conclusion

Nous avons présenté Nos contributions pour mesurer la similarité sémantique entre concepts biomédicaux issus du thésaurus MeSH. Nous avons commencé par une restructuration de MeSH en une seule arborescence à travers deux approches, la première avec adjonction des domaines de MeSH à une racine commune unique, et la seconde avec mappage des domaines de MeSH dans le réseau sémantique UMLS. Nous avons ensuite présenté notre mesure de similarité sémantique ainsi qu'un exemple illustratif de son utilisation.

CHAPITRE IV

Implémentation et tests

I. Introduction

Nous avons présenté dans le chapitre précédent nos contributions pour une mesure de similarité sémantique en passant par une mono-hiérarchisation du thésaurus MeSH, qui est, à la base composé de seize hiérarchies indépendantes. Dans ce chapitre, nous présentons les résultats de notre seconde contribution pour la restructuration de MeSH en une seule arborescence en utilisant le réseau sémantique UMLS, nous présentons ensuite les étapes de notre implémentation, l'environnement technologique, le corpus utilisé pour l'expérimentation, et enfin, nous présentons les résultats de l'évaluation de notre mesure.

I. Résultats de la restructuration de MeSH avec le réseau sémantique UMLS

Afin de restructurer MeSH en une seule arborescence hiérarchique nous avons proposé deux approches, la seconde qui consiste à mapper les domaines de MeSH dans le réseau sémantique UMLS nous a permis d'attribuer à chacun des domaines de MeSH un type sémantique UMLS. Les résultats de cette contribution sont présentés dans le tableau suivant :

Domaine MeSH	Type Sémantique UMLS
Anatomy	Body Part, Organ, or Organ Component
Organisms	Plant
Diseases	Disease or Syndrome
Chemicals and Drugs	Pharmacologic Substance
Analytical, Diagnostic and Therapeutic Techniques and Equipment	Therapeutic or Preventive Procedure
Psychiatry and Psychology	Mental or Behavioral Dysfunction
Phenomena and Processes	Natural Phenomenon or Process
Disciplines and Occupations	Biomedical Occupation or Discipline
Anthropology, Education, Sociology and Social Phenomena	Idea or Concept
Technology, Industry, Agriculture	Organic Chemical
Humanities	Idea or Concept

Information Science	Intellectual Product
Named Groups	Professional or Occupational Group
Health Care	Health Care Activity
Publication Characteristics	Intellectual Product
Geographicals	Geographic Area

Tableau IV.1 : Domaines de MeSH et types sémantiques associés dans UMLS

II. Evaluation

L'objectif de ces expérimentations est de mesurer les performances de notre mesure de similarité sémantique entre concepts biomédicaux comparativement à une mesure plus classique telle que la mesure de Leacock et Chodorow [Leacock et al., 98].

Plusieurs variantes de notre mesure de similarité ont été testées. Dans ce qui suit on posera $Sim_{comb(i)}$ la variante numéro (i) de notre similarité combinée :

$$Sim_{comb} = \propto Sim_{str} + (1-\propto) Sim_{inf}$$

Les variantes suivantes ont été évaluées :

- $Sim_{comb(1)}(c_1, c_2) = \propto \frac{1}{dist(c_1,c_2)} +$

$$(1-\propto)\left(\left(\frac{|S_i \cap S_j|}{|S_i \cup S_j|}\right) + \left(\frac{|D_i \cap D_j|}{|D_i \cup D_j|}\right) + \left(\frac{|S_{vois_i} \cap S_{vois_j}|}{|S_{vois_i} \cup S_{vois_j}|}\right) + \left(\frac{|D_{vois_i} \cap D_{vois_j}|}{|D_{vois_i} \cup D_{vois_j}|}\right)\right)$$

- $Sim_{comb(2)}(c_1, c_2) = \propto \left(-\log\left(\frac{dist(c_1,c_2)}{2*prof_{max}}\right)\right) +$

$$(1-\propto)\left(\left(\frac{|S_i \cap S_j|}{|S_i \cup S_j|}\right) + \left(\frac{|D_i \cap D_j|}{|D_i \cup D_j|}\right) + \left(\frac{|S_{vois_i} \cap S_{vois_j}|}{|S_{vois_i} \cup S_{vois_j}|}\right) + \left(\frac{|D_{vois_i} \cap D_{vois_j}|}{|D_{vois_i} \cup D_{vois_j}|}\right)\right)$$

Remarque : Chacune de ces deux variantes a été évaluée dans le contexte des deux mono-hiérarchies MeSH proposées en (section II.1 du chapitre précédent).

Pour évaluer notre mesure de similarité entre concepts, nous l'intégrons dans un processus d'indexation sémantique puis nous évaluons les résultats de recherche correspondants. En particulier le processus d'indexation sémantique est basé sur une étape de

désambiguïsation de termes d'index qui s'appuie sur la valeur d'un score associé au concept. Ce score, dit score de désambigüisation est une somme cumulée de valeurs de similarités entre concepts.

Dans nos tests :

- Les termes d'index sont des termes biomédicaux issus de Cxtractor.
- Nous avons utilisé le score de désambiguïsation de D.Dinh [Dinh, 12] formellement défini comme suit :

$$\begin{cases} s_k = \sum_{s_1 \in syn(c_1), s_2 \in syn(c_2)} sim(s_1, s_2)\,; si\ k \leq 2 \\ s_k = \sum_{s \in syn(c_k)} sim(s_{k-1}, s)\,, si\ k > 2 \end{cases}$$

Dans ce score que nous avons implémenté, sim(s_1,s_2) est successivement remplacée par $Sim_{comb(1)}$, $Sim_{comb(2)}$, et Sim_{LC}.

Les index sémantiques correspondants $Index\_Sim_{comb(1)}$, $Index\_Sim_{comb(2)}$, et $Index\_Sim_{LC}$ sont alors évalués et comparés à travers les résultats de recherche qu'ils fournissent.

La recherche est effectuée sur la plateforme Terrier en utilisant une sous-collection de la collection TREC Genomics.

Les comparaisons suivantes ont été réalisées pour les deux mono-hiérarchies MeSH :

- $Index\_Sim_{Comb(1)}$ VS $Index\_Sim_{LC}$
- $Index\_ Sim_{comb(2)}$ VS $Index\_Sim_{LC}$

Le schéma global des étapes de notre évaluation est illustré dans la figure IV.1 suivante :

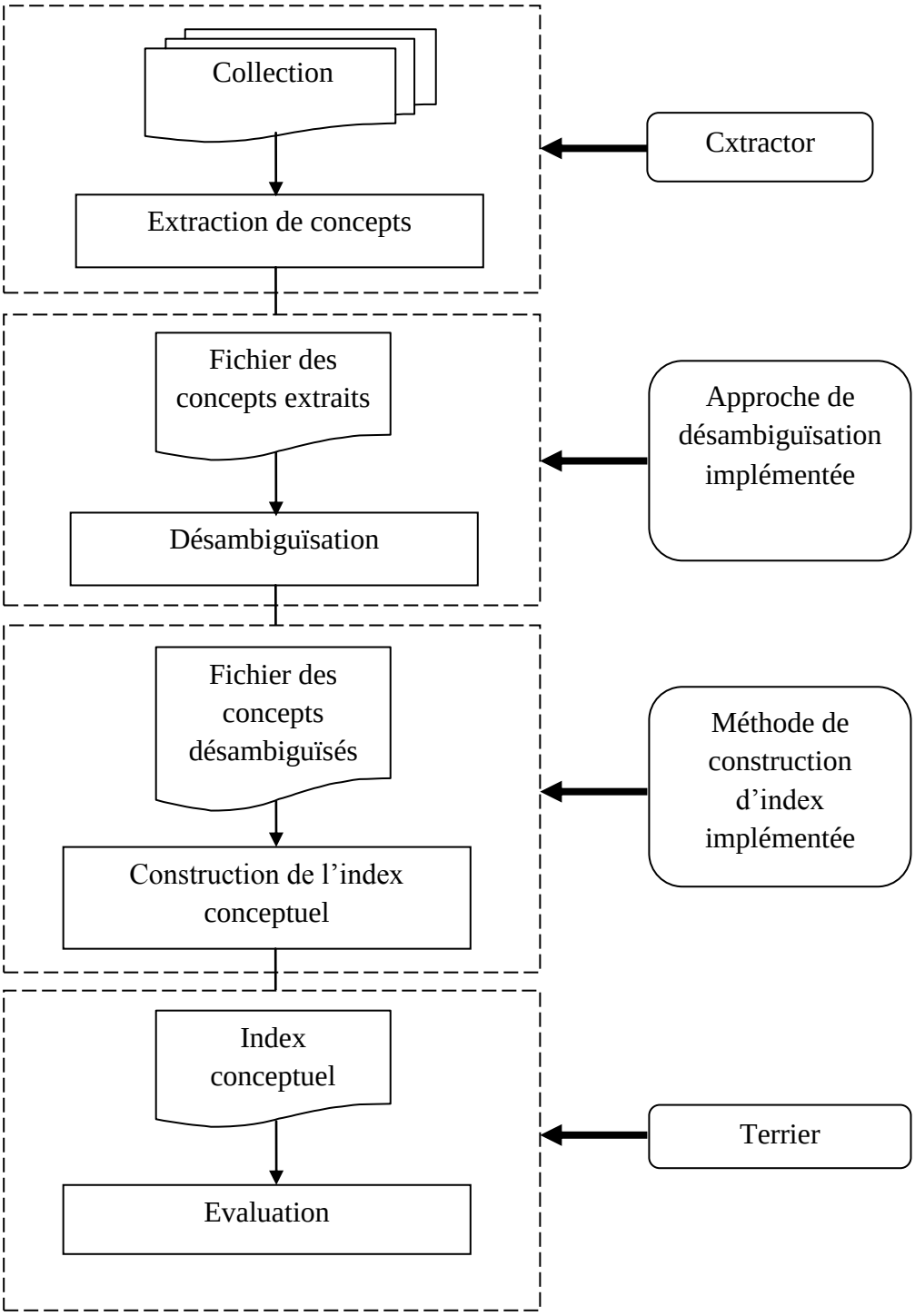

Figure IV.1 : Schéma global d'évaluation de notre mesure

III. Cadre d'évaluation

Pour nos expérimentations, nous avons utilisé une collection de documents, un ensemble de requêtes associées sous un protocole d'évaluation. Dans ce qui suit, nous présentons cette collection de document, les requêtes associées à la collection ainsi que le protocole d'évaluation.

III.1. Description de la collection de documents

Pour évaluer notre approche nous utilisons la collection TREC Genomics 2004 car elle représente les vrais besoins d'informations des professionnels de santé.

Le tableau suivant présente les statistiques de la collection TREC Genomics 2004. Il s'agit d'un sous-ensemble de la base bibliographique, intitulée MEDLINE4, des résumés d'articles de journaux biomédicaux entre 1994 et 2003. Le tableau suivant présente des statistiques sur cette collection de documents.

Nombre de documents	4.6 millions
Nombre de documents jugés	42255
Longueur moyenne du document	202
Nombre de requêtes	50
Longueur moyenne de la requête	17
Nombre de documents pertinents par requête	75

Tableau IV.2. Statistiques de la collection TREC Genomics 2004

Le nombre de documents était environ de 4.6 millions d'enregistrements représentant approximativement un tiers de la taille de MEDLINE jusqu'en 2004. La taille de la collection occupe 9.5 Giga octets au total. Il existe parmi ces enregistrements 1209243 (soit 26.3 %) qui n'ont pas de résumés.

Exemple de document :

```
<DOC>

<DOCNO>11005875</DOCNO>

<TITLE>Regional and strain-specific gene expression mapping in the
adult mouse brain. </TITLE>

<ABSTRACT>To determine the genetic causes and molecular mechanisms
responsible for neurobehavioral differences in mice, we used highly
parallel gene expression profiling to detect genes that are
differentially expressed between the 129SvEv and C57BL/6 mouse
strains at baseline and in response to seizure. In addition, we
identified genes that are differentially expressed in specific brain
regions. We found that approximately 1% of expressed genes are
differentially expressed between strains in at least one region of
the brain and that the gene expression response to seizure is
significantly different between the two inbred strains. The results
lead to the identification of differences in gene expression that
may account for distinct phenotypes in inbred strains and the unique
functions of specific brain regions.

</ABSTRACT>

</DOC>
```

Pour notre évaluation nous avons pris une sous collection de 1000 documents.

III.2. Description de l'ensemble des requêtes

Les requêtes sont collectées à partir des vrais besoins d'informations des biologistes. Un ensemble de 50 requêtes ont été créées et jugés par les utilisateurs. Chaque requête contient trois champs principaux :

– **ID** : identifiant de la requête,

– **TITLE** : besoin d'information bref ou requête courte,

– **NEED** : besoin d'information détaillé ou requête longue,

– **CONTEXT** : information supplémentaire sur la requête.

Exemple de requête :

<TOPIC>
<ID>2</ID>
<TITLE>Generating transgenic mice</TITLE>
<NEED>Find protocols for generating transgenic mice.</NEED>
<CONTEXT>Determine protocols to generate transgenic mice having a single copy of the gene of interest at a specific location.</CONTEXT>
</TOPIC>

Des jugements de pertinence sont associés aux requêtes selon le format suivant :

1	0	10077651	2
1	0	10084280	2
1	0	10084283	1
1	0	10088633	2
1	0	10226041	2
1	0	10228348	2
1	0	10318901	2

III.3. Protocole d'évaluation

L'évaluation des résultats de la recherche est effectuée selon le protocole TREC. Plus précisément, chaque requête est soumise au système de RI avec les paramètres fixés. Le système renvoie les 1000 premiers documents pour chaque requête. Les valeurs de précision P@1, P@2, P@3, P@4, P@5, P@10, P@15, P@20, P@30, P@100 et P@1000, ainsi que la R-Precision (précision réelle ou exacte) et la MAP (précision moyenne) sont calculées.

- **précision** à *X* premiers documents (dénotée *P@X*) représente la proportion des documents pertinents par rapport aux *X* premiers documents renvoyés par le SRI. Elle mesure la satisfaction de l'utilisateur concernant les *X* premiers documents pertinents.
- **R-Precision** (précision réelle ou exacte) correspond à la précision exacte calculée sur l'ensemble des documents pertinents retournés.
- **précision moyenne** (Mean Average Precision, dénotée *MAP*) correspond à la précision moyenne calculée sur l'ensemble des documents pertinents retournés. Elle mesure la capacité du modèle d'appariement ou d'un SRI à pouvoir sélectionner les documents pertinents, en réponse à un ensemble de requêtes.

V. Résultats Expérimentaux

Dans ce qui suit nous présentons les résultats de nos expérimentations pour les variantes de notre mesure de similarité entre concepts biomédicaux comparativement à la mesure de Leacock et Chodorow. Nous présentons dans un premier temps les résultats pour notre mesure combinée $Sim_{comb(1)}$, puis dans un second temps les résultats de la mesure combinée $Sim_{comb(2)}$.

V.1. Mesure combinée $Sim_{comb(1)}$

Dans ce qui suit nous présentons les résultats obtenus pour l'ensemble des requêtes avec les différentes valeurs de α, en utilisant une mono-hiérarchisation de MeSH avec l'adjonction à une racine commune pour la précision moyenne (tableau IV.3) et la précision à différents points (tableau IV.4), ensuite, en utilisant une mono-hiérarchisation de MeSH avec un mappage des seize domaines dans le réseau sémantique UMLS pour la précision moyenne (tableau IV.5) et la précision à différents points (tableau IV.6).

Mesure	Average Précision
Sim_{LC}	0.2009
α =0	0,1602
α =0.1	0,1647
α =0.2	0,1689
α =0.3	0,1689
α =0.4	0,1813
α =0.5	0,1709
α =0.6	0,1802
α =0.7	0,1784
α =0.8	0,1778
α =0.9	0,1792
α =1	0,1792

Tableau IV.3 : Résultats de la précision moyenne pour la mesure $Sim_{comb(1)}$ avec une restructuration de MeSH par adjonction des seize domaines à une racine commune

Mesure	Average Précision
Sim_{LC}	0.2009
α =0	0,1602
α =0.1	0,1599
α =0.2	0,1594
α =0.3	0,1634
α =0.4	0,1652
α =0.5	0,1656
α =0.6	0,1652
α =0.7	0,1653
α =0.8	0,1757
α =0.9	0,1774
α =1	0.1774

Tableau IV.4 : Résultats de la précision moyenne pour la mesure $Sim_{comb(1)}$ avec une restructuration de MeSH par mappage sur le réseau sémantique UMLS.

	Sim_{LC}	α=0	α=0.1	α=0.2	α=0.3	α=0.4	α=0.5	α=0.6	α=0.7	α=0.8	α=0.9	α=1
P@1	0,2105	0,1795	0,1795	0,1842	0,1842	0,2105	0,1842	0,2105	0,2105	0,2105	0,2105	0,2105
P@2	0,2237	0,1923	0,1923	0,1974	0,1974	0,2105	0,1842	0,2105	0,2105	0,2105	0,2105	0,2105
P@3	0,193	0,1709	0,1709	0,1754	0,1754	0,193	0,1667	0,2018	0,2018	0,193	0,193	0,193
P@4	0,1711	0,1282	0,1282	0,1316	0,1316	0,1513	0,1316	0,1513	0,1513	0,1513	0,1513	0,1513
P@5	0,1632	0,1231	0,1231	0,1263	0,1263	0,1474	0,1263	0,1421	0,1421	0,1421	0,1474	0,1474
P@10	0,1368	0,1205	0,1256	0,1263	0,1263	0,1316	0,1289	0,1289	0,1289	0,1263	0,1289	0,1289
P@15	0,1105	0,1145	0,1145	0,114	0,114	0,1158	0,1193	0,114	0,1123	0,1123	0,1158	0,1158
P@20	0,0987	0,0936	0,0949	0,0974	0,0974	0,0974	0,0974	0,0974	0,0961	0,0961	0,0961	0,0961
P@30	0,0825	0,0718	0,0726	0,0754	0,0754	0,0763	0,0772	0,0772	0,0763	0,0763	0,0763	0,0763
P@50	0,0589	0,0518	0,0523	0,0537	0,0537	0,0542	0,0547	0,0537	0,0532	0,0532	0,0532	0,0532
P@100	0,0339	0,0305	0,0308	0,0316	0,0316	0,0318	0,0313	0,0318	0,0316	0,0316	0,0316	0,0316
P@200	0,0197	0,0179	0,0181	0,0186	0,0186	0,0187	0,0189	0,0187	0,0186	0,0186	0,0186	0,0186
P@500	0,0095	0,0089	0,009	0,0092	0,0092	0,0093	0,0093	0,0093	0,0092	0,0092	0,0092	0,0092
P@1000	0,0048	0,0045	0,0045	0,0046	0,0046	0,0046	0,0046	0,0046	0,0046	0,0046	0,0046	0,0046

Tableau IV.5 : Résultats de la Précision @ X pour la mesure $Sim_{comb(1)}$ pour une restructuration de MeSH par adjonction des seize domaines à une racine commune unique.

	Sim_{LC}	α=0	α=0.1	α=0.2	α=0.3	α=0.4	α=0.5	α=0.6	α=0.7	α=0.8	α=0.9	α=1
P@1	0,2105	0,1795	0,1795	0,1795	0,1842	0,1842	0,1842	0,1842	0,1842	0,2105	0,2105	0,2105
P@2	0,2237	0,1923	0,1795	0,1795	0,1842	0,1711	0,1842	0,1711	0,1842	0,2105	0,2105	0,2105
P@3	0,193	0,1709	0,1624	0,1624	0,1667	0,1667	0,1667	0,1667	0,1754	0,193	0,193	0,193
P@4	0,1711	0,1282	0,1282	0,1282	0,1316	0,1316	0,1316	0,1316	0,1382	0,1447	0,1513	0,1513
P@5	0,1632	0,1231	0,1231	0,1231	0,1263	0,1263	0,1263	0,1263	0,1316	0,1474	0,1526	0,1526
P@10	0,1368	0,1205	0,1205	0,1179	0,1211	0,1211	0,1211	0,1211	0,1184	0,1211	0,1263	0,1263
P@15	0,1105	0,1145	0,1111	0,1077	0,1105	0,1123	0,1123	0,1123	0,1105	0,1105	0,1123	0,1123
P@20	0,0987	0,0936	0,0923	0,0897	0,0908	0,0921	0,0921	0,0921	0,0934	0,0934	0,0934	0,0934
P@30	0,0825	0,0718	0,0718	0,0701	0,0711	0,0728	0,0719	0,0728	0,0737	0,0737	0,0737	0,0737
P@50	0,0589	0,0518	0,0513	0,0503	0,0516	0,0532	0,0532	0,0532	0,0521	0,0521	0,0516	0,0516
P@100	0,0339	0,0305	0,0297	0,03	0,0308	0,0313	0,0311	0,0313	0,0311	0,0308	0,0305	0,0305
P@200	0,0197	0,0179	0,0179	0,0179	0,0184	0,0188	0,0188	0,0188	0,0187	0,0187	0,0186	0,0186
P@500	0,0095	0,0089	0,0089	0,0089	0,0091	0,0092	0,0092	0,0092	0,0092	0,0092	0,0092	0,0092
P@1000	0,0048	0,0045	0,0044	0,0044	0,0046	0,0046	0,0046	0,0046	0,0046	0,0046	0,0046	0,0046

Tableau IV.6 : Résultats de la précision à différents points X (Précision @ X) pour la mesure $Sim_{comb(1)}$ pour une restructuration de MeSH par mappage sur le réseau sémantique UMLS.

Les courbes représentant les différentes valeurs de α sont données comme suit :

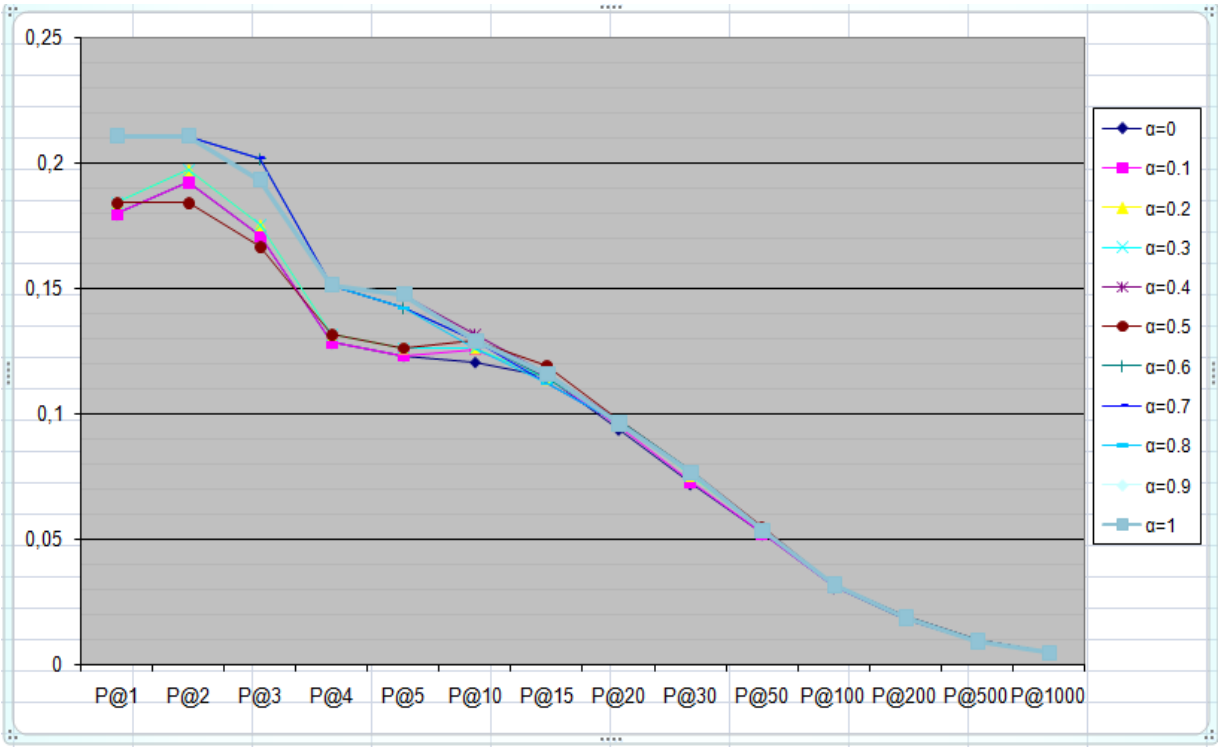

Figure IV.2 : Courbe représentant les différentes valeurs de α pour la mesure $Sim_{comb(1)}$ avec une restructuration MeSH à travers l'adjonction à une racine commune.

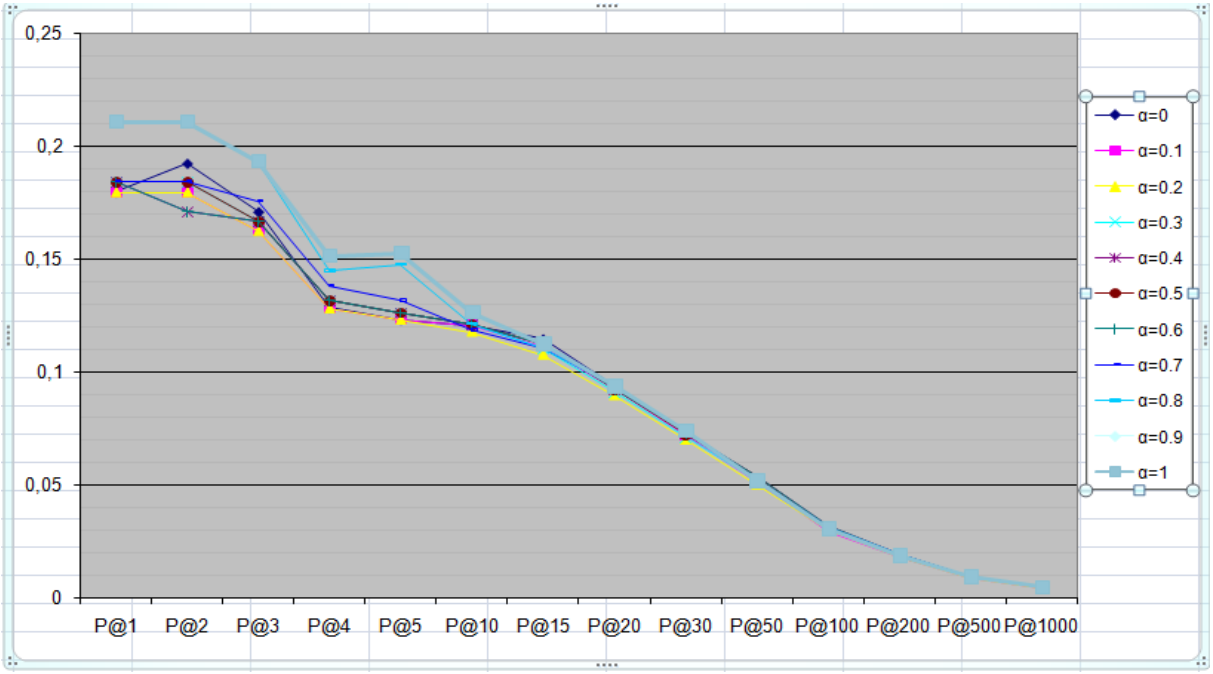

Figure IV.3 : Courbe représentant les différentes valeurs de α pour la mesure $Sim_{comb(1)}$ pour une restructuration MeSH avec mappage sur le réseau sémantique UMLS

A partir des courbes présentées ci-dessus, et des valeurs de la précision moyenne nous constatons que les valeurs de α ayant atteint les valeurs de précision maximale sont α=0.4 pour une restructuration de MeSH à travers l'utilisation d'une racine commune, et α=0.9 ou α=1 pour une restructuration de MeSH à travers le mappage de ses domaines sur le réseau sémantique UMLS. Les figures IV.4 et IV.5 présentent les graphes associés à la précision aux différents points et la précision moyenne pour la valeur α=0.4 et les figures IV.6 et IV.7 pour la valeur α=0.9 et α=1 (mêmes figures).

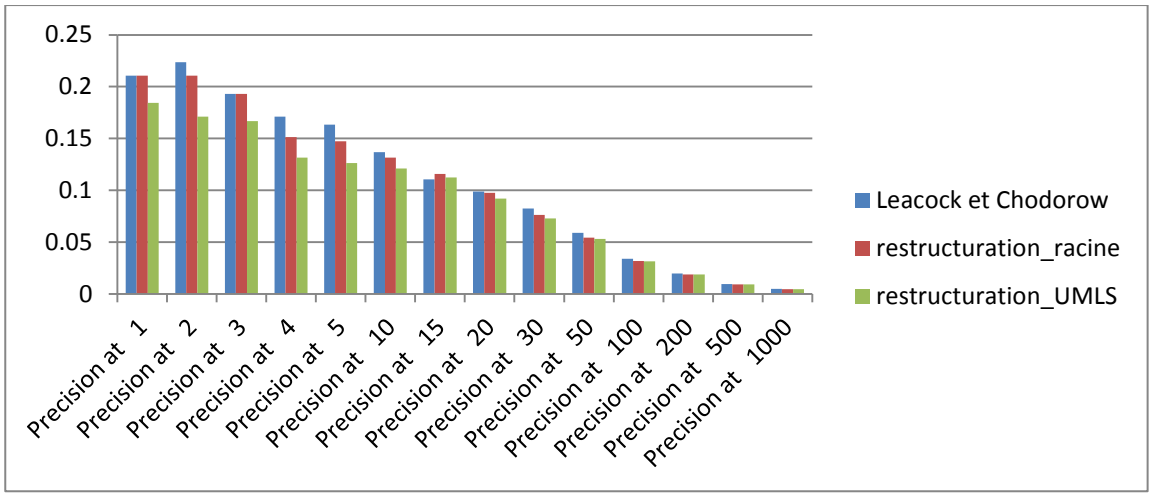

Figure IV.4 : Précision @ X pour une restructuration avec racine unique et une valeur de α=0.4

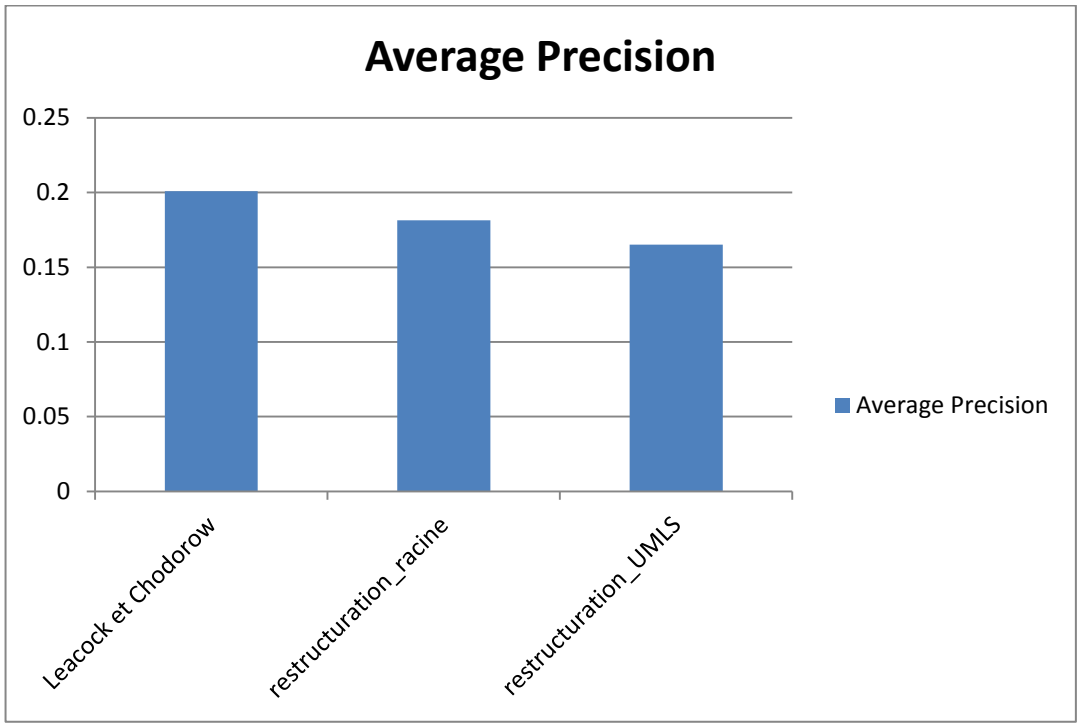

Figure IV.5 : Précision moyenne pour une restructuration avec racine unique et une valeur de α=0.4

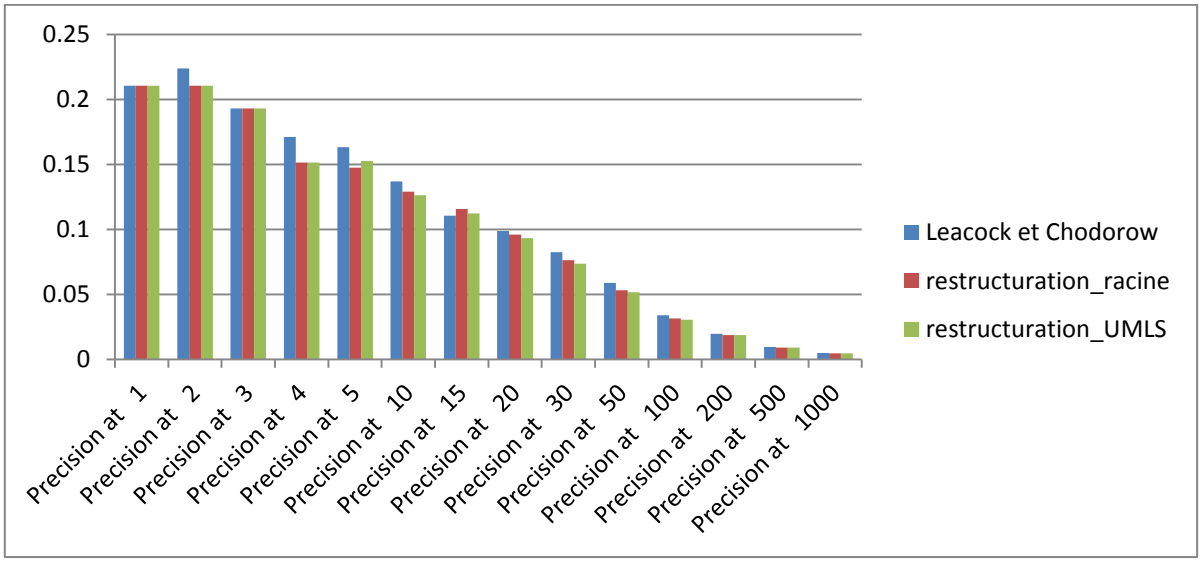

Figure IV.6 : Précision @ X pour une restructuration avec mappage sur le réseau sémantique UMLS et une valeur de α=0.9 et α=1

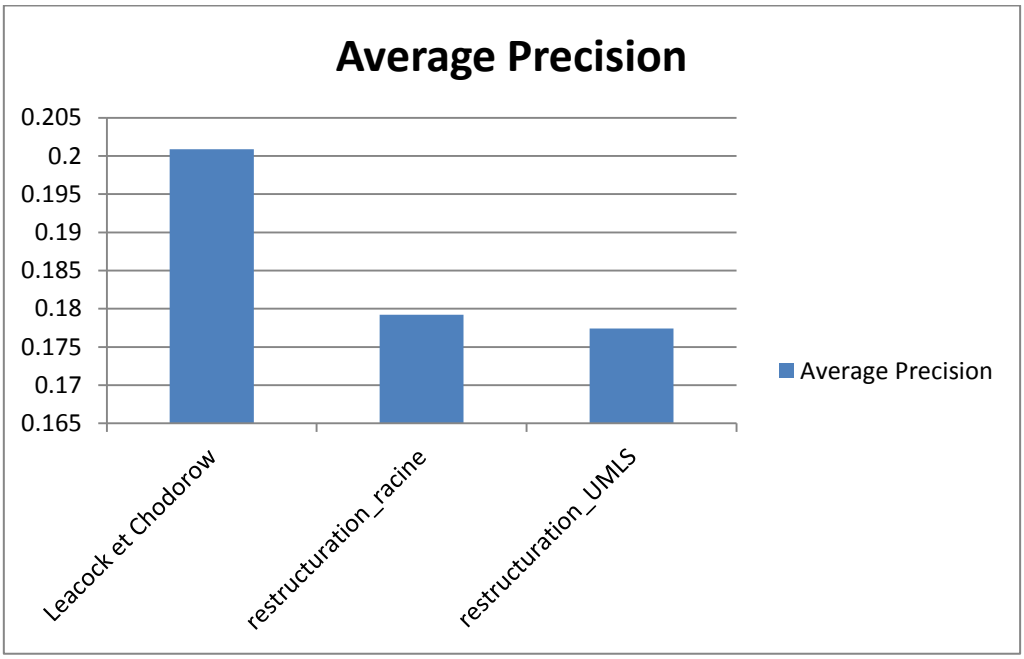

Figure IV.7 : Précision moyenne pour une restructuration avec mappage sur le réseau sémantique UMLS pour une valeur de α=0.9 et α=1

V.2. Mesure combinée $Sim_{comb(2)}$

Les tableaux suivants présentent les résultats obtenus pour l'ensemble des requêtes et les différentes valeurs de α en utilisant une mono-hiérarchisation de MeSH avec l'adjonction à une racine commune pour la précision moyenne (tableau IV.7) ainsi que la précision à différents points (tableau IV.8), et en utilisant une mono-hiérarchisation de MeSH avec un mappage des seize domaines dans le réseau sémantique UMLS pour la précision moyenne (tableau IV.9) et la précision à différents points (tableau IV.10).

Mesure	Average Précision
Sim_{LC}	0.2009
α =0	0,1602
α =0.1	0.1915
α =0.2	0.2005
α =0.3	0.1990
α =0.4	0.2030
α =0.5	0.2029
α =0.6	0.2027
α =0.7	0.2027
α =0.8	0.2023
α =0.9	0.2028
α =1	0.2009

Tableau IV.7 : Résultats de la précision moyenne pour la mesure $Sim_{comb(2)}$ et une restructuration de MeSH avec adjonction des domaines à une racine unique

Mesure	Average Précision
Sim_{LC}	0.2009
α =0	0,1602
α =0.1	0.2090
α =0.2	0.2059
α =0.3	0.2054
α =0.4	0.2051
α =0.5	0.2032
α =0.6	0.2025
α =0.7	0.2025
α =0.8	0.2025
α =0.9	0.2018
α =1	0,1998

Tableau IV.8 : Résultats de la précision moyenne pour la mesure $Sim_{comb(2)}$ et une restructuration de MeSH avec adjonction des domaines à une racine unique

	Sim_{LC}	α=0	α=0.1	α=0.2	α=0.3	α=0.4	α=0.5	α=0.6	α=0.7	α=0.8	α=0.9	α=1
P@1	0,2105	0,1795	0,2105	0,2368	0,2368	0,2368	0,2368	0,2368	0,2368	0,2368	0,2368	0,2105
P@2	0,2237	0,1923	0,1842	0,1842	0,1974	0,2105	0,2237	0,2105	0,2105	0,2105	0,2237	0,2237
P@3	0,193	0,1709	0,2105	0,2193	0,2193	0,2281	0,2193	0,2193	0,2193	0,2193	0,2193	0,193
P@4	0,1711	0,1282	0,1645	0,1776	0,1776	0,1908	0,1908	0,1908	0,1908	0,1908	0,1908	0,1711
P@5	0,1632	0,1231	0,1632	0,1684	0,1737	0,1842	0,1895	0,1895	0,1895	0,1895	0,1895	0,1632
P@10	0,1368	0,1205	0,1474	0,15	0,1474	0,15	0,1526	0,15	0,15	0,15	0,15	0,1368
P@15	0,1105	0,1145	0,1281	0,1316	0,1281	0,1281	0,1298	0,1316	0,1316	0,1298	0,1298	0,1105
P@20	0,0987	0,0936	0,1079	0,1079	0,1079	0,1079	0,1066	0,1079	0,1079	0,1066	0,1066	0,0987
P@30	0,0825	0,0718	0,0825	0,0833	0,0842	0,0842	0,0833	0,0833	0,0833	0,0833	0,0833	0,0825
P@50	0,0589	0,0518	0,0574	0,0574	0,0568	0,0568	0,0563	0,0563	0,0563	0,0563	0,0563	0,0589
P@100	0,0339	0,0305	0,0332	0,0332	0,0332	0,0332	0,0329	0,0329	0,0329	0,0329	0,0329	0,0339
P@200	0,0197	0,0179	0,0193	0,0193	0,0193	0,0192	0,0191	0,0191	0,0191	0,0191	0,0191	0,0197
P@500	0,0095	0,0089	0,0094	0,0094	0,0094	0,0094	0,0093	0,0093	0,0093	0,0093	0,0093	0,0095
P@1000	0,0048	0,0045	0,0047	0,0047	0,0047	0,0047	0,0047	0,0047	0,0047	0,0047	0,0047	0,0048

Tableau IV.9. Résultats de la Précision @ X pour la mesure $Sim_{comb(2)}$ et une restructuration de MeSH avec une racine commune.

	Sim_{LC}	α=0	α=0.1	α=0.2	α=0.3	α=0.4	α=0.5	α=0.6	α=0.7	α=0.8	α=0.9	α=1
P@1	0,2105	0,1795	0,3243	0,2973	0,2973	0,2973	0,2973	0,2973	0,2973	0,2973	0,2973	0,2973
P@2	0,2237	0,1923	0,2432	0,2432	0,2297	0,2297	0,2297	0,2297	0,2297	0,2297	0,2297	0,2297
P@3	0,193	0,1709	0,2252	0,2252	0,2252	0,2252	0,2162	0,2162	0,2162	0,2162	0,2162	0,2162
P@4	0,1711	0,1282	0,2027	0,1892	0,1892	0,1892	0,1892	0,1892	0,1892	0,1892	0,1824	0,1824
P@5	0,1632	0,1231	0,1892	0,173	0,173	0,173	0,1676	0,1676	0,1676	0,1676	0,1622	0,1622
P@10	0,1368	0,1205	0,1514	0,1459	0,1459	0,1459	0,1432	0,1432	0,1432	0,1432	0,1432	0,1405
P@15	0,1105	0,1145	0,1243	0,1189	0,1189	0,1189	0,1153	0,1117	0,1117	0,1117	0,1099	0,1099
P@20	0,0987	0,0936	0,1041	0,1	0,1	0,1	0,1	0,0986	0,0986	0,0986	0,0973	0,0973
P@30	0,0825	0,0718	0,0766	0,0775	0,0775	0,0775	0,0757	0,0757	0,0757	0,0757	0,0748	0,0748
P@50	0,0589	0,0518	0,0546	0,0568	0,0568	0,0568	0,0562	0,0562	0,0562	0,0562	0,0557	0,0557
P@100	0,0339	0,0305	0,0324	0,0327	0,0327	0,0327	0,0322	0,0322	0,0322	0,0322	0,0319	0,0319
P@200	0,0197	0,0179	0,0193	0,0196	0,0196	0,0197	0,0195	0,0195	0,0195	0,0195	0,0193	0,0193
P@500	0,0095	0,0089	0,0094	0,0094	0,0094	0,0094	0,0094	0,0094	0,0094	0,0094	0,0094	0,0094
P@1000	0,0048	0,0045	0,0047	0,0047	0,0047	0,0047	0,0047	0,0047	0,0047	0,0047	0,0047	0,0047

Tableau IV.10. Résultats de la précision à différents points X (Précision @ X) pour la mesure $Sim_{comb(2)}$ et une restructuration de MeSH avec mappage des domaines sur le réseau

Les courbes représentant les différentes valeurs de α sont données comme suit :

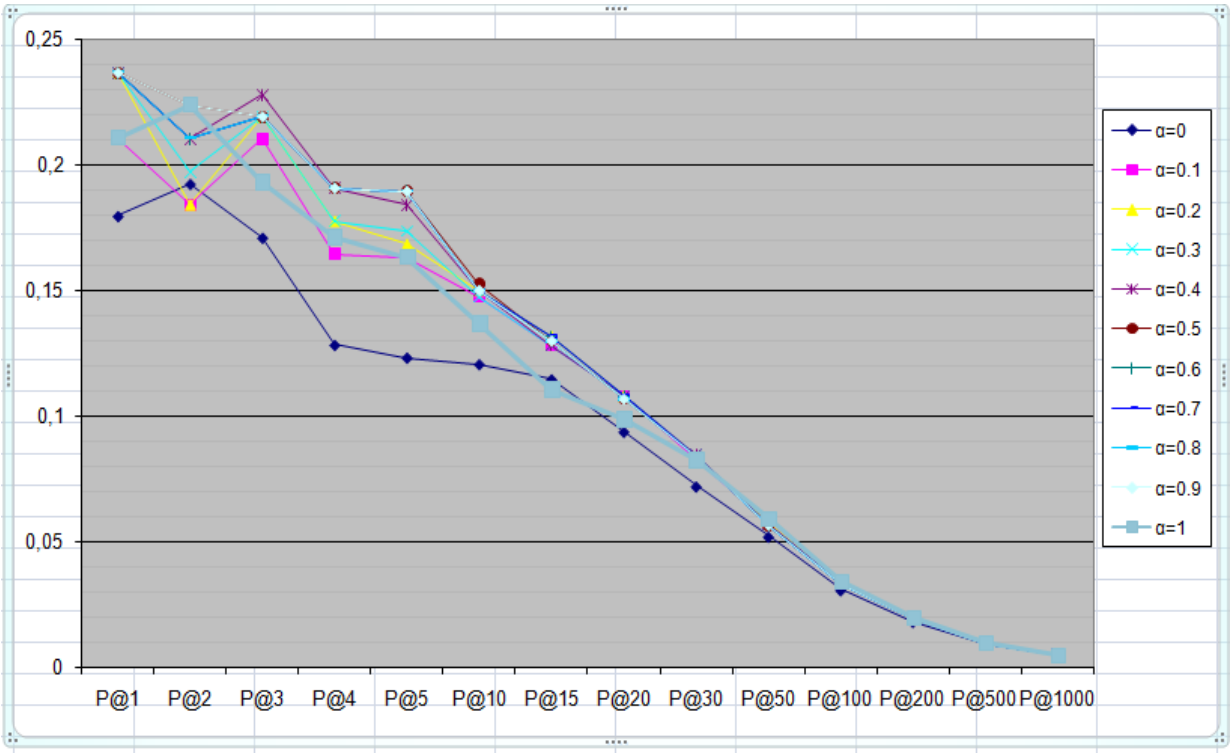

Figure IV.8 : Courbe représentant les différentes valeurs de α pour la mesure $Sim_{comb(2)}$ pour une restructuration MeSH à travers l'adjonction à une racine commune

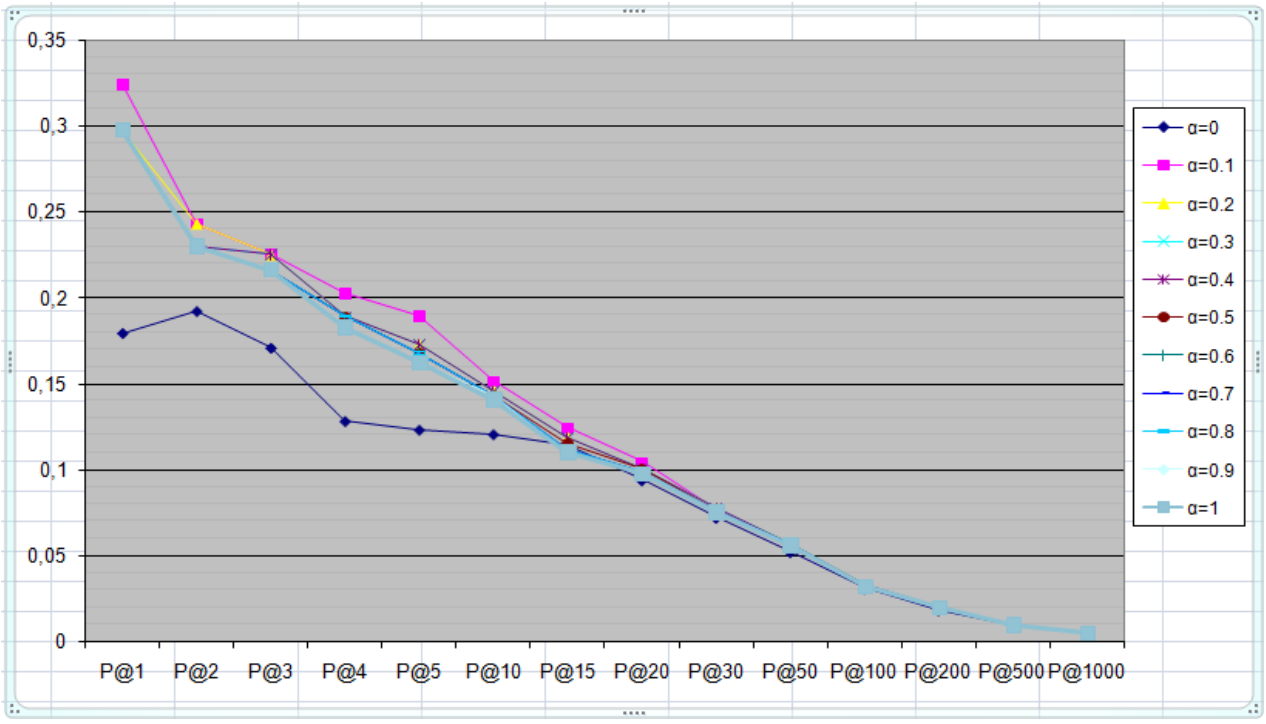

Figure IV.9 : Courbe représentant les différentes valeurs de α pour la mesure $Sim_{comb(2)}$ pour une restructuration MeSH avec mappage sur le réseau sémantique UMLS

Des courbes présentées ci-dessus, et des valeurs de précision moyenne, nous constatons que les valeurs de α ayant atteint les valeurs de précision maximale sont α=0.4 pour une restructuration MeSH à travers l'utilisation d'une racine commune, et α=0.1 pour une restructuration MeSH à travers le mappage de ses domaines sur le réseau sémantique UMLS. Les figures IV.10 et IV.11 présentent les graphes associés à la précision aux différents points et la précision moyenne pour la valeur α=0.4 et les figures IV.12 et IV.13 pour la valeur α=0.1.

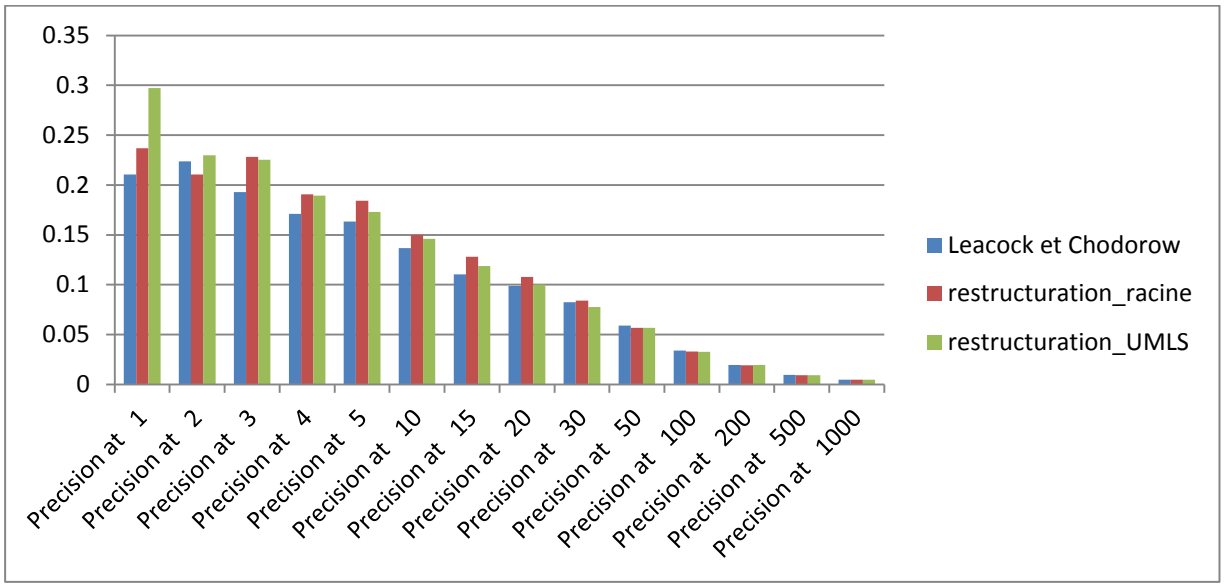

Figure IV.10 : Précision @ X pour une restructuration avec mappage sur le réseau sémantique UMLS et une valeur de α=0.4

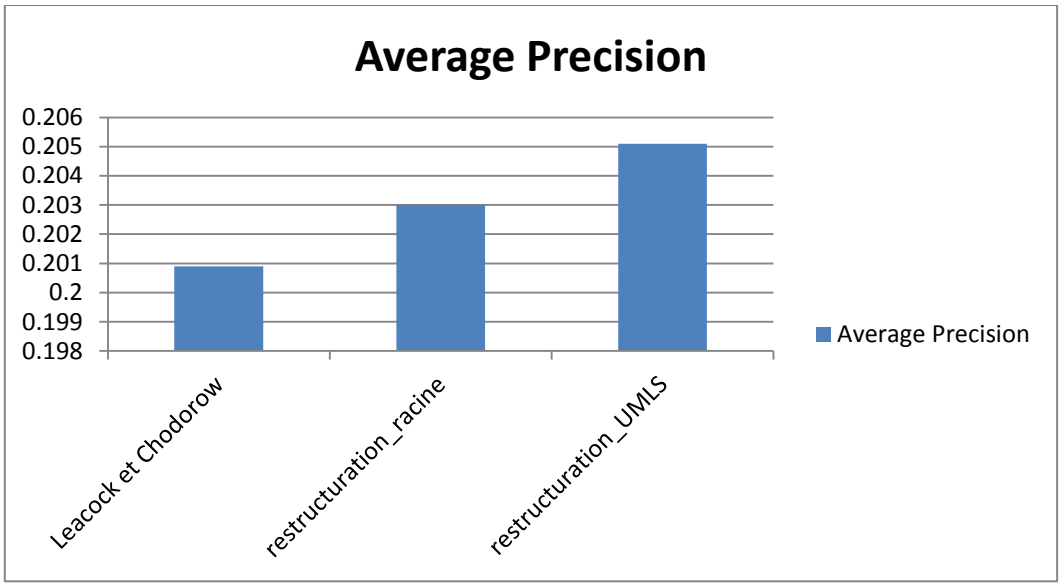

Figure IV.11 : Précision moyenne pour une restructuration avec mappage sur le réseau sémantique UMLS pour une valeur de α=0.4

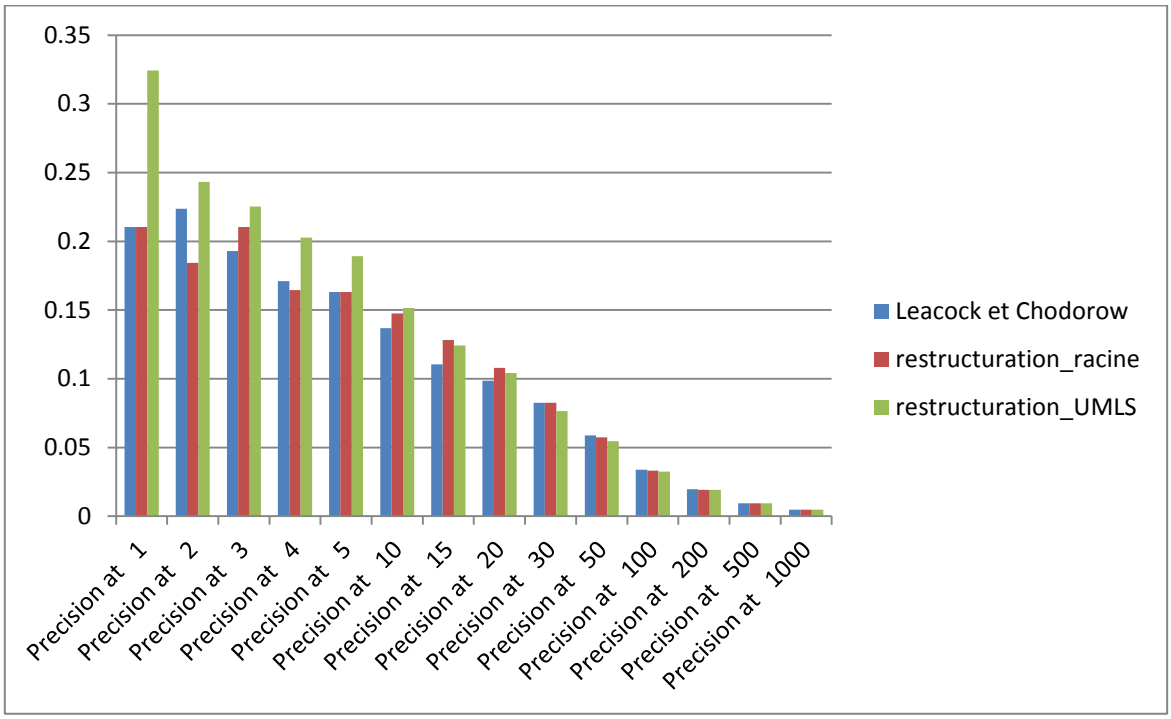

Figure IV.12 : Précision @ X pour une restructuration avec mappage sur le réseau sémantique UMLS et une valeur de α=0.1

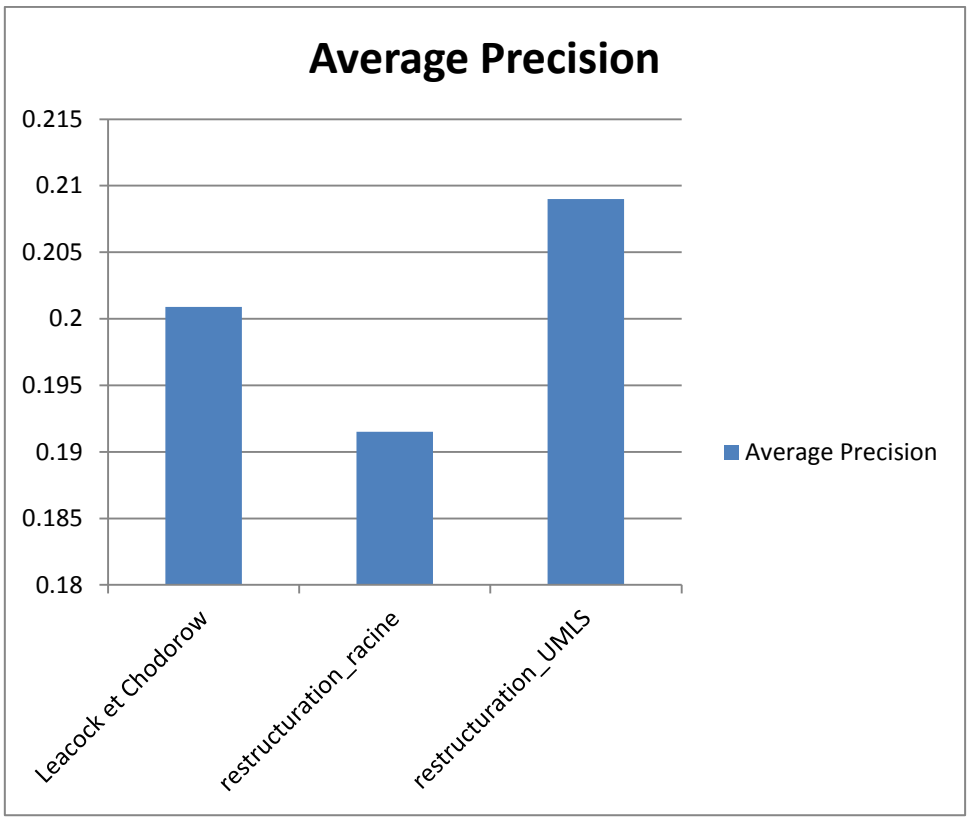

Figure IV.13 : Précision moyenne pour une restructuration avec mappage sur le réseau sémantique UMLS pour une valeur de α=0.1

A partir des résultats obtenus, nous constatons que la mesure combinée $Sim_{Comb(2)}$ donne de meilleurs résultats, et ce, pour les deux approches de restructuration de MeSH en une seule arborescence. Nous constatons que l'apport informationnel de notre mesure est très considérable.

VI. Conclusion

Dans ce chapitre nous avons présenté les étapes de notre évaluation les résultats de notre contribution pour la restructuration de MeSH en utilisant le réseau sémantique UMLS, le protocole d'évaluation ainsi que les résultats de nos expérimentations pour un ensemble de 1000 documents. Nos résultats sont très encourageants, et ce, pour les deux types d'approches de restructuration de MeSH en une seule arborescence.

CONCLUSION GENERALE

Conclusion Générale

Le travail présenté dans ce mémoire s'inscrit dans le cadre de la recherche d'information dans le domaine biomédical. Nous nous somme intéressées plus particulièrement à la notion de similarité sémantique entre concepts biomédicaux.

Notre objectif était de proposer une mesure de similarité sémantique entre concepts biomédicaux issus du thésaurus MeSH. Pour cela nous nous sommes basés sur les mesures existantes dans le domaine général à travers le réseau lexical WordNet que nous avons adapté à MeSH. Nous avons alors proposé deux approches de restructuration de MeSH en vue de cette adaptation, puis nous avons proposé une nouvelle mesure de similarité sémantique en combinant des caractéristiques structurelles et des caractéristiques informationnelles que nous avons intégré au sein d'un score de désambiguïsation avec différentes valeurs de poids associés.

Nous avons proposé un protocole d'évaluation de notre mesure de similarité dans le contexte de la recherche d'information sémantique, puis nous l'avons évaluée sur une sous collection de tests biomédicale (TREC Genomics). Les résultats obtenus sont encourageants.

En perspectives, il serait intéressant de valider nos tests sur une collection de taille réelle.

Bibliographie

[Al-Mubaid et al., 09] : Al-Mubaid H. and Nguyen H. A., Measuring Semantic Similarity between Biomedical Concepts within multiple ontologies‖, IEEE transactions on Systems, Man and Cybernetics, PART C Applications and Reviews, 2009.

[Amirouche et al., 11] : F.Amirouche, W.Azzoug, S.Chiout, M.Boghanem , Indexation sémantique de documents textuels, Ummto, Umbb, IRIT-SIG, 2011.

[Baziz, 05]: M. BAZIZ, Indexation conceptuelle guidée par ontologie pour la recherche d'information. Thèse de Doctorat en Informatique de l'université Paul Sabatier de Toulouse, décembre 2005.

[Budanitsky et al, 06]: BUDANITSKY A. & HIRST G. (2006). Evaluating wordnet-based measures of semantic distance. Computational Linguistics, 32(1), 13–47.

[Dinh et al., 10] : Duy Dinh, Lynda Tamine ; Vers un modèle de Recherche d'Information multi-terminologique des documents biomédicaux, IRIT- Université Paul Sabatier Toulouse III, 2010.

[Dinh, 12] : Accès à l'information biomédicale : vers une approche d'indexation et de recherche d'information conceptuelle basée sur la fusion de ressources termino-ontologiques. Thèse de Doctorat en Informatique de l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier).

[Fu et al. , 02] : Fu, Y., Bauer, T., Mostafa, J., Palakal, M., and Mukhopadhyay, S. (2002). Concept extraction and association from cancer literature. In *WIDM 2002*, pages 100–103.

[Hirst et al., 98]: Graeme Hirst and David St-Onge, Lexical chains as representations of context for the detection and correction of malapropisms, Department of Computer Science University of Toronto, Toronto, Ontario, Canada.

[Hliaoutakis, 05]: Angelos HLIAOUTAKIS, Semantic Similarity Measures in MeSH ontology and their application to Information Retrieval on Medline, 2005.

[Jiang et al., 97]: J.J. Jiang and D.W. Conrath. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistic*, Taiwan, 1998.

[Leacock et al, 98]: Leacock C., Chodorow M., « Combining Local Context and WordNet Similarity for Word Sense Identification », *An Electronic Lexical Database*p. 265-283, 1998.

[Lin, 93]: D. Lin. Principle-Based Parsing Without Overgeneration. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL'93)*, pages 112;120, Columbus, Ohio, 1993.

[Névéol et al., 06] : Névéol A., Rogozan A., Darmoni S., « Automatic indexing of online health resources for a French quality controlled gateway », *Inf. Process. Manage.*, vol. 42, n° 3, p. 695-709, 2006.

[Rada et al, 89]: ROY RADA, HAFEDH MILI, ELLEN BICKNELL, AND MARIA BLETTNER, Development and Application of a Metric on Semantic Nets, 1989.

[Resnik, 99]: O. Resnik. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity and Natural Language. *Journal of Artificial Intelligence Research*, 1999.

[Rodriguez, 00]: *Assessing Semantic Similarity among Spatial Entity Classes* . Thèse de doctorat, University of Maine, p.168.

[Seco et al, 04]: Nuno Seco, Tony Veale, and Jer Hayes. An intrinsic information content metric for semantic similarity in wordnet. 2004.

[Serres, 02] : Cours « Problématique Générale de la Recherche d'Information », URFIST Bretagne- Pays de Loire, Alexandre Serres, 2002. http://www.uhb.fr/urfist/

[Tversky, 77]: Amos Tversky, Feature of Similarity, Psychological Review, Hebrew University: Jerusalem, Israel, 1977.

[Voorhees, 93]: Voorhees, E. M. Using WordNet to disambiguate word senses for text retrieval. Association for Computing Machinery Special Interest Group on Information Retrieval. (ACM-SIGIR-1993) : 16thAnnual International Conference on Research and Development in Information Retrieval, 171–180. (1993).

[Wu et al., 94]: Zhibiao WU and Martha PALMER. Verb semantics and lexical selection. In 32nd. Annual Meeting of the Association for Computational Linguistics, pages 133–138, New Mexico State University, Las Cruces, New Mexico, 1994.

ANNEXE I

Cxtractor 1.0.3

L'annexe qui suit est principalement consacrée à la présentation de Cxtractor (Version 1.0.3) une plateforme de RI de haute performance et évolutive qui a comme tâche principale l'extraction de concepts biomédicaux à partir de textes.

I. Introduction

Développé par Duy Dinh [Dinh, 2012] , Cxtractor est intégré dans la plateforme BioSIR (Biomedical Semantic Information Retrieval). Il représente logiciel open source entièrement écrit en java. Dans ce logiciel sont implémentées plusieurs méthodes d'extraction de concepts à partir des documents biomédicaux, à savoir les méthodes d'extraction basées sur les modèle de RI (ex : TF_IDF, BM25, etc). Il utilise plusieurs terminologies pour l'extraction de concepts biomédicaux (ex : MeSH, SNOMED, GO ou UMLS).

II. Installation de Cxtractor

Cxtractor est téléchargé à partir du site http://www.softpedia.com/progDownload/Cxtractor-Download-230918.html. Télécharger. Il faut télécharger cxtractor-1.0.3.zip et cxtractor-ressources.zip, puis décompresser les deux fichiers et Copier les répertoires « lib » et « nlpdata » dans le répertoire « cxtractor-1.0.3/cxtractor ».

III. Structure de Cxtractor

Cxtractor contient un l'ensemble des répertoires suivants :

- bin\ : contient l'ensemble des scripts nécessaires pour démarrer Cxtractor.
- config\ : contient les fichiers de configuration de Cxtractor (le fichier settings properties sample contient la plupart des propriétés de configuration de Cxtractor).
- doc\ : contient la documentation relative à Cxtractor.
- examples\ : contient des exemples de document donnés en entrée pour l'extraction de concepts.
- lib\ : contient les différentes librairies externes utilisées par Cxtrator.
- nlpdata\ : contient les différentes ressources qu'utilise Cxtractor pour extraire les concepts (Mesh, UMLS, SNOMED, etc)
- output\ : contient les résultats lors de l'extraction.

- screenstrats\ : contient quelques captures d'écran.
- scripts\ : contient un fichier en script shell.
- src\ : contient le code source java des programmes de Cxtractor.
- tests\ : contient les collections qu'on donne pour l'extraction de concepts.

IV. Lancement de Cxtractor sur une ligne de commande

La liste complète des options pour lancer Cxtractor sur une ligne de commande est donnée comme suit :

java -jar cxtractor.jar [-r|--recursive] [-c|--clean] [-f|--file] [-d|--folder] input [-e|--doctype documentType] [-o|--output output] [-t|--terminology terminology] [-X|--cxMethod method] [-w|--wModel weightingModel] [-v|--version]

Exemple d'utilisation:

« java -jar cxtractor.jar -r -c -d tests -o output -X TerrierSpearmanExtractor Option Long Option Value (y/n) Description »

Les options sont décrites comme suit :

- -r --recursive no Recursively processing
- -c --clean no Clean all previous data
- -h --help no Print this usage information
- -f --file yes Extracting concepts from a file
- -t --terminology yesTerminology used
- -w --wModel yes Weighting model (PL2 by default)
- -X --cxMethod yes Extraction method (MaxMatcherExtractor by default)
- -d --folder yes Extracting concepts from a directory
- -e --doctype yes Document type (file, trec, html)
- -o --output yes Output directory
- -v --version no Version number

V. Entrées de Cxtractor

Les documents d'entrée peuvent être dans les formats suivants : .txt, .html, ou les formats de TREC. (il faut les configurer dans le fichier de configuration

/config/settings.properties.sample). Lors de l'exécution, les paramètres de configuration sont chargés automatiquement.

Nous donnons ci-dessous un exemple de documents sous le format de TREC. Chaque document TREC contient des balises particulières, par exemple :

- DOC représente le document,
- DOCNO représente l'identifiant unique du document,
- TITLE correspond au titre du document,
- ABSTRACT correspond au résumé du document.

Exemple de document :

<DOC>

<DOCNO>10928726</DOCNO>

<TITLE>A randomized controlled trial of Moderation-Oriented Cue Exposure. </TITLE>

<ABSTRACT>OBJECTIVE: A randomized controlled trial was conducted to examine the effectiveness of Moderation-Oriented Cue Exposure (MOCE) in comparison to Behavioral Self-Control Training (BSCT). The main hypothesis was that MOCE would be more effective than BSCT among a sample of problem drinkers aiming at moderate drinking. A subsidiary hypothesis was that MOCE would be relatively more effective than BSCT among problem drinkers with higher levels of alcohol dependence. METHOD: Clients (N = 91; 75% men) were randomly allocated to either MOCE or BSCT. Treatment was delivered in weekly sessions by two trained therapists, in a nested design in which therapists switched to the alternative treatment modality approximately halfway through the trial. Follow-up was carried out 6 months following posttreatment assessment, with 85% successful contact. RESULTS: There was no evidence for the general superiority of MOCE over BSCT. The subsidiary hypothesis was not confirmed. A subsample of clients (n = 14) showing levels of dependence at baseline above the commonly accepted cut-point for a moderation goal (Severity of Alcohol Dependence Questionnaire [SADQ] > 29) showed outcomes at least as favorable as those below the cut-point. The validity of self-reports of alcohol consumption and problems was supported by significant relationships with liver function tests (gamma-glutamyl transferase and alanine transferase). CONCLUSIONS: These results provide no grounds for the replacement of BSCT by MOCE in routine, moderation-oriented treatment practice. Assuming they prefer it to abstinence and that it is not contra-indicated on other grounds, there seems no reason why clients showing a higher level of dependence (SADQ = 30-45) should not be offered a moderation goal.

</ABSTRACT>

</DOC>

VI. Résultats de Cxtractor

Le format des résultats d'extraction retournés par Cxtractor est comme suit :

<DOCNO> DOC1

rank|CUI|concept name (preferred/non-preferred terms)|score

rank|CUI|concept name (preferred/non-preferred terms)|score

rank|CUI|concept name (preferred/non-preferred terms)|score

rank: représente le rang du concept

Où :

- **CUI :** représente l'identifiant unique du concept
- **concept name (preferred/non-preferred terms):** le nom du concept préféré.
- **Score :** le poids du concept dans le document.

VII. Exemple d'utilisation de Cxtractor

Avant de lancer l'exécution il fau placer le dossier contenant les documents à partir des quels on veut extraire les concepts dans le répertoire « ….\cxtractor-1.0.3\cxtractor\tests\trec »

1. Lancer la commande : « **cd dossier_de_cextractor\cxtractor-1.0.3\cxtractor »**
2. **Lancer** la commande : « **java -jar cxtractor-1.0.3.jar -r -c -d tests/trec/test** »
   test représente le dossier contenant les documents à partir desquels on veut extraire les concepts. Cette commande va utiliser par défaut le thésaurus MeSH pour l'extraction de concepts. Si on veut utiliser une autre terminologie on peut spécifier dans la commande la terminologie à utilisé exemple : java -jar cxtractor-1.0.3.jar -r -c -d tests/trec/ -o output/ -t snomed.
   Cette figure montre l'utilisation de ces deux commandes :

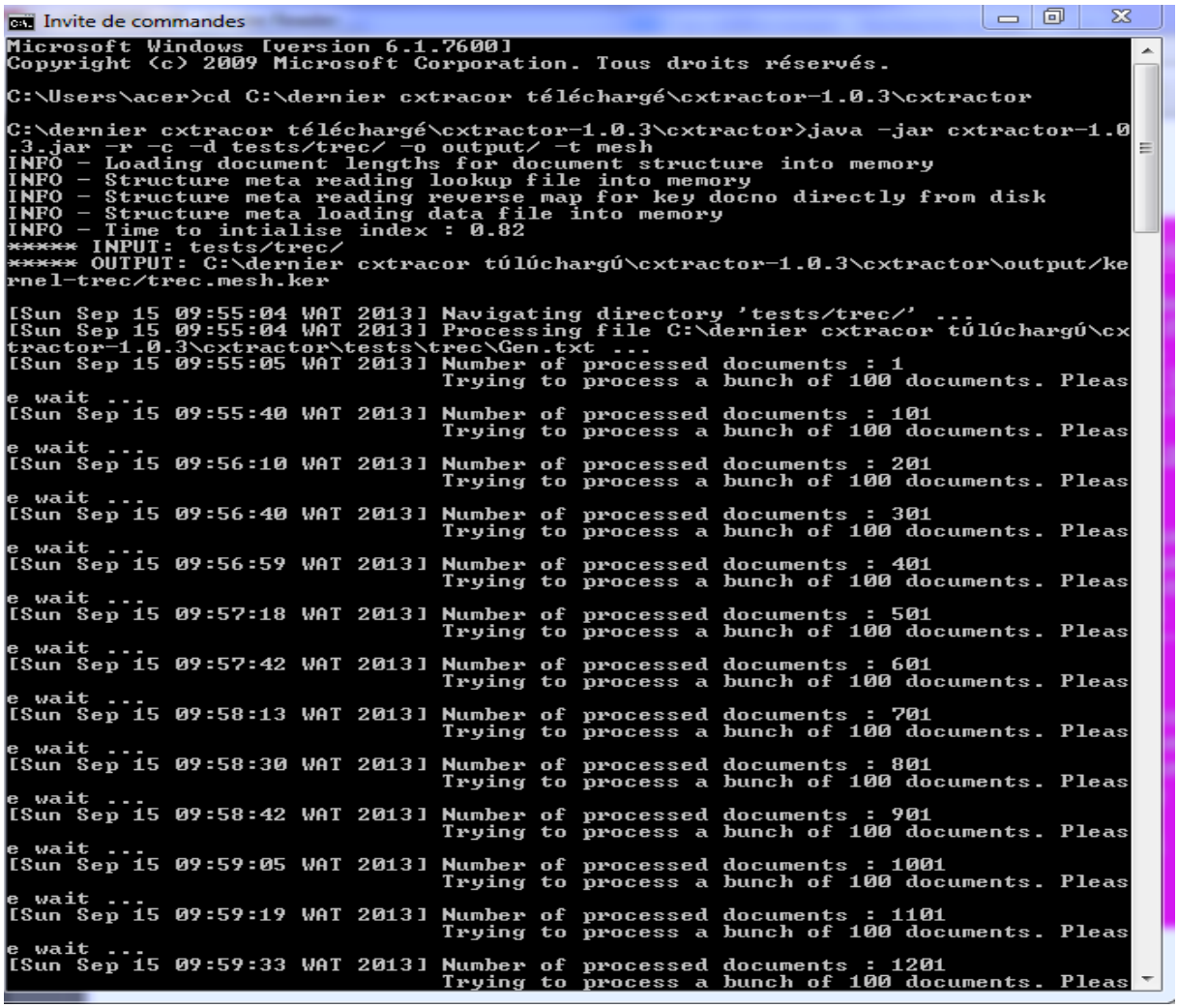

Figure 1 : Lancement de Cxtractor

Les résultats de l'extraction seront sauvegardé dans le répertoire « \cxtractor1.0.3\cxtractor\tests\trec\test » dans le fichier « trec.mesh.txt ». Ce fichier contient les résultats d'extraction. A ce fichier est associé un fichier nommé « trec.mesh.ker.doccount » qui contient le nombre de documents traités pour l'extraction.

Par exemple, pour le document « 10928726 » présenté plus haut les résultats sont les suivants :

```
<DOCNO>  10928726
0|C0001898|Alanine|24,9351
1|C1096777|Randomized Controlled Trial|20,3211
2|C0023901|Liver Function Tests|19,9158
3|C0034394|Questionnaires|19,6674
4|C0001948|Alcohol Drinking|16,9139
5|C0025663|Methods|16,4256
6|C0011546|Dependency (Psychology)|16,3984
7|C0376287|Behavior Control|16,2367
8|C0684271|Drinking|14,4327
9|C0010439|Cues|13,3298
10|C0018017|Goals|11,7591
11|C0030705|Patients|10,4221
12|C0025266|Men|9,5162
13|C0039796|Therapeutics|8,0145
14|C0040676|Transferases|7,7171
15|C0001975|Alcohols|6,3566
```

ANNEXE II

TERRIER 3.5

I. Présentation

TERRIER (TERabyte RetrIEveR) est un moteur de recherche robuste et efficace, développé par le département informatique de l'université Glasgow de Scotland. Utilisé avec succès dans :

- La recherche Ad-hoc
- La recherche web
- La recherche multilingue

Terrier est un logiciel Open Source écrit en Java. Il offre une plate forme idéale destinée à l'indexation de volumes importants de documents: jusqu'à 25 millions de documents.

Comme tous moteurs de recherche, terrier permet :

- L'indexation classique : Extraction des mots clés des documents appartenant à une collection et les stocker dans un index.
- Recherche des documents pertinents pour répondre aux requêtes formulés par l'utilisateur.
- Evaluation des résultats de la recherche.

Terrier 3.5 est téléchargeable sur le lien : http://www.terrier.org. Il contient un l'ensemble des dossiers suivants permettant son utilisation :

- bin/ Ensemble de scripts pour exécuter terrier
- doc/ Documentation relative à terrier
- etc/ Fichiers de configuration de terrier, le fichier **terrier.properties.sample** contient la plupart des propriétés de configuration de terrier
- lib/ Classes compilés de terrier et les différentes librairies externes utilisées par terrier

- share/ Liste des mots vides (stopword-list.txt) et des exemples de documents à tester sur terrier
- scr/ Code source Java de terrier
- var/
    - ✓ index/ Structures de données après indexation (fichier inverse, fichier lexicon, index direct, index documents)
    - ✓ result/ Résultats de la recherche et l'évaluation

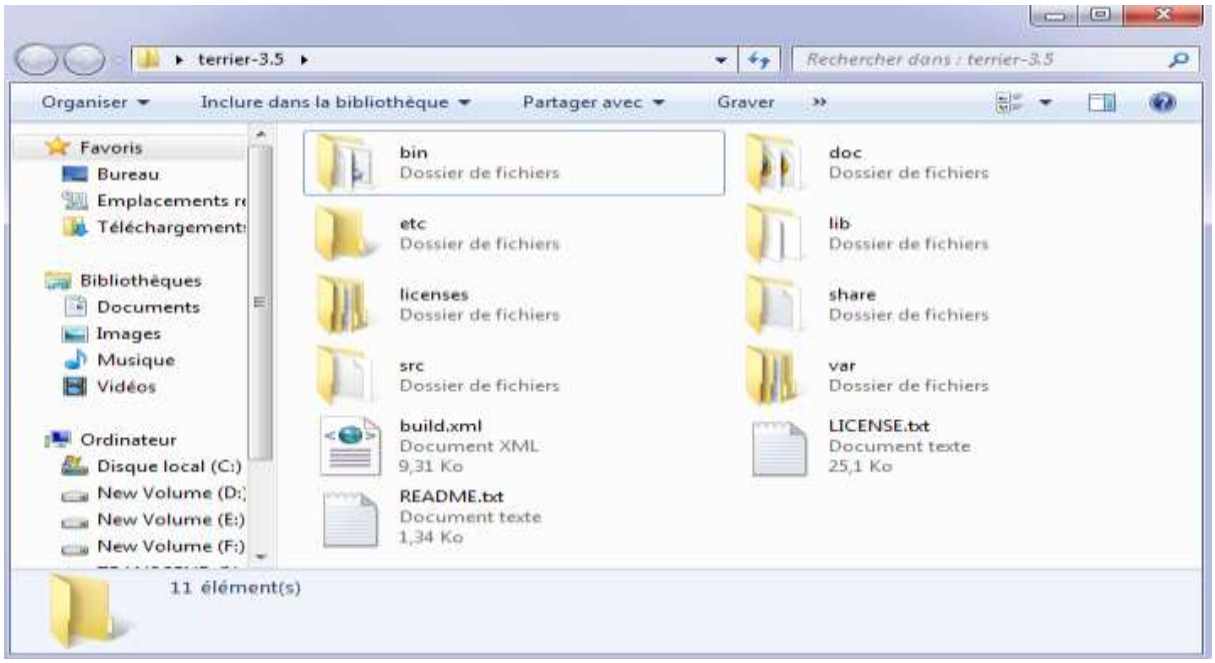

Figure1 : Dossiers de Terrier 3.5

II. Architecture de Terrier

L'architecture de Terrier est résumée dans la figure suivante :

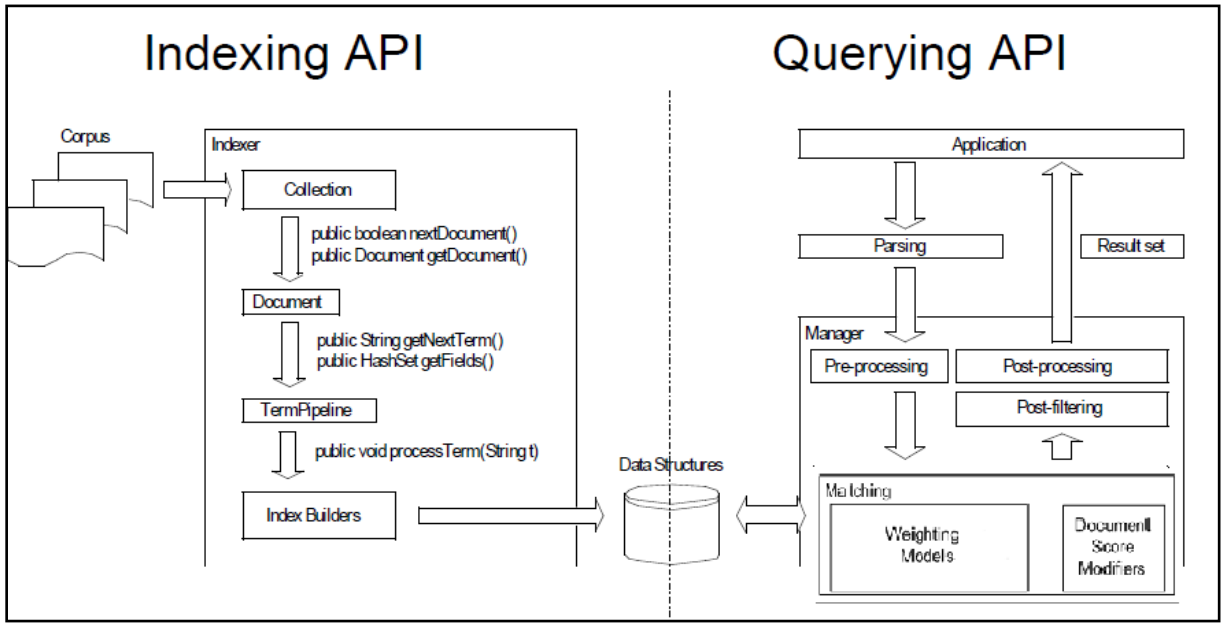

Figure 2 : Architecture de Terrier

L'architecture de l'indexation est illustrée dans la figure suivante :

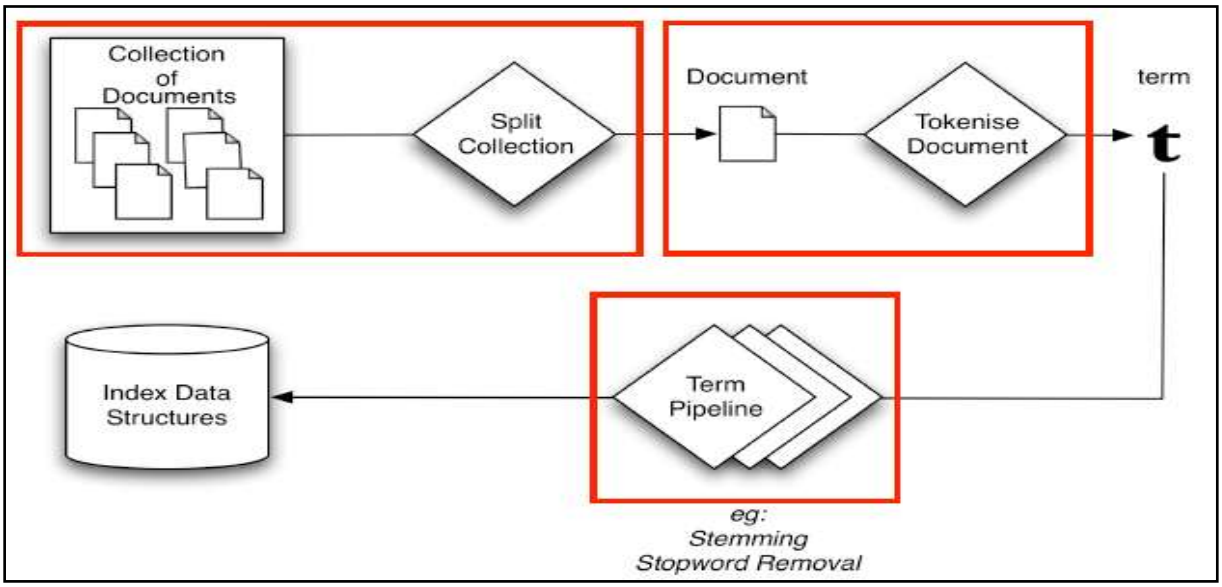

Figure 3 : Architecture d'indexation de Terrier

III. Utilisation de Terrier

Dans cette section, nous décrivons l'utilisation de Terrier pour l'indexation, la recherche et l'évaluation :

III.1. Indexation

L'indexation dans terrier requiert les étapes suivantes :

1. Supprimer le contenu du dossier index dans var dans terrier (s'il est plein).
2. Initialiser de terrier pour une nouvelle indexation (collection ou corpus Trec) en utilisant la commande **« trec-setup <path de la collection ou le corpus à indexer>** ».
3. Indexer avec la commande : **trec_terrier –i**

III.2. Recherche

La recherche des documents pertinents dans terrier passe elle aussi par plusieurs étapes comme suit :

1. Ajouter la propriété trec.topics dans **terrier.properties:**

   trec.topics=<Path vers le fichier txt contenant les requêtes>
2. Dans l'invité de commandes exécuter: **trec_terrier –r**
3. Le fichier résultat (.res) est dans ..\terrier-3.5\var\result\ x.res

III.3. Evaluation

L'évaluation Selon le protocole TREC se fait en suivant les étapes suivantes :

1. Pour chaque requête, les 1000 premiers documents restitués par le système sont examinés.
2. Les précisions P@x à différents points x (x = 5, 10, 20, 50, 100, 200, 500, 1000) ainsi que la précision moyenne Average Precision sont calculées.
3. Spécifier dans **terrier.properties** le chemin vers le Rel_Ass, qui contient les documents pertinents pour chaque requête. **trec.qrels=<path vers fichiers Rel_Ass.qrels>**
4. Dans l'invité de commandes exécuter: **trec_terrier -e**
5. Le fichier évaluation (.eval) est dans : ..\terrier-3.5\var\result\ x.eval 59