République **A**lgérienne **D**émocratique et **P**opulaire
Ministère de l'**E**nseignement **S**upérieur et de la **R**echerche **S**cientifique

UNIVERSITE **M**OULOUD **M**AMMERI DE **T**IZI-**O**UZOU

FACULTE DU **G**ENIE **E**LECTRIQUE ET D'INFORMATIQUE
DEPARTEMENT D'INFORMATIQUE

Mémoire de Fin d'Etudes de MASTER ACADEMIQUE

Domaine : **Mathématiques et Informatique**
Filière : **Informatique**
Spécialité : **Conduite de projet informatique**

Présenté par
Younes MEZAOUR
Arezki HAFID

Thème

Désambiguïsations des termes ambigus dans des documents textuels.

Mémoire soutenu publiquement le ……/……/ 2016 devant le jury composé de :

Président : M Prénom NOM

Encadreur : M^elle^ Samia Iltache
Co-Encadreur : M Prénom NOM

Examinateur : M Prénom NOM

Examinateur : M Prénom NOM

C’est avec un immense plaisir que je dédie ce travail,

A ma fierté, ma chère et mon adorable mère que j’adore, pour son sacrifice, son soutien tout au long de mes études et pour toutes les valeurs magnifiques qu’elle m’a donné durant toute ma vie,

A mon cher père et ma chère adorable sœur « **Massilva**»,

Dieu les gardes pour moi,

A ma très cher grand mère **Fatma,**

que dieu la garde et la protège

A ma chère fiancée que je n’ai pas encore rencontré

A ma très cher ami **Younes** et toute sa famille,

A mes camarades **Kidousse ,djura, Hicham**

et tout ce qui sont partis a l’étranger.

A tous mes amis sans exception ;

Juba,lhadi,Amar,idir,amirouche,boukhalfa, nordine,mouloud,yahia

A tous mes coéquipiers **Ghiles,Arezki** et a toutes la famille **OTM** joueurs, entraineurs et dirigeants.

Arezki

C’est avec un immense plaisir que je dédie ce travail,

A ma chère et mon adorable mère que j’adore et à mon cher père pour leur sacrifice, leur soutien tout au long de mes études et pour toutes les valeurs magnifiques qu’ils m’ont inculqués durant toute ma vie,

A mon cher frère « **Juba**», et mes chers sœurs « **Nadia, Tassadit et Sabrina**»,
Qui m’ont toujours soutenus et crus en moi,
A la mémoire de mon cher cousin « **Kamel**» et mes grands-parents.

A mon très cher ami **Arezki** et toute sa famille,
A tous mes amis sans exception ;
Djura,Kidousse,lhadi,amar,idir,amirouche,boukhalfa, nordine,mouloud,yahia ainsi que tous ceux que je n’ai pas pu citer.

Younes

Sommaire

Chapitre II: désambigüisation

Chapitre III: Conception et réalisation

Introduction Générale

Introduction générale

Actuellement, le monde connait une avancée technologique considérable dans tous les secteurs et cela grâce à l'informatique qui est une science qui étudie les techniques du traitement automatique de l'information. Elle joue un rôle important dans la société d'information d'aujourd'hui.

L'ambiguïté inhérente aux langues naturelles est un problème récurrent dans le domaine du Traitement Automatique du Langage. On peut, en effet, rencontrer différents types d'ambiguïté en fonction du niveau d'analyse linguistique où l'on se situe : au niveau syntaxique, du fait des différentes manières possibles d'agencer ceux-ci dans une même langue (catégories syntaxiques, problèmes de rattachements), au niveau sémantique, avec les différents types d'ambiguïtés lexicales dues aux différents sens des mots (homonymies, polysémie, etc.). S'ajoute à cela le fait qu'une langue peut faire l'objet de différents types d'usages, avec des conséquences importantes sur la manière de gérer les informations.

Les ambiguïtés rencontrées par un système de traitement automatique des langues peuvent être levées par une analyse linguistique complète de leur contexte d'apparition ;
c'est le principe de **la désambiguïsation.**

La désambiguïsation sémantique des mots est une tâche intermédiaire fondamentale pour la plupart des applications de traitement automatique du langage telles que la traduction automatique, la recherche d'information, l'acquisition automatique de connaissances, la compréhension automatique, l'interaction homme-machine, le traitement de la parole, etc.

D'une manière générale, la désambiguïsation sémantique des mots consiste à associer une occurrence donnée d'un mot ambigu avec l'un des sens de ce mot.

L'objectif de notre projet présenté dans ce mémoire s'inscrit dans le contexte de la désambigüisation des termes ambigus dans des documents textuels. Notre travail consiste à faire une comparaison entre deux algorithmes de désambiguïsation, à savoir l'Algorithme de désambiguïsation *Left to right et l'*Algorithme de désambiguïsation par translation d'une fenêtre de termes, dans le but d'analyser les résultats afin d'évaluer les performances de chaque algorithme.
Par conséquent, on utilisera pour la réalisation de ce travail le corpus brown et le corpus semcor.

Notre mémoire est structuré en quatre chapitres. Le premier chapitre présente l'indexation, ses concepts et ses différentes étapes. Le deuxième chapitre présente l'essentiel de notre travail, commençant par l'introduction, présenter des approches de désambiguïsation et illustrer et détailler les deux algorithmes utilisés. Puis la description des Mesures permettant le calcul de Similarité entre concepts appartenant à une ressource externe.
Enfin, le dernier chapitre est consacré à l'implémentation des deux algorithmes utilisés, leur évaluation sur le corpus **Brown** puis comparaison de leurs résultats en terme de précision. Finalement, analyse des résultats de la comparaison, et évaluation des performances des algorithmes implémentés, représentation de l'environnement de développement **(NetBeans)**, et le corpus **brown** et le corpus **semcor**.

Chapitre I
INDEXATION

CHAPITRE 1 : INDEXATION

1. Introduction :

La numérisation des documents et le développement des technologies internet engendre une augmentation incessante de la masse de documents disponibles; Face à cette masse documentaire, le futur lecteur se sent désorienté et a besoin d'outils pour l'aider à filtrer les documents pour accéder aux documents pertinents. l'indexation permet de donner a un document une représentation exploitable automatiquement par une machine.

1.2. Définition

- L'indexation consiste à donner accès aux documents à partir d'une indication concernant leur contenu et/ou leur nature (forme, type).

- Dans la recherche d'information classique, l'indexation consiste à extraire à partir de l'information textuelle de documents et de la requête un ensemble de descripteurs de manière à faciliter la recherche et à la rendre plus efficace. Ces descripteurs sont des unités textuelles significatives dans le document. Dans une indexation classique, les descripteurs d'un document peuvent être des mots simples ou des mots composés.

1.3. Problématique de l'indexation

L'indexation par des mots clés est généralement imprécise. Cette imprécision est due au problème de l'ambiguïté sémantique des mots du langage naturel. En effet, un même mot peut posséder plusieurs sens et différents mots peuvent avoir une même signification avec une syntaxe différente. De ce fait, des documents bien qu'ils soient pertinents et contenant des mots sémantiquement équivalents mais lexicalement différents (synonymes) des mots de la requête, ne seront pas retrouvés. Par ailleurs, des éléments non pertinents, contenant des mots lexicalement identiques mais sémantiquement différents (homonymes et polysemie) des mots de la requête seront retournés à l'utilisateur.

Une solution pour palier aux limites de l'indexation à base de mots clés est la prise en compte de la sémantique des termes d'indexation. Ce type d'indexation passe du niveau des mots au niveau des concepts pour mieux décrire le contenu d'un document (ou d'un élément) et de la requête. Ainsi on parle d'indexation sémantique ou conceptuelle où l'appariement consiste à comparer la représentation conceptuelle de la requête avec les représentations conceptuelles des documents. Contrairement aux systèmes classiques à base de mots clés, dans une représentation conceptuelle l'appariement peut exploiter les relations entre concepts. Ce qui peut améliorer la performance des systèmes de RI. En effet, le rapprochement entre le document et la requête se fait via des concepts similaires mais pas nécessairement « identiques ». La RI sémantique ou conceptuelle est caractérisée par l'utilisation des ressources sémantiques (thésaurus, ontologies, etc.) dans la phase d'indexation .

2. Le processus d'indexation

2.1 Qu'est ce que l'Indexation en RI :

L'indexation est le processus qui consiste à **décrire** et à **caractériser** un document à l'aide de la représentation du contenu de celui-ci. Sa finalité est :

- d'indiquer dans une forme concise la teneur du document,
- de permettre une recherche efficace des informations contenues dans une collection de documents sans avoir à analyser chaque texte de document à chaque interrogation ou recherche.

Mathématiquement, un index est une *relation* qui relie chaque document à l'ensemble des **mots clés** ou **descripteurs** décrivant le thème qu'il **traite** (Aboutness):

$$Index : doc_i \{kw \longrightarrow traite_j\}$$

La relation inverse permet de capturer, pour chaque mot clé, le document qu'il **décrit**:

$$Index^{-1} : kw_j \{doc \longrightarrow décrit_i\}$$

Ou :
Kw représente les mots composant un document.

Dans un processus de RI, la requête et les documents à l'état brut sont difficilement exploitables. Afin de rendre la recherche possible, une étape primordiale s'avère nécessaire. Cette étape consiste à construire une représentation interne pour chaque document de la collection et de même pour la requête. Ces représentations seront utilisées ultérieurement (dans la fonction de correspondance) par le SRI. Pour ce faire des techniques et des modèles sont mis en œuvre. Ces techniques permettent de décrire les documents et la requête par un ensemble de termes d'indexation ou de descripteurs. Ces descripteurs reflètent au mieux le contenu du document. Cette étape est appelée l'indexation.

Ainsi, l'indexation consiste à analyser les documents et la requête afin d'extraire un ensemble de descripteurs (Salton, 1970), (Rijsbergen, 1979). Ces descripteurs sont des unités qui possèdent un pouvoir discriminant dans le document. Le processus d'indexation peut être manuel, semi-automatique ou automatique.

2.1.1 Pondération des Termes

La pondération est l'une des fonctions fondamentales en RI. Elle est la clé de voûte de la majorité des modèles et approches de RI proposés depuis les années 1960. Le poids d'un terme dans un document traduit l'importance de ce terme dans le document. Si certaines méthodes proposent d'introduire des éléments linguistiques dans l'indexation des documents, la grande majorité des approches et systèmes opérationnels, se base sur les aspects statistiques. Ces méthodes tirent leur origine de la loi de Zipf et de la conjecture de Kuhn.

2.2 Les techniques d'indexation

2.2.1. indexation manuelle

Dans une **indexation manuelle,** chaque document de la collection est examiné par un spécialiste du domaine ou un documentaliste afin d'identifier les descripteurs Malgré que l'indexation manuelle donne des résultats jugés satisfaisants (Ren et al., 1999), ce type d'indexation est en général, difficile à adopter. En effet, grâce au développement des nouvelles technologies le nombre de documents croît d'une manière incessante ce qui nécessite un temps important pour les indexer.

2.2.2. indexation automatique

Dans une **indexation automatique,** le processus d'indexation est entièrement informatisé. Il met en œuvre des méthodes et des techniques issues des Traitements Automatiques de la Langue Naturelle (TALN).
Comparés aux résultats obtenus par une indexation manuelle, les résultats obtenus par une indexation automatique sont souvent jugés insatisfaisants (Jacquemin et al., 2002). Pour remédier à cette limite, dans (Jacquemin et al.,2002), les auteurs proposent de coupler l'indexation automatique avec l'indexation manuelle. Ainsi, les résultats obtenus par une indexation automatique sont exposés à un documentaliste pour les valider ou pour les enrichir. Ce type d'indexation est appelé **indexation semi-automatique** ou indexation supervisée.

En général, l'indexation automatique se fait en plusieurs étapes :

1. L'analyse lexicale (tokenization en anglais) : consiste à découper le document en unités lexicales. Chaque unité lexicale est une séquence de caractères entourée par des séparateurs d'unités.

2. L'élimination des mots vides : dans cette étape les mots d'usage général et grammatical (les mots vides) sont éliminés. Du fait que ces mots apparaissent d'une manière uniforme dans les documents ils sont non utiles pour l'indexation et ils doivent être éliminés. On distingue deux techniques pour l'élimination des mots vides : l'utilisation des stoplist ou des anti-dictionnaires et l'utilisation des mesures statistiques.

3. La normalisation : la lemmatisation consiste à prendre la forme canonique du mot. Dans le document les mots peuvent apparaître sous différentes formes. Par exemple, citer, citation, citations, etc. La normalisation permet de substituer chaque mot par sa racine.la lemmatisation est l'une des formes de normalisation dans ce cas,la racine d'un mot est soit la forme infinitive si le mot est un verbe, soit la forme singulier si le mot est un nom. L'utilisation de la lemmatisation contribue à l'amélioration des performances des SRI[1] .

4. L'analyse syntaxique et morphosyntaxique : L'objectif de cette étape est de repérer des mots composés. Cette analyse syntaxique se base sur des patrons (templates) pour extraire les

[1] SRI :Système de Recherche d'information

mots composés. Dans ce processus une catégorie grammaticale est associée à chaque mot (ou groupe de mots) et des patrons sont manuellement construits. Ces patrons sont ensuite projetés dans les documents afin de détecter les séquences qui satisfont ces patrons syntaxiques.
Nous signalons que dans la littérature il existe d'autres méthodes dites méthodes statistiques qui sont utilisées pour l'extraction des mots composés. Ces méthodes utilisent des mesures statistiques et n'utilisent pas d'analyse linguistique.

2.3 Indexation sémantique et indexation conceptuelle en RI:

Pour remédier aux limites de l'indexation classique basée sur des mots clés, deux alternatives peuvent être distinguées : l'indexation sémantique et l'indexation conceptuelle

1. L'indexation sémantique en RI se base historiquement sur les algorithmes de désambiguïsation (Word Sense Disambiguation WSD) pour affecter un sens à un mot. Alors que l'indexation conceptuelle se base sur des méthodes d'identification de concepts dans un corpus textuel (appariement de concept) en utilisant une ressource sémantique externe comme des ontologies et des thésaurus.

2. L'indexation conceptuelle se réfère à la construction de structures conceptuelles à partir des textes. Cette structure rend possible une extension de la représentation des documents (ou requêtes) via les différentes relations sémantiques.
Nous considérons que l'indexation conceptuelle peut être vue comme une généralisation de l'indexation sémantique, dans la mesure où les concepts aussivéhiculent des sens des mots ou les termes.

3. Utilisation des ressources sémantiques externes

Nous avons vu dans la section précédente que les concepts peuvent être utilisés dans le processus d'indexation pour représenter les documents à la place des mots simples. Avec l'avènement des ontologies et autres ressources sémantiques externes, tels les réseaux sémantiques, des méthodes d'identification des descripteurs (d'indexation) basées sur les concepts sont apparues. La ressource lexico-sémantique (concepts + relations sémantiques) peut alors être utilisée pour identifier dans le texte du document, les termes qui correspondent aux concepts de cette ressource. Ces techniques ont l'avantage de s'affranchir des aléas de la combinaison des groupes de mots, comme vu précédemment, avec des règles morphosyntaxiques ou statistiques. En effet, ces règles génèrent du bruit dans la mesure où elles regroupent des mots sans que le groupe ait un sens, même s'ils semblent corrects d'un point de vue syntaxique ou co-occurrent dans le texte. L'utilisation du vocabulaire contrôlé issu d'une ressource externe, permet donc d'éliminer ces "faux termes". Dès lors, le problème franchit cette contrainte et touche à d'autres problèmes encore plus fins. Il ne s'agit plus seulement de trouver les termes composés qui occurrent dans les documents ou les requêtes, mais de trouver un moyen de les attacher à des concepts de l'ontologie ou du réseau sémantique (Concept Mapping). Ce qui peut présenter des avantages à plusieurs égards. D'une part, ceci offre la possibilité d'enrichir les représentations des requêtes et des documents avec des concepts de l'ontologie liés sémantiquement à ceux des documents et des requêtes. D'autre part, la ressource offre ainsi un moyen de représenter les documents et les requêtes dans un même référentiel.

L'inconvénient de ces méthodes est qu'elles supposent que tous les concepts des documents sont couverts par la ressource. D'où la nécessité d'avoir une ressource sémantique disposant d'une terminologie suffisamment riche pour couvrir le domaine dont traitent les documents de la collection. Dans le cas de domaine général, la base de données lexicographique WordNet est une des ressources les plus utilisées en RI, notamment dans le cas de collections de documents de type "news", tels que les journaux et magasines. Dans le cas de domaine spécifique, tel le domaine médical, le réseau conceptuel médical UMLS est un des plus utilisés. Dans Medline, les documents sont décrits par au moins une quinzaine de termes médicaux représentant des concepts du thesaurus MeSH1. L'objectif est d'unifier la terminologie utilisée par les indexeurs pour la description des documents et d'assurer ainsi leur couverture par MeSH.

3.1 Types de ressources sémantiques

Différents types de ressources sémantiques peuvent être distingués parmi lesquels se trouvent les taxonomies, les thésaurus, et les ontologies.

1 .***Taxonomie***

La taxinomie est la forme la plus simple des vocabulaires contrôlés, elle représente sous la forme d'une hiérarchie simple de terme Les liens hiérarchiques dans une taxonomie correspondent à des liens de spécialisation /généralisation.

2 .Thésaurus

Un thésaurus constitue un dictionnaire hiérarchisé des vocabulaires contrôlés. Ce vocabulaire est normalisé, il présente les termes génériques ou spécifiques à un domaine de connaissances. Ces termes dénotent les concepts d'un domaine particulier. Dans un thésaurus, les termes sont organisés dans une hiérarchie de concepts liés par des relations sémantiques. Les relations couramment présentes dans un thesaurus sont des relations taxonomiques (spécialisation/généralisation), d'équivalence (synonymie), d'association (proximité sémantique, proche-de, relié-à, …).

3. Ontologie

La définition la plus citée présente une ontologie comme étant « une spécification explicite et formelle d'une conceptualisation partagée » (Gruber, 1993). En d'autre terme, une ontologie est une représentation formelle d'un domaine. C'est une conceptualisation, dans le sens où elle fournit un vocabulaire formalisé de concepts et de leurs relations.

On distingue deux types d'ontologie : les ontologies légères et les ontologies lourdes. Ces ontologies se distinguent par la présence ou non d'axiomes (Mothe et al., 2007). Les ontologies légères sont constituées uniquement de concepts et de relations entre les concepts. Contrairement aux ontologies légères, les ontologies lourdes sont dites formelles.Ces ontologies intègrent en plus des concepts et des relations, des règles d'inférence et les axiomes.

Dans (Van Heijst et al., 1997), la classification des ontologies se repose sur deux critères : la structure de la conceptualisation et le sujet de la conceptualisation.
Pour le premier critère, ils distinguent trois catégories, à savoir (i) les ontologies terminologiques (lexiques, glossaires…), (ii) les ontologies d'information (schéma

d'une BD) et (iii) les ontologies des modèles de connaissances. Pour le deuxième critère, ils distinguent quatre catégories :

a. les ontologies d'application : contiennent toutes les informations nécessaires pour modéliser les connaissances pour une application particulière.

b. *les ontologies de domaine :* fournissent un ensemble de concepts et de relations décrivant les connaissances d'un domaine spécifique.

c. *les ontologies génériques* (dites aussi de haut niveau) : similaires aux ontologies de domaine, mais les concepts qui y sont définis sont plus génériques et décrivent des connaissances telles que l'état, l'action, l'espace et les composants.

d.. les ontologies de représentation ou méta-ontologies : fournissent des primitives de formalisation pour la représentation des connaissances. Elles sont généralement utilisées pour écrire les ontologies de domaine et les ontologies de haut niveau.

3.2. Exemple de ressources sémantiques :

Dans cette section, nous présentions des exemples de ressources sémantiques12 utilisées dans la recherche d'information. Nous mettons l'accent sur la définition des concepts dans ces ressources.

*3.2 1.*Wordnet[2] :

3.2.1.1. introduction

Des programmes issus du monde de l'Intelligence Artificielle ont également établi des passerelles avec WordNet. Le WordNet est utilisable librement, y compris pour un usage commercial, ce qui en a favorisé une diffusion très large. Plusieurs autres ressources linguistiques ont été constituées (manuellement ou automatiquement) à partir de, en extension à, ou en complément à WordNet. L'ensemble de ces ressources linguistiques constitue un système complet couvrant des aspects lexicaux, syntaxiques et sémantiques. Combinées, ces ressources fournissent un point de départ intéressant pour des développements sémantiques dans le cadre du Web sémantique, tels que la recherche d'information, l'inférence pour la compréhension automatique de textes, la désambiguïsation lexicale ou la résolution d'anaphores.

3.2.1.2. Définition

WordNet est une base de données lexicale développée depuis 1985 par des linguistes du laboratoire des sciences cognitives de l'université de Princeton. C'est un réseau sémantique de la langue anglaise, qui se fonde sur une théorie psychologique du langage. La première version diffusée remonte à juin 1991. Son but est de répertorier, classifier et mettre en relation de diverses manières le contenu sémantique et lexical de la langue anglaise. Des versions de WordNet pour d'autres langues existent, mais la version anglaise est cependant la plus complète à ce jour.
WordNet est distribué sous une licence libre, permettant de l'utiliser commercialement ou à des fins de recherche.

[2] https://wordnet.princeton.edu/wordnet/download

La dernière version distribuée en avril 2013 est la 3.1. Cette version est par ailleurs consultable en ligne.[14]

3.2.1.3. Principe

WordNet est donc un réseau lexical où :

- Les synsets sont les noeuds.
- Les relations sémantiques entre synsets sont les arcs.

3.2.1.4. Les synsets

La composante atomique sur laquelle repose le système entier est le synset (*synonyme)* c'est un groupe de mots interchangeables, dénotant un sens ou un usage particulier. La version 2.0 de WordNet définit ainsi le nom commun anglais' *car'* à l'aide de cinq synsets comme il est montré dans la figure.

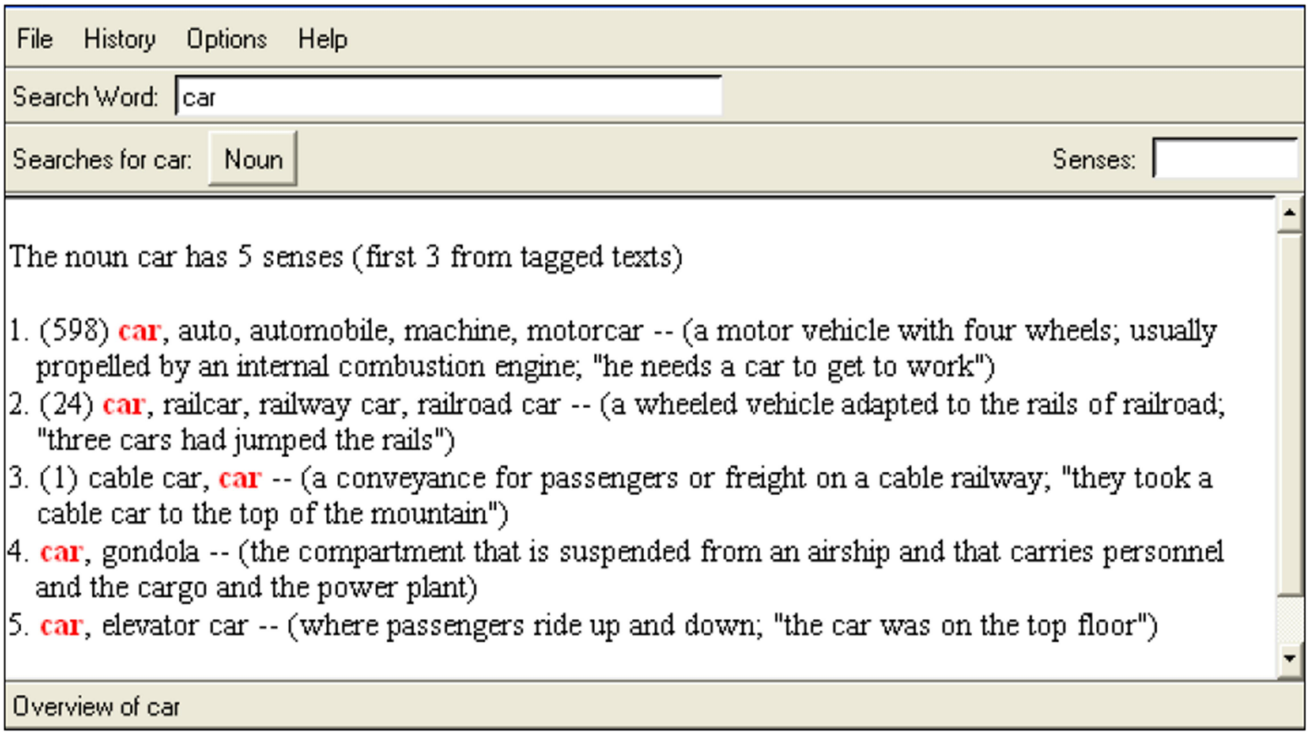


Figure.1 : les différents sens du mot car.

Chaque synset dénote une acception différente du mot *car*, décrite par une courte définition. Une occurrence particulière de ce mot dénotant par exemple le premier sens (le plus courant), dans le contexte d'une phrase ou d'un énoncé, serait ainsi caractérisée par le fait qu'on pourrait remplacer le mot polysémique par l'un ou l'autre des mots du synset sans altérer la signification de l'ensemble.

3.2.1.5 Les relations sémantiques

L'hyperonymie

L'hyperonymie est la relation *sémantique* hiérarchique d'un lexème à un autre selon laquelle l'extension du premier terme, plus général, englobe l'extension du second, plus spécifique. Le premier terme est dit hyperonyme de l'autre, ou super ordonné par rapport à l'autre. C'est le contraire de l'hyponymie. [15]

L'hyponymie

L'hyponymie est une Relation d'inclusion entre deux mots dont l'un est l'hyponyme de l'autre. La relation d'hyponymie est l'expression linguistique de la relation logique d'inclusion d'une classe dans une autre (La Linguistique, Paris, Denoël, 1969, p. 193).

On peut aussi définir les hyponymie comme la relation sémantique d'un lexème à un autre selon laquelle l'extension du premier est incluse dans l'extension du second. Le premier terme est dit hyponyme de l'autre. C'est le contraire de l'hyperonymie.

La méronymie

La méronymie est une relation sémantique entre mots d'une même langue. Des termes liés par méronymie sont des méronymes. La méronymie est une relation partitive hiérarchisée : une relation de partie à tout. Un méronyme X d'un mot Y est un mot dont le signifié désigne une sous-partie du signifié de Y. La relation inverse est l'holonymie. WordNet inclus trois types de méronymie :

- X est un composante de Y.
- X est un élément de Y.
- X est le matériau dont Y est constitué.[17]

L'holonymie

L'Holonymie est une relation sémantique entre mots d'une même langue. Des termes liés par holonymie sont des holonomes. L'holonymie est une relation partitive hiérarchisée : un holonyme A d'un mot B est un mot dont le signifié désigne un ensemble comprenant le signifié de B. La relation inverse est la méronymie.[18]

La Synonymie

La synonymie est un rapport de similarité sémantique entre des mots ou des expressions d'une même langue. La similarité sémantique indique qu'ils ont des significations très semblables. Des termes liés par synonymie sont des synonymes.

Il existe des bases de données de synonymes, présentées comme des dictionnaires, librement téléchargeables. On en trouve aussi vendues ou consultables sous la forme de livres, de logiciels, ou de web, ou des jeux.[19]

L'antonymies

Deux items lexicaux sont en relation d'antonymie si on peut exhiber une symétrie de leurs traits sémantiques par rapport à un axe. La symétrie peut se décliner de différentes manières, selon la nature de son support. On distingue plusieurs supports qui sont autant de type d'antonymie :

- ➔ Les antonymes complémentaires.
- ➔ Les antonymes scalaires.
- ➔ Les antonymes duals. [20]

La Troponymie

La troponymie est une relation sémantique entre deux verbes, l'un décrivant de manière plus précise l'action de l'autre. Le premier verbe est dit troponyme du second. [21]

3.2.1.6 Quelques données statistiques

Dans WordNet, les synsets sont reliés par des relations sémantiques. La relation de synonymie est la relation de base dans WordNet. Elle relie les termes d'un même noeud. Les noeuds (les synsets) sont reliés entre eux par des relations sémantiques telles que, la relation de composition (partie-tout) et la relation hyponymie-hyperonyme (est-un) (Fellbaum, 1998). Dans sa version 3.0 WordNet contient 155287 termes organisés en 117659 synsets. Le Tableau 2.2 présente des statistiques sur le nombre des mots et de concepts dans WordNet 3.0.

Categorie	Mots	Synset	Paire Mot-Sens
Nom	117798	82115	149312
Verbe	11529	13767	25042
Adjectif	21479	18156	30002
Adverbe	4481	3621	5580
Total	155287	117659	206941

Tableau 1. –Statistiques sur les nombres de mots et de synsets dans WordNet 2.1.

3.2.1.7. Les limites du WordNet

Informations manquantes
WordNet ne précise pas l'étymologie, la prononciation, les formes de verbes irréguliers et ne contient que des informations limitées sur l'usage des mots.

Profusion de sens pour un mot donné
La contrepartie de son importante couverture est que WordNet est très précis dans le sens des définitions. On a une granularité très fine des sens. Par exemple, le verbe *to give* (« donner ») n'a pas moins de 44 sens. Une telle profusion ne facilite pas une tâche de désambiguïsation lexicale.

Absence de relations pragmatiques
WordNet ne matérialise pas d'une façon formelle tout le sens contenu dans les définitions des termes. Par exemple, l'information qu'un chat ne rugit pas figure dans la définition, mais ne se retrouve formalisée dans aucune relation. De même, des relations pragmatiques telles que savon / bain (SOAP#1 / BATH#2) sont absentes de WordNet.

3.2.1.8. Quelque domaine d'application

L'utilisation de WordNet en recherche d'informations :

- Pour représenter les documents.
- Pour étendre la requête de l'utilisateur (ajout de synonymes, par exemple pour augmenter le rappel, c'est-à-dire la proportion de documents pertinents rapportés).

➔ Acquisition de relations sémantiques.

➔ Désambiguïsation sémantique.

➔ Pour l'étiquetage sémantique de corpus.

➔ Pour la structuration et catégorisation des documents.

Ressource lexicale permet d'incorporer certaines connaissances lexico-sémantiques au traitement des textes :
□□Pour indexer les documents.

□□Pour réduire les différences de vocabulaire, entre les textes et les questions sur ces textes.

En général WordNet est utilisé :

□□Pour la recherche d'informations.

□□Pour l'extraction d'informations.

□□Pour les systèmes de questions/réponses.

□□Pour enrichir la représentation avec des synonymes, hyperonymes,….

*3.2.2.*UMLS [3]

UMLS (Unified Medical Language System)15 est un thésaurus résultant de la fusion de plusieurs sources terminologiques (UMLS knowledge sources) du domaine médical telles que MESH, SNOMEDCT et RXNORM. UMLS est formé de deux composantes principales:

[3] https://www.nlm.nih.gov/research/umls/licensedcontent/downloads.html

Le méta-thésaurus : il regroupe principalement les concepts (2125396 concepts) et les termes (7581706 termes) associés à ces concepts. Ces termes sont écrits dans une ou plusieurs langues. Ces concepts et termes sont issus de différentes ressources sémantiques. Des variations syntaxiques et lexicales des termes sont parfois données.

Le réseau sémantique : le réseau sémantique définit l'organisation des concepts et les relations entre ces concepts. Dans UMLS, les concepts sont organisés en classes. A chaque concept au moins une classe est associée. Ces classes forment des types sémantiques (135 types sémantiques). Ces types sont reliés entre eux par des relations sémantiques (54 relations).
Un concept de UMLS est identifié par un identificateur unique (CUI : Concept Unique Identifier). Un concept est relié à une ou plusieurs chaînes de caractères (STR), les termes qui dénotent le concept. Les STRs sont liés à une langue (LAT : Language of term) et à un indicateur (ISPREF : Atom status - preferred (Y) or not (N) for this string within this concept) qui indique si le terme est préféré ou non.
Pour chaque concept, la source du concept est mentionnée (SAB : Abbreviated source name).

3.2.3. Gene Ontology[4]

Gene Ontology16(GO) est une initiative de plusieurs banques de données génomiques visant à construire une ontologie générique pour décrire le rôle d'un gène/protéine
(« gene product»). G O e st l a f usion d e trois sous-ontologies :processus biologiques, fonctions moléculaires et composants cellulaires. Dans Gene Ontology , les concepts sont structurés avec quatre types de relations :
subsomption « is-a », méronymie « part-of », synonymie et une relation transversale «*regulates* ».
Un concept de GO est décrit par 5 champs : l'identifiant du concept (**id**), le terme (**name**), la sous-ontologie (**namespace**), définition (**def**), et les relations sémantiques (**isa**, **relationsh)**

3.2.4. YAGO[5]

YAGO17 est une base de connaissances généraliste sémantique, qui est développée au Max Planck Institute. YAGO (Suchanek et al., 2007) vise à construire une ontologie généraliste et extensible. YAGO s'appuie sur les entités et relations extraites à partir du réseau lexical WordNet et l'encyclopédique Wikipedia18. Les entités sont des objets (personnes, villes…) et les relations représentent des faits.
Par exemple les entités « Elvis Presley » et « Grammy Award » sont reliées pare la
relations « hasWonPrize »

Le contenu de YAGO a été extrait automatiquement de Wikipedia et unifié avec la sémantique de Wordnet avec une précision 95%. Tous les objets (villes, personnes, mêmes les URLs) sont représentés comme des entités dans le modèle de YAGO. YAGO n'utilise pas la totalité de la hiérarchie des catégories de Wikipedia mais associe les catégories feuilles à la taxonomie de Wordnet pour établir des relations de type « subClassOf ».

[4] *https://fr.wikipedia.org/wiki/Gene_Ontology*

[5] https://en.wikipedia.org/wiki/YAGO_(database)

3.3. Les concepts

De façon générale, il existe une littérature riche qui traite de la notion de concept selon les communautés (linguistiques, sciences de la cognition, etc.). Dans ce qui suit, nous ferons un rapide tour d'horizon sur les différentes définitions qui lui sont attribuées, avant de passer aux concepts tels qu'ils sont définis et utilisés en RI.

3.3.1 Qu'est ce qu'un concept?

Selon le dictionnaire de l'académie française, *"Le concept regroupe les objets qu'il définit en une même catégorie appelée « classe » "*. De façon général, le terme *concept* est souvent utilisé comme se référant à toute notion, de l'idée au lexème, en passant par l'entité et la catégorie. Selon Medin [Medin, 89], globalement un concept est une idée qui inclut tout ce qui est caractéristiquement associé à elle. Dans la littérature, il existe différents points de vues traitant de la notion de concept selon les communautés (sciences de la cognitions, linguistique). Nous donnons en annexe un résumés sur les principales théories issues de la science de la cognition ainsi que le point de vue linguistique.

3.3.2. Les concepts en RI

En recherche d'information, le processus de représentation ou d'indexation dans la majorité des systèmes actuels, utilise l'indexation classique basée mots clés et représente les documents dans le niveau symbole (chaînes de caractères) si l'on se réfère au triangle sémiologique. Dans ce type d'indexation, le concept est alors réduit au mot clé et le sens des mots n'est pas pris en compte. Dans les systèmes récents supportant un mécanisme d'indexation conceptuelle, les concepts sont le plus souvent prédéfinis et pré ordonnés dans des structures conceptuelles telles les hiérarchies de concepts ou les ontologies. Ces structures sont souvent construites manuellement par des spécialistes du domaine qu'elles couvrent. Un concept, qui correspond à un nœud de cette structure, peut alors différer d'une structure à une autre. Comme exemples de concepts, on peut citer ceux définis dans le thesaurus médical MeSH et l'ontologie générique WordNet.

Conclusion

La complexité du traitement textuelle des documents est évidente. L'indexation nous permet une représentation meilleure du document en reformulant son contenu sous une forme plus adaptée à son exploitation.

Dans ce chapitre, nous avons présenté les principales étapes du processus d'indexation des documents et les ressources sémantiques externes, qui permettent un accès efficace à la représentation des documents.

Chapitre II
DESAMBIGUISATION

CHAPITRE :2 DESAMBIGUISATION

1. Introduction

La désambiguïsation sémantique consiste à déterminer le sens correct des mots d'un texte. Elle peut en effet permettre d'améliorer de nombreuses applications comme l'extraction d'informations des documents. les approches existantes montre qu'elle permet de mieux comprendre certaines propriétés et, nous avons décidé d'implémenter deux algorithmes de désambiguïsation simple et tenté de comprendre leur limites autrement qu'en analysant leur performances brutes en terme de précision.

2. Désambiguïsation lexicale

Une tâche de désambiguïsation lexicale consiste à choisir le sens le plus approprié pour chaque mot d'un texte, de nombreux travaux traitant la désambigüisation existent dans la littérature.

2.1. Indexation Sémantique_Désambigüisation :

Afin d'assurer la recherche dans des conditions acceptables de coût et d'efficacité, l'indexation est l'étape qui consiste à analyser le document lors de l'organisation du fond documentaire afin de produire un ensemble de mots clés.
Dans notre travail, on prend en compte la sémantique des termes d'indexation. Ce type d'indexation passe du niveau des mots au niveau des concepts ou des sens des mots pour décrire le contenu.

Nous présentons dans ce qui suit quelques approches proposées pour l'indexation sémantique.

2.2.1. Le modèle *DocCore [Baaziz,2005]*

Le modèle DocCore comme variante des réseaux sémantiques

La représentation proposé pour le document, appelé *noyau sémantique de document,* est une représentation analogue à celle du réseau sémantique , avec la différence que les arcs ne soient pas étiquetés, mais quantifiés par une valeur réelle dénotant la proximité sémantique entre les deux concepts (nœuds) et le sens de l'arc est bidirectionnelle puisque la relation de proximité sémantique entre les deux valeurs est symétrique. Cette valeur de proximité sémantique est calculée en utilisant des mesures de similarité sémantique. Ceci permet notamment d'estimer l'importance du terme dans le document non seulement par sa fréquence d'apparition, mais également par les valeurs de proximité sémantique qu'il a avec le reste des termes dans le document.

L'objectif du modèle est de représenter le contenu sémantique de documents. L'approche consiste à projeter les documents sur une ontologie linguistique générale, telle que WordNet. Il s'agit d'identifier pour chaque document les *représentants* de concepts de l'ontologie. Ces derniers peuvent être des mots simples ou des groupes de mots. Un critère de cooccurrence (*CF.IDF*) est utilisé pour extraire les concepts importants. Un deuxième critère qui est la similarité sémantique entre concepts, permet de les désambiguïser via le réseau sémantique de l'ontologie. Le résultat de cet appariement entre le document et l'ontologie est un ensemble de concepts avec des liens pondérés entre eux, formant le *noyau sémantique de document* qui représente au mieux le contenu sémantique du document.

2.2.2 Les travaux de Khan L. [Khan et al,2004]

Dans (Khan et al., 2004), Khan propose une indexation conceptuelle en se basant sur une ontologie spécifique au domaine du sport. Son idée se base sur l'utilisation de l'ontologie dans le processus d'indexation afin de désambigüiser les termes représentant d'un document. Khan propose un algorithme de désambiguïsation basé sur deux principes : la co-occurrente et la proximité sémantique. Il utilise les définitions suivantes :

- ➔ le score d'un élément : chaque concept C_i est composé d'une liste complémentaire de synonymes $C_i = \{l_1 , l_2 ,\ldots, l_n \}$. Les mots-clés dans le texte sont appariés avec chaque élément l_j du concept. Le score pour l'élément l_j du concept C_i est le nombre de mots clés l_j s'accordant avec l'annotation du document.

$$Score\_element_{ij} = \frac{nombre\ des\ mots\ clés\ de\ \ l_j\ \ appariés}{nombre\ des\ mots\ clés\ de\ \ l_j}$$

- ➔ le score d'un concept correspondant au plus grand $Score\_element_{ij}$:

$$Score\_concept_i = \max\ Score\_element_{ij}\ \text{où}\ 1 \leq j \leq n$$

- ➔ le score d'une région dans l'ontologie : est égale à la somme des $Score\_concept_i$ sélectionnés.

- ➔ la distance sémantique, SD(Ci , Cj) entre les concepts Ci et Cj correspond au plus court chemin entre les deux concepts dans l'ontologie et es égale à 1 s'ils sont directement reliés.

- ➔ le score de propagation : s'il existe une corrélation entre un concept C_i avec un ensemble de concepts $\{C_{j+1} , C_{j+2} ,\ldots, C_n \}$, le score de propagation S_i est le suivant :

$$S_i = Score\_concept_i + \sum_{k=j}^{k=n} \frac{Score\_concept_k}{SD(C_i , C_j)}$$

2.2.3. Algorithme de désambiguïsation*Left to right [Dinh et al, 2012]*

Cet algorithme de désambiguïsation consiste à sélectionner le sens le plus adéquat d'un concept dans le contexte local du document (Dinh *et al.,* 2010). Cette méthode est basée sur les définitions et notations suivantes (Dinh *et al.,* 2010) :
Cette méthode se base sur les définitions suivantes :

1) **Définition 1 :** Un mot est une chaîne de caractère alphanumérique séparée par un espace.

2) **Définition 2 :** Un terme composé d'un ou plusieurs mots détermine une unitélinguistique dans le vocabulaire de MeSch.

3) **Définition 3 :** Un concept représente une classe sémantique d'un objet et se compose d'un ou plusieurs termes synonymes.

4) **Définition 4 :** Le sens d'un concept est représenté par un nœud, indiqué par le numéro d'arbre dans la poly-hiérarchie. L'ensemble de sens d'un concept *c* est désigné par *syn(c)*.

Cette méthode de désambiguïsation est basée sur les hypothèses suivantes :

1) **H1 :** l'unicité du sens d'un concept dans le document (Gale *et al.,* 1992),

2) **H2 :** la corrélation des sens des concepts voisins : les sens associés à des concepts voisins sur une fenêtre (contexte) sont sémantiquement proches les uns des autres,

3) **H3 :** la priorité du sens est définie selon la précédence des concepts : le concept le plus à gauche détermine le sens global de la suite du discours, ce qui crée une chaîne sémantique du discours à partir du début jusqu'à la fin du document.

De proche en proche le sens du concept dans le document est calculé par la similarité entre celui-ci et son voisin précédent désambiguïsé. En se basant sur l'hypothèse (H1), une fois que le concept est désambiguïsé, son sens est propagé pour toutes ses occurrences dans le document. En considérant la liste de *n* concepts du document,
$Ln = \{c_1, c_2, ..., c_n\}$, nous utilisons la formule suivante pour identifier le sens optimal du concept c_k:

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

Où :
***s_k*:** le sens du concept *ck*,
***syn*(c_k) :** l'ensemble de sens du concept *ck*,
***sim*(s_1, s_2) :** similarité basée sur les hiérarchies de *s*1 and *s*2.

La similarité entre deux sens de deux concepts est calculée en utilisant la similarité de graphes des hiérarchies de concepts associés selon la formule de (Leacock *et al.,*1998):

$$sim(s_1, s_2) = -\log \frac{length(s_1, s_2)}{2 * D}$$

Où :
***length*(*s*1, *s*2)** est le chemin le plus court entre *s*1 and *s*2, and *D* est le niveau le plus profond de la hiérarchie.

2.2.4. Algorithme de désambiguïsation par translation d'une fenêtre de termes [*Harathi,2010*]

Dans cette approche, le processus de désambiguïsation se compose de deux étapes :

2.2.4.1. Identification des concepts

Dans cette étape, les concepts sont détectés dans une unité textuelle d'un texte. Une unité textuelle à l'intérieur d'un document, peut représenter une phrase, un paragraphe.
Le processus d'identification des concepts se décompose en deux principales étapes :

1. L'analyse syntaxique : l'objectif de cette étape est d'analyser le texte afin de fournir des mots étiquetés et lemmatisés. Dans cette étape, on associe à chaque mot sa racine (lemme).

2. Recherche des concepts : cette étape consiste à identifier les concepts à partir des termes. La recherche des concepts nécessite une projection sur WORDNET qui renvoi les synsets associés aux termes.

2.2.4.2. Désambiguïsation des termes

La désambiguïsation consiste à déterminer, parmi les différents concepts dénotés par un terme, celui qui correspond au mieux à ce terme dans son contexte d'énonciation. Ce contexte désigne le contexte d'apparition qui inclut les termes voisins d'un terme à désambiguïser et pour lequel l'unité maximale considérée est la phrase (Yaroswsky,1993).
Formellement on définit un contexte comme un ensemble termes qui apparaissent ensemble dans une phrase d'un texte.

$$contexte = \{t1,\ldots, tk,\ldots tn\}$$

Où $t_k$représente un terme et *k* désigne son ordre d'apparition dans la phrase.

On dénote par $concept_\Omega$ (tk) l'ensemble des concepts dans la ressource sémantique Ω qui sont dénotés par le terme tk.

$$concept_\Omega(tk) = \{ck1,\ldots, cki\}$$

Où $\mathbf{C_k^{\,i}}$ est l'i-ième concept dénoté par le terme tk .

Le problème de la désambiguïsation consiste à sélectionner une seule combinaison des concepts parmi les combinaisons possibles des concepts dans un contexte donné. Une combinaison est définie comme suit :

$$combinaison = \left\{ c_1^{i_1}, \ldots, c_k^{i_k}, \ldots c_n^{i_n} \right\},\ avec\ 1 \leq i_k \leq |\ concept_\Omega(t_k)\ )\ |$$

Où ck_i^kdésigne le i_k-ième concept dénoté par le terme tk.

La désambigüisation est un problème combinatoire. Le nombre des combinaisons possibles est :

$$\prod_1^{k=n} \left|\ concept_\Omega(t_k)\ \right|.$$

Autrement dit, on détermine la combinaison dans laquelle les concepts sont sémantiquement très proches. La proximité sémantique entre les concepts peut être évaluée par l'utilisation des mesures de similarités sémantiques. Ainsi, le concept sélectionné correspond au concept qui maximise la similarité sémantique avec les autres concepts du même contexte.

Soit CB= { c1,…, ck,…cn} une combinaison de concepts, on définit la similarité sémantique entre les concepts de CBcomme la valeur de la mesure de similarité entre tous les concepts de CB.

$$\text{SIM (CB)} = \sum_{i=1}^{i=n} \sum_{j=i+1}^{n} sim_\Omega(c_i,\ c_j)$$

L'identification des concepts consiste à sélectionner la combinaison CB_{max}qui est la valeur maximal de CB entre ses concepts.

Notre méthode consiste à utiliser une fenêtre de taille fixe et à la translater. Par exemple pour désambigüiser les termes du contexte {t1, t2, t3, t4, t5} et en prenant une fenêtre de taille 3, l'identification des concepts se déroule de la façon suivante :

1. identification des concepts {c_1, c_2} à partir des combinaisons de la fenêtre { $concept_\Omega(t_1)$, $concept_\Omega(t_2)$, $concept_\Omega(t_3)$}.

2. identification du concept {c_3} à partir des combinaisons de la fenêtre {c_2, $concept_\Omega(t_3)$, $concept_\Omega(t_4)$}.

3. identification des concepts {c_4, c_5} à partir des combinaisons de la {c_3, $concept_\Omega(t_4)$, $concept_\Omega(t_5)$}.

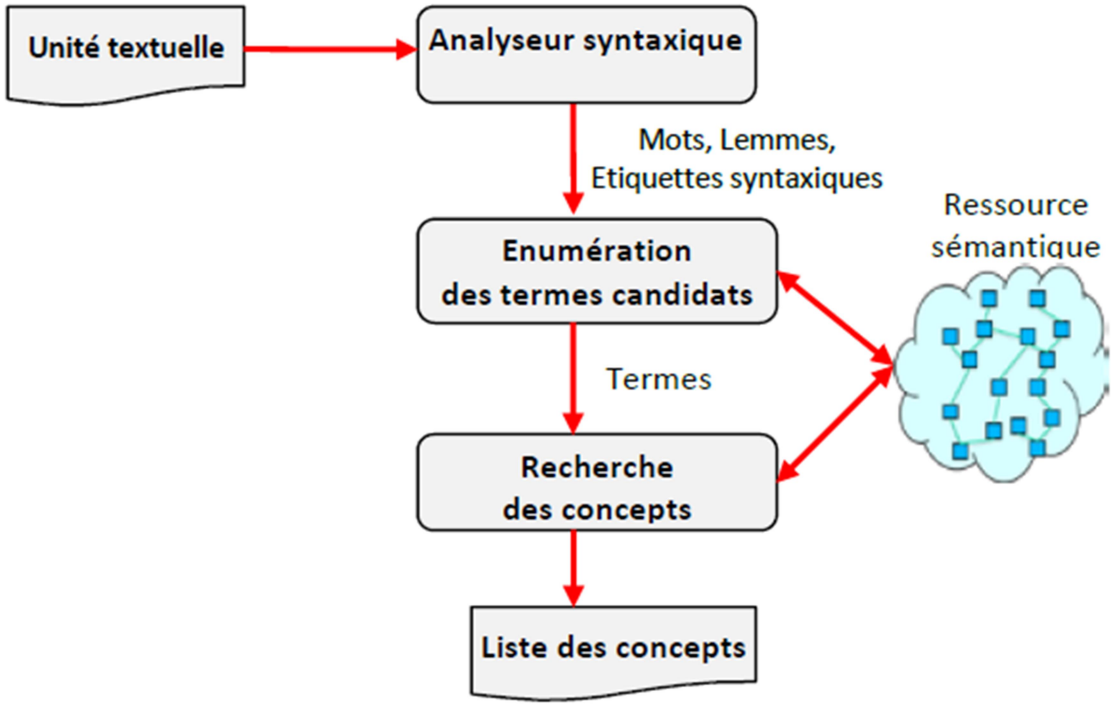


Figure 2 ·Processus d'identification des concepts.

3. Description des Mesures de Similarité

Les mesures de similarité entre concepts sont utiles dans la phase de sélection des meilleurs concepts lors de la construction du Noyau sémantique de Document. Elles servent à sélectionner parmi les différents concepts candidats pour un terme donné, celui qui représente au mieux le sens du terme dans un contexte local (défini par les autres concepts du document). Globalement, dans la littérature, il existe une bonne dizaine de mesures permettant de calculer, dans une hiérarchie de concepts, une valeur de similarité sémantique entre le deux concepts à comparer. Un état de l'art sur ces mesures est fait dans [Budanitsky, 99] et [Patwardhan et al., 03]. Nous avons choisi d'utiliser, pour l'évaluation, la mesure de Lesk [Lesk, 86], de Resnik [Resnik, 95], de Leacock et Chodorow [Leacock et al., 94] et la mesure de Lin [Lin, 98]. Ces mesures utilisent, soit le principe de Lesk (nombre de mots en commun), soit la notion du plus court chemin entre les deux concepts (noeuds) à comparer [Leacock et

al., 94] [Lin, 98], ou encore, la notion de chemin combinée avec la notion de quantité d'information, telle que introduite par Resnik. Le choix de ces mesures a été fait, (1) en sélectionnant les mesures connues dans la littérature comme étant les plus performantes, (2) de façon à varier le type de mesure utilisée. Nous décrivons dans la suite les trois dernière mesures, à savoir, Resnik, Leacock et Chodorow (Lch) et Lin. La mesure de Lesk adaptée a été déjà décrite dans la section 4.2.4. Pour l'implémentation de ces mesures, nous avons utilisé de Jason Rennie [Rennie, 00] et Similarity0.12de Pedersen, les packages QueryData1.382 3 Patwardhan et Michelizzi [Pedersen et al., 03].

3.1. Méthode basée sur le comptage de lien (Edge counting)

Cette méthode considère la position où se trouvent les concepts sur la taxonomie. Elle se base sur l'hypothèse est que plus il y a de liens entre deux concepts et plus ils sont similaires.

3.1.1. La mesure de Resnik

Resnik [Resnik, 99] a introduit la notion de *Contenu d'Information* (Information Content ou *IC*) des concepts (appliquée aux seuls noms) en utilisant le sous-ensemble correspondant à la hiérarchie *est-un (is-a)* ou hyperonymie de WordNet. L'idée principale derrière cette mesure est que deux concepts sont sémantiquement liés ou proches, proportionnellement à la quantité d'information qu'ils partagent. La quantité d'information est déterminée par le contenu d'information du plus spécifique concept (noeud de la hiérarchie) qui subsume les deux concepts à comparer qu'il appelle *lcs* (pour Least Common Subsumer). Elle est définie comme suit :

$$Sim_{\_resnik}(c_1, c_2) = IC(lcs(c_1, c_2)$$

Le contenu d'information (*IC*) d'un concept est estimé en calculant sa fréquence dans un large corpus. Il est défini comme le négatif du log de sa probabilité :

$$IC(concept) = -\log(P(concept)$$

La fréquence d'un concept dans la hiérarchie, inclut la fréquence de tous ces descendants puisque une occurrence ajoutée à un concept est aussi ajoutée aux concepts qui le subsument. Par conséquent, les concepts qui se trouvent dans la partie supérieure de la hiérarchie vont avoir les plus grandes fréquences que ceux qui se trouvent dans le niveau le plus spécifique (en bas de la hiérarchie). Ce qui justifie le moins (-) du log affecté par Resnik pour favoriser les concepts spécifiques qui se trouvent en bas de la hiérarchie.

3.1.2. La Mesure de Leacock et Chodorow

La mesure de Leacock et Chodorow [Leacock et al., 94] est une mesure basée sur le chemin. Elle dépend de la longueur du plus court chemin entre concepts dans une hiérarchie *est-un* (*is-a*). Le plus court chemin est celui qui comprend le plus petit nombre de noeuds intermédiaires. Cette valeur est inversement proportionnelle à la profondeur maximale de l'arbre notée *D* qui représente la taille du plus long chemin de la feuille au noeud racine dans la hiérarchie. Cette mesure est définie comme suit :

$$Sim\_lch(c_1, c_2) = \max[-\log(length(c_1, c_2) / (2.D))]$$

Où length(**C_1 ,C_2)** est le plus court chemin entre deux noeuds et *D* la profondeur maximale dans la taxonomie (égale à 16 dans WordNet 1.7).
Exemple : dans la Figure 4-6 ci-dessus, le plus court chemin entre *credit card* et *medium of exchange* est celui qui passe par le noeud *Credit*. La mesure est alors calculée comme suit :

$$Sim\_lch(credit\ card, medium\ of\ exchange) = -\log(1/(2x16))$$

3.1.3. La mesure de Lin

Selon le *Théorème de Similarité* de Lin [Lin, 98], la similarité entre deux concepts est mesurée par le ratio du contenu d'information nécessaire pour mesurer la "communalité" des deux concepts, sur le montant du contenu d'information nécessaire pour décrire chacun des deux concepts.
La communalité entre deux concepts dépend du contenu d'information (*IC*) de leur plus spécifique subsumer (*lcs*) et du contenu d'information des deux concepts eux même. Cette mesure est proche de celle de Jiang et Conrath même si elles sont développées séparément :

$$Sim\_lin(c_1, c_2) = \frac{2.IC(lcs(c_1, c_2))}{IC(c_1) + IC(c_2)}$$

Ceci peut être vu comme le contenu d'information de l'intersection des deux concepts (multiplié par deux) qui est divisé sur leur somme, ce qui est aussi analogue à la mesure de Dice.

3.1.4. La mesure de Wu-Palmer

(Wu et al., 1994). Dans une ontologie, la similarité est définie par rapport à la distance qui sépare deux concepts dans la hiérarchie et également par leur position par rapport à la racine. La similarité entre c1 et c2 est :

$$sim_{Wu\_Palmer}(c_1, c_2) = \frac{2 * prof(c)}{dist(c_1, c) + dist(c_2, c) + 2 * prof(c)}$$

Où c est le concept le plus spécifique qui subsume les deux concepts c1 et c2 , prof (c) est le nombre d'arcs qui sépare c de la racine et dist(ci, c) est le nombre d'arcs qui séparent ci de c.

3.1.5. La mesure de Zargayouna

(Zargayouna et al., 2004). Cette mesure de similarité est inspirée de celle de (Wu et al., 1994). Le lien père-fils est ainsi privilégié par rapport aux autres liens de voisinage en adaptant la mesure de Wu-Palmer.
L'adaptation de la mesure est faite au travers de la fonction de calcul du degré de spécialisation d'un concept (spec) qui mesure sa distance par rapport à l'anti-racine.

$$sim_{Zargayoun}(c_1, c_2) = \frac{2 * prof(c)}{dist(c_1, c) + dist(c_2, c) + 2 * prof(c) + spec(c_1, c_2)}$$

$$spec(c_1, c_2) = prof_b(c) * dist(c_1, c) * dist(c_2, c)$$

Où $prof_b(c)$ correspond au nombre maximum d'arcs qui séparent le plus petit ancêtre commun du concept « virtuel » représentant l'anti-racine.

3.1.6. La mesure de Hirst-St.Onge.

(Hirst et al., 1998). L'idée de base de cette mesure est que si deux concepts sont reliés entre eux par un chemin très court et qui « ne change pas la direction » alors les deux concepts sont similaires. Le calcul de la similarité est basé sur le poids du chemin le plus court qui mène d'un concept à un autre et le nombre de changements de directions.

$$sim_{Hirst-St.Onge}(c_1, c_2) = T - chemin - k \times D$$

3.2. Méthode basée sur le contenu informatif

La notion de contenu informatif (CI) a été pour la première fois introduite par
Resnik (Resnik, 1995). Le contenu informatif d'un concept traduit la pertinence d'un concept dans le corpus en tenant compte de sa spécificité ou de sa généralité.

La fréquence de concepts dans le corpus est calculée pour retrouver le contenu informatif. Cette fréquence regroupe la fréquence d'apparition du concept lui même ainsi que des concepts qu'il subsume. Resnik (Resnik, 1999) défini le contenu informationnel comme suit :

$$CI(c) = -log(P(c))$$

P(c) est définie comme la probabilité de retrouver un mot du corpus qui soit une instance du concept c:

$$P(c) = \frac{Freq(c)}{N}$$

Où N est la taille totale d'échantillon de texte et Freq(c) est la fréquence d'occurrence des mots dénotant le concept c dans la collection.

3.2.1 La mesure de Resnik (Resnik, 1999).

La similarité entre deux concepts est liée avec l'information qu'ils partagent en commun, indiquée par le plus spécifique concept $psc(c1\ ,\ c2\ )$ qui les subsume. Le concept le plus spécifique est supposé être le concept qui a le contenu informatif le plus grand. La similarité entre deux concepts est proposée comme suit :

$$sim_{ResnikCI}(c_1, c_2) = CI\left(psc(c_1, c_2)\right)$$

3.2.2. La mesure de Lin (Lin , 1998).

La similarité entre deux concepts est mesurée par le ratio du contenu d'information nécessaire pour mesurer la « communalité » des deux concepts, sur le montant du contenu d'information nécessaire pour décrire chacun des deux concepts. La similarité entre deux concepts c1 et c2 est :

$$sim_{Lin}(c_1, c_2) = \frac{2 \times CI(psc(c_1, c_2))}{CI(c_1) + CI(c_2)}$$

3.3. *Méthodes hybrides*

Ces méthodes sont fondées sur un modèle mixte qui combine entre des approches basées sur le comptage des liens en plus du contenu informatif qui est considéré comme facteur de décision.

3.3.1.Jiang et Corath (Jiang et al., 1997).

La similarité est définie comme une distance sémantique qui tient compte aussi des contenus d'informatifs dans la fonction de la similarité. La distance sémantique est calculée comme suit:

$$\text{distance}\,(c_1, c_2) = 2 \times \text{CI}(psc(c_1, c_2)) - (\text{CI}(c_1) + \text{CI}(c_2))$$

$$\text{sim}_{\text{Jiang\_Corath}}(c_1, c_2) = \frac{1}{\text{distance}\,(c_1, c_2)}$$

3.4 Comparaison des mesures de similarité sémantique

La performance des mesures de similarité sémantique dépend de la qualité de la ressource sémantique et du corpus utilisé. Dans le travail de (Ventresque, 2004) , les mesures de Wu-Palmer (Wu et al., 1994), Leacock-Chorodow (Leacock et al., 1998), Hirst-St.Onge (Hirst et al., 1998), (Resnik, 1995), Resnik (Resnik, 1999), et Lin (Lin , 1998), Jiang-Corath (Jiang et al., 1997) ont été expérimentées avec la ressource sémantique Wordnet. L'évaluation de ces mesures est réalisée par corrélation par rapport aux jugements humains en s'inspirant des travaux de (Miller et al., 1991). Le Tableau 3.6 présente la corrélation entre ces mesures et les jugements humains. Les mesures de Jiang-Corath et la méthode de Leacock-Chorodow apparaissent comme étant les meilleures globalement.

Mesure	Corrélation MC
Wu-Palmer	0.74
Leacock -Chorodow	**0.82**
Resnik -EDGE	0.77
Hirst-St.Onge	0.68
Resnik - Contenu informatif	0.77
Lin	0.81
Jiang-Corath	**0.84**

Tableau 3. Corrélation entre les mesures de similarité sémantique et les jugements humains par Miller et Charles (MC) (Miller et al., 1991).

Cette comparaison est limitée par la qualité du corpus sur lequel sont calculés les contenus informatifs (Zargayouna, 2005). En effet, les mesures hybrides nécessitent un corpus qui doit contenir des occurrences des mots avec une répartition permettant de calculer la probabilité de retrouver un mot du corpus qui soit une instance du concept donnée.

Conclusion

La désambigüisation consiste à déterminer quel est le sens le plus approprié pour chaque mot d'un texte dans son emplacement prédéfini.
Dans ce chapitre, nous avons présenté quelques algorithmes de désambigüisation, et nous avons expliqué leur fonctionnement.
Par conséquent, nous avons choisi l'algorithme ***Left to right*** et l'algorithme par ***Translation de fenêtre*** et nous expliquons leurs implémentations dans le chapitre suivant.

CAPITRE III
CONCEPTION ET REALISATION

1. Introduction :

Dans le contexte de notre projet, nous avons choisi d'implémenter deux algorithmes de désambiguïsations comparer leurs résultats et analyser leurs performances. nous avons retenus les algorithmes « *Left to right* » et celui « *par translation de fenêtre* ».

Pour ce faire nous avons utilisé :

- **JAVA** comme langage de programmation.
- **Netbeans** comme environnement de développement.
- **WORDNET 2.1** comme base de données lexicale.
- **RiTalibrary**
- L'API de **JAWS.**
- La Mesure de similarité de **Leacock et Chodorow**

1.1. NetBeans

NetBeansest un environnement de développement intégré (EDI), placé en open source par Sun en juin 2000 sous licence CDDL et GPLv2 (Common Devlopment and Distribution License).en plus de JAVA, NetBeans permet également de supporter différents autres langages, comme Python, C, C++, JavaScript, XML, Ruby, PHP et HTML. Il comprend toutes les caractéristiques d'un IDE moderne (éditeur en couleur, projets multi-langage, refactoring, éditeur graphique d'interfaces et de pages Web).Conçu en java, NetBeans est disponible sous Windows, Linux, Solaris (sur x86 et SPARC), Mac OS X et open VMS.

En 1977, NetBeans nait d'Elfi, un projet d'étudiants dirigé par la Faculté de mathématiques et de physique de 1'Université Charles de Pragues. Plus tard, une société se forme autour du projet et édite des versions commerciales de 1'EDI NetBeans, jusqu'à ce qu'il soit acheté par Sun en 1999.Sun place le projet sous double licence CDDL et GPL v2 en juin de l'année suivante.

Choix de la mesure de similarité

Nous avons pu tester les différentes mesures de similarités qui sont incluses dans la Librairie **WS4J**[1] , notamment **WuPalmer, Lesk, Lin** et **Leacock et Chodorow.** Nous avons constater que les résultats différent d'une mesure a l'autre. Nous avons retenu **Leacock et Chodorow** car c'est la mesure qui a donné de meilleur résultat.

[1] **WS4J :Word Similarity for JAVA**

Description des étapes de notre application

Les deux algorithmes que nous avons implémentés s'appuient tous les deux sur un modèle simple, ils prennent en entrée un texte. en utilisant la librairie **RiTa** le texte sera découpé en un ensemble de phrases, qui seront à leur tour découpé en mots pour enfin en extraire les noms.une fois la liste des noms obtenus, on procède a la lemmatisation de ces derniers grâce a **RiTa Pling.**

L'étape suivante consiste a projeter tous les noms de chaque phrase sur **WordNet** en utilisant la librairie de **JAWS** qui nous permettra d'obtenir la liste des synsets de chaque nom.

La dernière étape qui est la désambigüisation nous permet d'obtenir le meilleur synset pour chaque nom, pour ce faire, on compare les synsets du nom a désambigüiser avec ceux du nom (des noms) voisin a l'aide de la mesure de similarité **Leacock et Chodorow** qui nous renvoi un score.

Les scores obtenus seront a leurs tour comparé entre eux pour garder le score maximal (le taux de similarité le plus élevé).

2. Présentation des deux approches :

2.1. Algorithme « Left To Right»

Entrées : Document D

Sorties : Synset correspondant aux noms de D

1. $L_n \leftarrow ( )$ {initialiser la liste des noms extrait du document}
2. {Découper le texte en phrases}
3. Phrases ← extrairePhrases(D)
4. **Pour** p ∈ Phrases **Faire**
5. {Extraire les noms pour chaque phrase}
6. L_{nom} ← extraireNoms(p) {Liste de noms dans P}
7. $L_n \leftarrow L_n \cup L_{nom}$
8. **Fin Pour**
9. **Pour** Nom $\in L_{nom}$ **Faire**
10. {Extraire les Synset pour chaque nom}
11. $L_c \leftarrow$ extreaireSynset(Nom) {Liste de synset dans P}
12. **Fin Pour**
13. **Pour** p ∈ Phrases **Faire**
14. {Désambiguïser de proche en proche avec propagation de sens}
15. **Pour** K=1 ; K ≤ n ; K++ **Faire**
16. **Si** K=1 **Alors**
17. $S[1, P] \leftarrow \text{argmax} \sum_{\substack{s_1 \in Syn(L_p[1]) \\ s_2 \in Syn(L_p[2])}} Sim(s1,s2)$
18. Propager(S[1, P],L_n)
19. **Si Non**
20. $S[k, P] \leftarrow \text{argmax} \sum_{s_k \in Syn(L_p[k])} Sim(s_{k-1},s_k)$ {Sk-1 précédemment désambiguïsé}
21. Propager(S[k, P], Ln)
22. **Fin si**
23. **Fin Pour**
24. **Fin Pour**

25. Retourner S

2.1.2 Illustration et discussion

Dans cette section nous illustrons par un exemple le processus de désambiguïsation *Left to Right* des noms de la phrase ci dessous.

> "**Wi-Fi** is a **technology** that allows electronic **devices** to connect to a wireless **LAN** (**WLAN**) **network**."

Exemple 1.1 exemple d'une phrase d'un document a désambiguïsé

Nous extrayons quatre noms dans la phrase, qui seront lemmatiser et projeter sur **WordNet** pour obtenir leurs synsets respectifs.

Les résultats sont donnés dans le tableau suivant :

Terme	Nombre de concepts dénotés par le terme dans WordNet	Numéro d'ordre du synset dans WordNet	Glossaire[2]
Wi-Fi	**1**	**1**	a local area network that uses high frequency radio signals to transmit and receive data over distances of a few hundred feet; uses Ethernet protocol
technology	**2**	**1**	the practical application of science to commerce or industry
		2	the discipline dealing with the art or science of applying scientific knowledge to practical problems
Devices	**5**	**1**	An instrumentality invented for a particular purpose

[2]Le glossaire : Il contient une définition du concept avec éventuellement un ou plusieurs exemples du monde réel (Fellbaum, 1998).

		2	Something in an artistic work designed to achieve a particular effect
		3	Any clever maneuver
		4	any ornamental pattern or design (as in embroidery)
		5	an emblematic design (especially in heraldry)
LAN	**1**	**1**	A local computer network for communication between computers…
WLAN	**1**	**1**	A local area network that uses high frequency radio signals to transmit and receive data over distances of a few hundred feet; uses Ethernet protocol
network	**5**	**1**	An interconnected system of things or people.
		2	(Broadcasting) a communication system consisting of a group of broadcasting stations that all transmit the same programs
		3	an open fabric of string or rope or wire woven together at regular intervals
		4	a system of intersecting lines or channels
		5	(electronics) a system of interconnected electronic components or circuits

Dans cet exemple, le programme commence par désambiguïser les deux premiers termes « **Wi-Fi**» et « **technology**» en appelant la méthode : **WordSimilarity** (**Technology,Wi-Fi**) ou se fait le calcul de similarité entre synsets, cette dernière nous retourne le résultat suivant :

$$\mathbf{SIM_{LeacockChodorow}(Technology_{synset1},Wi\text{-}Fi_{synset1})} = 1.05$$

$$\mathbf{SIM_{LeacockChodorow}(Technology_{synset2},Wi\text{-}Fi_{synset1})} = 0.80$$

La valeur de la mesure de similarité retenue sera :

$$\mathbf{SIM_{LeacockChodorow}(Technology_{synset1}\ et\ Wi\text{-}Fi_{synset1})} = 1.05$$

Le synset retenu pour le nom -- **TECHNOLOGY** -- est le N° 1

Son sens dans WordNetest:**(the practical application of science to commerce or industry).**

Pour la désambiguïsation du terme suivant « **DEVICE**» le programme fait appel à la méthode**WordSimilarity** (**Device,Technology, Indice**) (ou**indice**représente le numéro d'ordre du synset retenu du terme précédemment désambiguïser).

$\mathbf{SIM_{LeacockChodorow}}(\mathbf{Device}_{synset1}, \mathbf{Technology}_{synset1}) = 1.39$

$\mathbf{SIM_{LeacockChodorow}}(\mathbf{Device}_{synset2}, \mathbf{Technology}_{synset1}) = 1.61$

$\mathbf{SIM_{LeacockChodorow}}(\mathbf{Device}_{synset3}, \mathbf{Technology}_{synset1}) = 1.29$

$\mathbf{SIM_{LeacockChodorow}}(\mathbf{Device}_{synset4}, \mathbf{Technology}_{synset1}) = 1.29$

$\mathbf{SIM_{LeacockChodorow}}(\mathbf{Device}_{synset5}, \mathbf{Technology}_{synset1}) = 1.20$

La valeur de la mesure de similarité retenue sera

$\mathbf{SIM_{LeacockChodorow}}(\mathbf{Device}_{synset2}, \mathbf{Technology}_{synset1}) = 1.61$

Le synset retenu pour le nom -- **DEVICE** -- est le N° 2

Son sens dans WordNet est: **(something in an artistic work designed to achieve a particular effect).**

De la même maniéré on obtiendra la valeur de similarité maximale pour :

$\mathbf{SIM_{LeacockChodorow}}(\mathbf{Wlan}_{synset1}, \mathbf{Network}_{synset5}) = 2.30$

Le synset retenu pour le nom -- **NETWORK** -- est le N° 5

Son sens dans WordNet est: **((electronics) a system of interconnected electronic components or circuits)**

2.2 Algorithme « Par translation de fenêtres »

Entrées : Document D

Sorties : Synset correspondant aux noms de D

1. $L_n \leftarrow ( )$ {initialiser la liste des noms extrait du document}
2. {Découper le texte en phrases}
3. Phrases ← extrairePhrases(D)
4. **Pour** p ∈ Phrases **Faire**
5. {Extraire les noms pour chaque phrase}
6. L_{nom} ← extraireNoms(p) {Liste de noms dans P}
7. $L_n \leftarrow L_n \cup L_{nom}$
8. **Fin Pour**
9. **Pour** Nom ∈L_{nom} **Faire**
10. {Extraire les Synset pour chaque nom}
11. L_c ← extreaireSynset(Nom) {Liste de synset dans P}
12. **Fin Pour**
13. **Pour** p ∈ Phrases **Faire**
14. {faire translater une fenêtre de taille 3 pour la désambiguïsation des termes}
15. **Pour** K=1 ; K< n ; K++ **Faire**
16. **Si** K=1 **Alors**
17. {Identification des synset pour les deux premiers termes}
18. $S[1, P] \leftarrow \text{argmax} \sum \text{Sim}(\mathbf{S_1}, \mathbf{S_2}, S_3)$
19. $S[2, P] \leftarrow \text{argmax} \sum \text{Sim}(\mathbf{S_1}, \mathbf{S_2}, S_3)$

    $\mathbf{S_1} \in \text{Syn}(L_p[1])$
    $\mathbf{S_2} \in \text{Syn}(L_p[2])$
    $\mathbf{S_3} \in \text{Syn}(L_p[3])$
20. **Fin Si**
21. **Si** K ≥ 3 **Alors**
22. {identification du synset pour K^{eme} nom de la phrase P}
23. $S[k, P] \leftarrow \text{argmax} \sum \text{Sim}(S'_{k-1}, \mathbf{S_k}, S_{k+1})$

    $\mathbf{S_k} \in \text{Syn}(L_p[k])$
24. **Fin Si**
25. **Fin Pour**

    $\mathbf{S'_{k-1}}$: synset du nom précédemment désambiguïser
26. {identification des synsets pour les deux derniers mots de la phrase P}
    K← N {Nombre de noms appartenant à P}
27. $S[k, P] \leftarrow \text{argmax} \sum \text{Sim}(S'_{k-2}, \mathbf{S_{k-1}}, \mathbf{S_k})$
28. $S[(k-1), P] \leftarrow \text{argmax} \sum \text{Sim}(S'_{k-2}, \mathbf{S_{k-1}}, \mathbf{S_k})$

    $\mathbf{S_k} \in \text{Syn}(L_p[k])$
29. **Fin Pour**
30. **Retourner S**

    S'_{k-2} : synset du nom désambiguïsé

2.2.2Illustration et discussion

Pour illustrer cette approche on reprend l'**Exemple 1.1**

1. identification des synsets pour les noms{**Wi-Fi**, **Technology**} à partir des combinaisons de lafenêtre {$\mathbf{Wi\text{-}Fi}_{\mathbf{Synset}}$,$\mathbf{Technology}_{\mathbf{Synset}}$$\mathbf{Device}_{\mathbf{Synset}}$}.On commence par désambiguïser les deux premiers termes.

$\mathbf{SIM}_{\mathbf{LeacockChodorow}}(\mathbf{Technology}_{\text{synset1}}, \mathbf{Wi\text{-}Fi}_{\text{synset1}}) = 1.05$

$\mathbf{SIM}_{\mathbf{LeacockChodorow}}(\mathbf{Technology}_{\text{synset2}}, \mathbf{Wi\text{-}Fi}_{\text{synset1}}) = 0.80$

$\mathbf{SIM}_{\mathbf{LeacockChodorow}}(\mathbf{Technology}_{\text{synset1}}, \mathbf{Device}_{\text{synset1}}) = 1.38$

$\mathbf{SIM}_{\mathbf{LeacockChodorow}}(\mathbf{Technology}_{\text{synset1}}, \mathbf{Device}_{\text{synset2}}) = 1.60$

$\mathbf{SIM}_{\mathbf{LeacockChodorow}}(\mathbf{Technology}_{\text{synset1}}, \mathbf{Device}_{\text{synset3}}) = 1.29$

$\mathbf{SIM}_{\mathbf{LeacockChodorow}}(\mathbf{Technology}_{\text{synset1}}, \mathbf{Device}_{\text{synset4}}) = 1.29$

$\mathbf{SIM}_{\mathbf{LeacockChodorow}}(\mathbf{Technology}_{\text{synset1}}, \mathbf{Device}_{\text{synset5}}) = 1.20$

$\mathbf{SIM}_{\mathbf{LeacockChodorow}}(\mathbf{Technology}_{\text{synset2}}, \mathbf{Device}_{\text{synset1}}) = 1.04$

$\mathbf{SIM}_{\mathbf{LeacockChodorow}}(\mathbf{Technology}_{\text{synset2}}, \mathbf{Device}_{\text{synset2}}) = 1.38$

$\mathbf{SIM}_{\mathbf{LeacockChodorow}}(\mathbf{Technology}_{\text{synset2}}, \mathbf{Device}_{\text{synset3}}) = 1.04$

$\mathbf{SIM}_{\mathbf{LeacockChodorow}}(\mathbf{Technology}_{\text{synset2}}, \mathbf{Device}_{\text{synset4}}) = 0.98$

$\mathbf{SIM}_{\mathbf{LeacockChodorow}}(\mathbf{Technology}_{\text{synset2}}, \mathbf{Device}_{\text{synset5}}) = 0.91$

La valeur de la mesure de similarité retenue sera

$\mathbf{SIM}_{\mathbf{LeacockChodorow}}(\mathbf{Technology}_{\text{synset1}}, \mathbf{Device}_{\text{synset2}}) = 1.60$

Le synset retenu pour le nom « **Technology**» est le N°1

Son sens dans WordNet est: **(the practical application of science to commerce or industry)**

2. identification du synset pour le nom {**Device**} à partir des combinaisons de lafenêtre {$\mathbf{Technology}_{synset1}$,$\mathbf{Device}_{Synset}$,$\mathbf{LAN}_{Synset}$}.

Dans cette étape notre programme récupère le numéro d'ordre du synset retenu pour «Technologie»,après avoir calculé toutes les similarités sémantiques entre les différentes combinaisons possibles, nous renvoi le résultat suivant :

$\mathbf{SIM}_{\mathbf{LeacockChodorow}}(\mathbf{Device}_{synset1}, \mathbf{LAN}_{synset1}) = 1.89$

Son sens dans WordNet est: **an instrumentality invented for a particular purpose.**

3. identification des concepts {**Network**, **WLAN**} à partir des combinaisons de la fenêtre {$\mathbf{LAN}_{synset1}$,$\mathbf{WLAN}_{Synset}$,$\mathbf{Network}_{Synset}$}.

$\mathbf{SIM}_{\mathbf{LeacockChodorow}}(\mathbf{Network}_{synset1}, \mathbf{WLAN}_{synset1}) = 0.98$

$\mathbf{SIM}_{\mathbf{LeacockChodorow}}(\mathbf{Network}_{synset2}, \mathbf{WLAN}_{synset1}) = 0.38$

$\mathbf{SIM}_{\mathbf{LeacockChodorow}}(\mathbf{Network}_{synset3}, \mathbf{WLAN}_{synset1}) = 1.49$

$\mathbf{SIM}_{\mathbf{LeacockChodorow}}(\mathbf{Network}_{synset4}, \mathbf{WLAN}_{synset1}) = 1.89$

$\mathbf{SIM}_{\mathbf{LeacockChodorow}}(\mathbf{Network}_{synset5}, \mathbf{WLAN}_{synset1}) = 2.30$

$\mathbf{SIM}_{\mathbf{LeacockChodorow}}(\mathbf{Network}_{synset1}, \mathbf{LAN}_{synset1}) = 1.04$

$\mathbf{SIM}_{\mathbf{LeacockChodorow}}(\mathbf{Network}_{synset2}, \mathbf{LAN}_{synset1}) = 1.49$

$\mathbf{SIM}_{\mathbf{LeacockChodorow}}(\mathbf{Network}_{synset3}, \mathbf{LAN}_{synset1}) = 1.60$

$\mathbf{SIM}_{\mathbf{LeacockChodorow}}(\mathbf{Network}_{synset4}, \mathbf{LAN}_{synset1}) = 2.07$

$\mathbf{SIM}_{\mathbf{LeacockChodorow}}(\mathbf{Network}_{synset5}, \mathbf{LAN}_{synset1}) = 2.59$

La valeur de la mesure de similarité retenue sera

$\mathbf{SIM}_{\mathbf{LeacockChodorow}}(\mathbf{Network}_{synset5}, \mathbf{LAN}_{synset1}) = 2.59$

Le synset retenu pour le nom « NETWORK » est le N°5

Son sens dans WordNet est: **((electronics) a system of interconnected electronic components or circuits)**

3. évaluation des résultats :

Afin de pouvoir évaluer les algorithmes que nous avons implémentés, il est nécessaire d'utiliser un corpus. Peu de corpus sont étiquetés manuellement par rapport aux sens de WordNet. Nous avons opté pour le corpus SemCor où chaque mot est étiqueté manuellement avec sa catégorie syntaxique (POS : nom, verbe, adjectif, adverbe) ainsi que le numéro du sens correspondant dans WordNet.

3.1. Corpus Brown

Le Brown Corpus est une collection de texte dans le domaine de la linguistique de corpus. Il contient 500 échantillons de texte en langue anglaise, totalisant environ un million de mots, compilées à partir des ouvrages publiés aux États-Unis.

Le Corpus est composé de 500 échantillons, répartis dans 15 genres en proportion approximative du montant publié en 1961 dans chacun de ces genres. Toutes les œuvres échantillonnées ont été publiés en 1961; aussi loin que pouvait être déterminée, ils ont été d'abord publiés alors, et ont été écrits par des locuteurs natifs de l'anglais américain.

Le corpus à l'origine contenait 1,014,312 mots échantillonnés à partir de 15 catégories de texte(Presse, Religion, Fiction, Humour…etc.).

3.2. Corpus SemCor

Le SemCor (Miller, 1995) est un sous-ensemble du corpus anglais Brown (Kucera & Francis, 1979) qui contient 234 000 mots annotés au niveau sémantique grâce au Princeton WordNet. L'annotation porte au total sur 352 textes. Pour 186 d'entre eux, 192 639 mots (soit l'ensemble des noms, verbes, adjectifs et adverbes) sont annotés. Sur les 166 autres, seulement 41 497 verbes sont annotés.Ce corpus permet par exemple un début d'apprentissage automatique pour des tâches de désambiguïsation lexicale.
Le SemCor utilise le format de balisage structuré SGML (Standard Generalized Markup Language) normalisé et publié par l'ISO en 1986, et qui permet d'encoder le contenu d'un texte en attribuant des balises qui délimitent et identifient chacun des éléments du texte. Par exemple dans le SemCor, labalise **<s snum="id">** identifie une phrase, un mot est encodé entre **<wf>** et **</wf>.**

.3.3. Exemple de document Brown/SemCor

Nous donnons un exemple d'un document extrait du corpus Brown et son équivalent dans le corpus SemCor

corpus Brown

```
<J01 0820  7>    The observable characteristics of planetary radio
<J01 0830  3> radiation are the intensity, the polarization, and
<J01 0830 10> the direction of arrival of the waves.
```

corpus SemCor

```
<s snum=24>
<wfcmd=ignore pos=DT>The</wf>
<wfcmd=done pos=JJ lemma=observable wnsn=1
lexsn=5:00:00:noticeable:00>observable</wf>
<wfcmd=done pos=NN lemma=characteristic wnsn=1
lexsn=1:09:00::>characteristics</wf>
<wfcmd=ignore pos=IN>of</wf>
<wfcmd=done pos=JJ lemma=planetary wnsn=1
lexsn=3:01:00::>planetary</wf>
<wfcmd=done pos=NN lemma=radio_radiationwnsn=1
lexsn=1:19:00::>radio_radiation</wf>
<wfcmd=done pos=VB lemma=be wnsn=2 lexsn=2:42:06::>are</wf>
<wfcmd=ignore pos=DT>the</wf>
<wfcmd=done pos=NN lemma=intensity wnsn=1
lexsn=1:07:03::>intensity</wf>
<punc>,</punc>
<wfcmd=ignore pos=DT>the</wf>
<wfcmd=done pos=NN lemma=polarization wnsn=1
lexsn=1:19:00::>polarization</wf>
<punc>,</punc>
<wfcmd=ignore pos=CC>and</wf>
<wfcmd=ignore pos=DT>the</wf>
<wfcmd=done pos=NN lemma=direction wnsn=1
lexsn=1:15:00::>direction</wf>
<wfcmd=ignore pos=IN>of</wf>
<wfcmd=done pos=NN lemma=arrival wnsn=1 lexsn=1:04:00::>arrival</wf>
<wfcmd=ignore pos=IN>of</wf>
<wfcmd=ignore pos=DT>the</wf>
<wfcmd=done pos=NN lemma=wave wnsn=3 lexsn=1:11:00::>waves</wf>
<punc>.</punc>
</s>
```

3.4. Evaluation

Dans les tableaux ci-dessous on présente les résultats des numéros d'ordres des synsets retenus pour chaque noms dans les textes « **J01**[3] » «**J60**[4] » ayant été pris dans la collection du corpus Brown. Nous avons pris en exemple trois phrases du corpus Brown

- La première colonne est le numéro de la phrase dans le texte.
- la deuxième colonne représente les noms de chaque phrase extrait du **SemCor.**
- la troisième colonne renvoi les numéros d'ordre des synsets retenus par la désambigüisation manuelle dans le corpus SemCor.
- La quatrième colonne représente les numéros d'ordre des synsets retenus avec l'algorithme de désambiguïsation 'Left to right'.
- La cinquième colonne représente les numéros d'ordre des synsets retenus avec l'algorithme de désambiguïsation 'par translation de fenêtre'.

.

Text1:J01

	Numéro du synset retenu			
	Terme	SemCor	Algorithme *Left to right*	Algorithme *Translation de fenétres*
Phrase1	Radio_Emission	1	1	1
	Moon	1	1	6
	Planets	1	1	1
	Source	4	4	3
	Information	3	1	1
	Bodies	1	8	8
	Atmospheres	5	6	6

[3] J01 Cornell--H. Mayer - Radio,Emission of the Moon and Planets

[4] J60 Robert A. Futterman - The Future of Our Cities

	Numéro du synset retenu			
	Terme	SemCor	Algorithme *Left to right*	Algorithme *Translation de fenétres*
Phrase2	Results	1	2	2
	Observations	1	3	3
	Radio_Emission	1	1	1
	Moon	1	1	4
	Conductivity	1	1	1
	Surface	1	1	3
	Layer	2	1	2
	Variation	1	5	6
	Infrared_emission	1	1	1
	Eclipses	1	-	-
	Garstung	1	-	-

	Numéro du synset retenu			
	Terme	SemCor	Algorithme *Left to right*	Algorithme *Translation de fenétres*
Phrase3	Measurements	1	1	1
	Limits	1	5	1
	Characteristics	1	2	1
	Surface	1	1	4
	Materials	1	2	3
	Moon	1	1	2

TEXT2 J60

	Numéro du synset retenu			
	Terme	SemCor	Algorithme *Left to right*	Algorithme *Translation de fenétres*
Phrase1	Density	1	2	1
	Core	2	1	5
	Rapid-transit	1	1	1
	Systems	1	1	1
	Prices	2	4	4

	Numéro du synset retenu			
Phrase2	Terme	SemCor	Algorithme *Left to right*	Algorithme *Translation de fenétres*
	Bulding	2	-	-
	Freeways	1	1	1
	Garages	1	1	1

	Numéro du synset retenu			
Phrase3	Terme	SemCor	Algorithme *Left to right*	Algorithme *Translation de fenétres*
	Interchage	1	1	3
	Los_angeles	1	-	-
	Freeways	1	1	1
	Accesses	3	4	4
	Acres	1	1	1
	Land	1	3	3
	One-eighth	1	-	-
	Square_mile	1	1	1
	Area	6	2	4
	Size	1	2	4
	Rockefeller_center	1	-	-
	New York	1	1	1

4.. Estimation et comparaison des résultats

Nous avons utilisé dix documents extrait du corpus BROWN .

Le tableau 1 présente les résultats obtenus sur l'ensemble des textes traités. Ce tableau illustre les taux de réussite des algorithmes de désambigüisations en comparant avec les resultats donnés parle Corpus SemCor. .

Documents	Algorithme *Left to right*	Algorithme *par translation des fenêtres*
Texte J01	**58%**	**59%**
Texte J60	**57%**	**55%**
Texte J59	**48%**	**49%**
Texte J52	**47%**	**47%**
Texte J53	**45%**	**47%**
Texte E01	**41%**	**48%**
Texte J21	**41%**	**44%**
Texte E24	**42%**	**48%**
Texte E29	**54%**	**55%**
Texte F03	**44%**	**47%**

Tableau 1 : performance des deux algorithmes Retenus.

Apres la comparaison avec les résultats du SemCor, nous constatons que les deux algorithmes réalisent des taux de réussites moyens entre 41% et 59%.

En ce qui concerne 'algorithme left to right ,les résultats varient entre 41% et 58% , ce qui est un taux moins bon comparant aux résultats enregistrer avec l'algorithme par translation de fenêtre qui varient entre 44% et 59%.

L'algorithme *Left to Right* utilise la propagation du sens d'un terme dans un texte, ceci peut s'avérer bénéfique pour l'efficacité de désambigüisation dans le cas ou le sens propager est correcte. Cependant, dans le cas contraire, le sens incorrect retenu aura un impact sur la désambiguïsations su reste du document .

Un meme nom peut avoir des sens différents dans le même document.la propagation du sens retenu d'un nom pour tout le document peut générer des erreurs.

La différence de la procédure de désambiguïsation des deux algorithmes se situe sur la taille de la fenêtre du terme a désambiguïser ; l'algorithme *Left to Right* utilise une fenêtre de deux terme (celui a désambiguïser et celui de gauche précédemment désambiguïser),quant a ; l'algorithme *par translation de fenêtres* ;il utilise une fenêtre de » trois termes (le terme a désambiguïser, le prochain qui le succède dans la phrase et celui de gauche précédemment désambiguïser).ce qui procure a l'algorithme par translation de fenêtre un contexte de désambigüisation plus large, ce qui explique la pertinence de ses résultats.

Nous avons enregistré des taux moyens, ceci est dû à plusieurs critères qui influent sur les résultats qui eux dépendent de l'efficacité des différentes étapes du processus de désambigüisation.

Voici quelques critères qui peuvent engendrer des erreurs :

- Concernant la projection sur WordNet, certains noms existant dans SemCor, sont introuvables dans la base donné lexicale de WordNet, prenons l'exemple du nom [**Garstung**] de la phrase2 du texte **J01** ,qui nous retourne aucun résultat lors de sa projection sur WordNet.

- La lemmatisation des noms retenus n'est pas toujours efficace, car elle peut transformer certains noms en des termes qui n'ont pas d'entrer dans le dictionnaire WordNet, par exemple le nom [**Eclipses**] de la phrase2 du texte **J01** , sa lemmatisation avec **RiTa Pling** est [**Eclips**] qui donne un résultat erroné lors de sa projection sur WordNet.

- Dans l'étape Extraction des noms qui a son importance dans le processus de désambiguïsation, la librairie utilisé n'est pas toujours fiable car certains noms sont ignorés, prenons l'exemple de la phrase2 du document **J60** , les noms extrait par SemCor sont :{**Building, Freeways, Garages**}, alors que notre algorithme n'a pu extraire que {**Freeways, Garages**}.

- au cours du processus de désambigüisation, nous avons pu observer quelques cas particulier par rapport au calcul de similarité avec **Leacock et Chodorow** ou la valeur de la mesure de similarité maximale est similaire entre deux ou plusieurs synsets. Dans ce cas, nous choisissons par défaut de prendre le premier synset retrouvé. ce synset ne représente pas toujours le sens correcte.

Exemple

$\mathbf{SIM}_{\mathbf{LeacockChodorow}}(\mathbf{Technology}_{\text{synset1}}, \mathbf{Device}_{\text{synset1}}) = 1.38$

$\mathbf{SIM}_{\mathbf{LeacockChodorow}}(\mathbf{Technology}_{\text{synset1}}, \mathbf{Device}_{\text{synset3}}) = 1.29$

$\mathbf{SIM}_{\mathbf{LeacockChodorow}}(\mathbf{Technology}_{\text{synset1}}, \mathbf{Device}_{\text{synset4}}) = 1.29$

$\mathbf{SIM}_{\mathbf{LeacockChodorow}}(\mathbf{Technology}_{\text{synset1}}, \mathbf{Device}_{\text{synset5}}) = 1.20$

$\mathbf{SIM}_{\mathbf{LeacockChodorow}}(\mathbf{Technology}_{\text{synset2}}, \mathbf{Device}_{\text{synset1}}) = 1.04$

$\mathbf{SIM}_{\mathbf{LeacockChodorow}}(\mathbf{Technology}_{\text{synset2}}, \mathbf{Device}_{\text{synset2}}) = 1.38$

$\mathbf{SIM}_{\mathbf{LeacockChodorow}}(\mathbf{Technology}_{\text{synset2}}, \mathbf{Device}_{\text{synset3}}) = 1.04$

$\mathbf{SIM_{LeacockChodorow}}(\mathbf{Technology}_{\text{synset2}}, \mathbf{Device}_{\text{synset4}}) = 0.98$

$\mathbf{SIM_{LeacockChodorow}}(\mathbf{Technology}_{\text{synset2}}, \mathbf{Device}_{\text{synset5}}) = 0.91$

Au terme de l'expérience de désambiguïsation sémantique des documents textuels, nous avons pu constater que l'efficacité de nos algorithmes dépend de tous les critères précédemment cités, car ils ont un impact directe sur l'enchainement des noms dans une phrase (le fait d'ignorer un nom dans une phrase réduit la pertinence de désambigüisation des autres nom dans la même phrase)

Conclusion :

Nous avons fait l'étude des différentes approches de désambiguïsation et nous avons opté pour l'implémentation des deux algorithmes de désambiguïsation « left to right » et « l'algorithme par translations de fenêtres ».

Les performances des deux algorithmes en comparaison avec le corpus SemCor sont moyennes. Les résultats obtenus sont moyens a cause de certains critères qui peuvent être améliorer.

CONCLUSION GENERALE

Conclusion générale

La recherche sémantique a pour objectif d'améliorer la précision de recherche par la compréhension de l'objectif de recherche et la signification contextuelle des termes tels qu'ils apparaissent dans l'espace de données recherché, afin de générer des résultats plus pertinents.

Nous avons présentés dans ce mémoire, les différentes approches de désambigüisation des documents textuels qui utilisent des techniques différentes en s'intéressant particulièrement à la désambigüisation sémantique afin de comparer les principaux algorithmes des approches proposées.
Le but de la recherche sur la désambigüisation sémantique des textes est donc de trouver un algorithme permettant de trouver le meilleur sens d'un terme dans le texte avec le plus grand taux de réussite possible.

Au terme de ce travail, on a pu constater que les algorithmes implémentés enregistrent des taux de documents correctement désambigüisés moyen pour les textes du Corpus Brown. L'algorithme par *translation de fenêtre* est légèrement meilleur par rapport à celui de *Left to right* qui réalise un taux plus faible par rapport au corpus SemCor, il est donc considéré comme le moins bon.

Suite à notre travail quelques perspectives sont envisagées dont, l'utilisation de corpus plus grand pour les tests, améliorer le processus d'indexation des noms (Pos Tagger), augmenter la taille de la fenêtre pour désambiguïser un terme ambiguë et évaluer l'impact de cette taille sur les résultats du processus de désambiguïsation et améliorer le processus de lemmatisation.
Nous avons observé que la possibilité de filtrer les mots du contexte en retenant les noms appartenant au domaine relatif au document pourrait être bénéfique.

Au terme de ce travail,

Nos remercions avant tout le bon Dieu tout puissant de nous avoir donné patience, courage et volonté pour réussir notre mémoire,

Nous tenons à exprimer notre profonde gratitude à Mlle ***Iltache Samia****, enseignante à l'UMMTO, pour avoir encadré et dirigé notre travail. Nous la remercions pour ses conseils et ses remarques constructives qui nous ont permis d'améliorer la qualité de nos travaux,*

Nos vifs remerciements vont aux membres du jury pour nous avoir fait l'honneur d'examiner et d'évaluer notre travail avec le poids de leurs compétences,

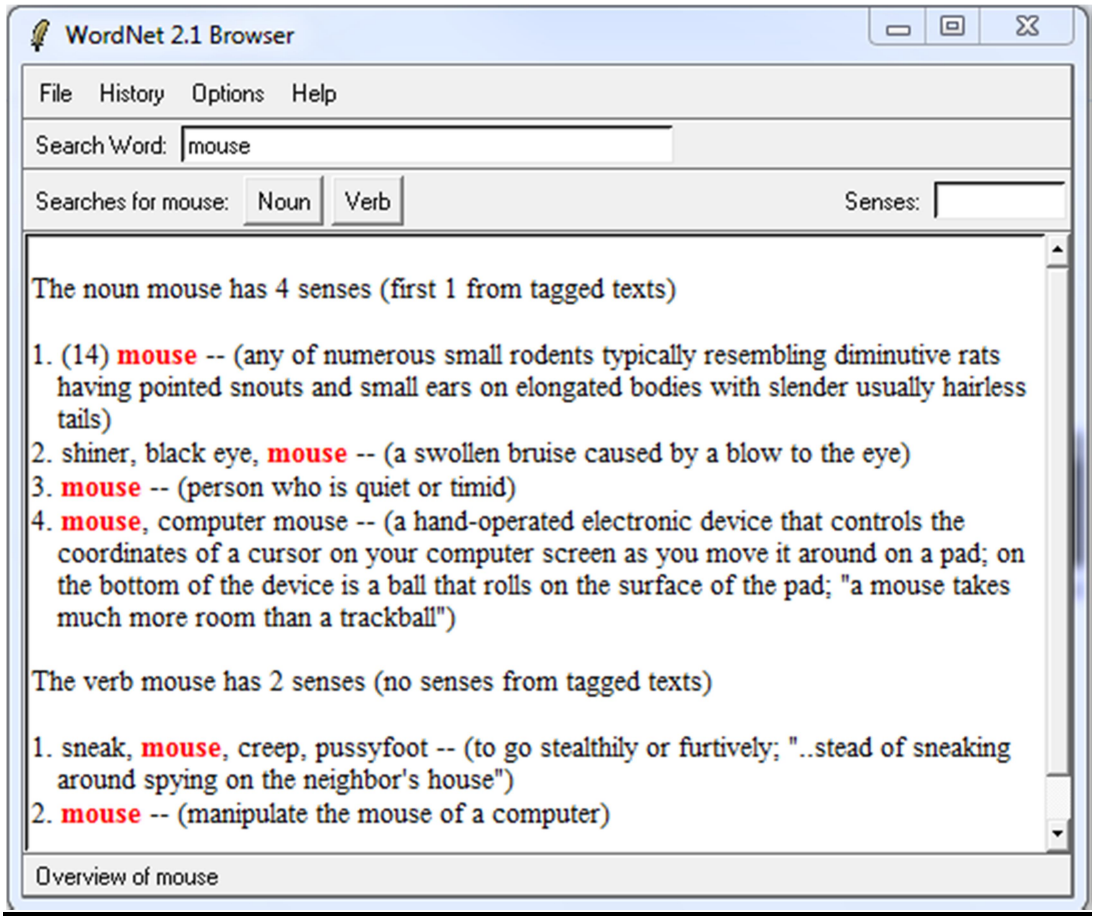
A tous les enseignants qui ont assuré notre formation durant notre parcours universitaire, nous souhaitons qu'ils trouvent dans ce travail l'expression de notre infinie reconnaissance,

Nous souhaitons également exprimer notre gratitude à tous ceux qui de près ou de loin ont participé à l'élaboration du présent travail .

ANNEXE

A. WordNet[1] :

WordNet constitue une base lexicale ou un réseau lexical[14] développée depuis 1985 par des linguistes du laboratoire des sciences cognitives de l'université de Princeton (Fellbaum, 1998).



[1] Des exemples de ressources sémantiques peut être trouvé sur:
http://en.wikipedia.org/wiki/Ontology_%28information_science%29

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

[14] En recherche d'information, WordNet est considéré comme un réseau sémantique et lexical (Zargayouna, 2005) où une ontologie linguistique (Boubekeur, 2008), (Baziz, 2005).

ANNEXE

Le Tableau 3.1 présente un exemple du concept dans WordNet. Le terme Java correspondant à trois concepts.

1.Java -- (an island in Indonesia south of Borneo; one of the world's most densely populated regions)
2. coffee, java -- (a beverage consisting of an infusion of ground coffee beans; "he ordered a cup of coffee")
3. Java -- (a simple platform-independent object-oriented programming language used for writing applets that are downloaded from the World Wide Web by a client and run on the client's machine)

Tableau – Exemple de concept dans WorNet.

Nombre de sens	Nombre de concepts multi mots (2-9 mots)	%
1	56286	89,035%
2	6238	9,867%
3	375	0,593%
>=4	319	0,506%
Total	63218	100%

Tableau Répartition de la polysémie sur les concepts multi mots dans WordNet 2.1

ANNEXE

Projection de document sur une Ontologie

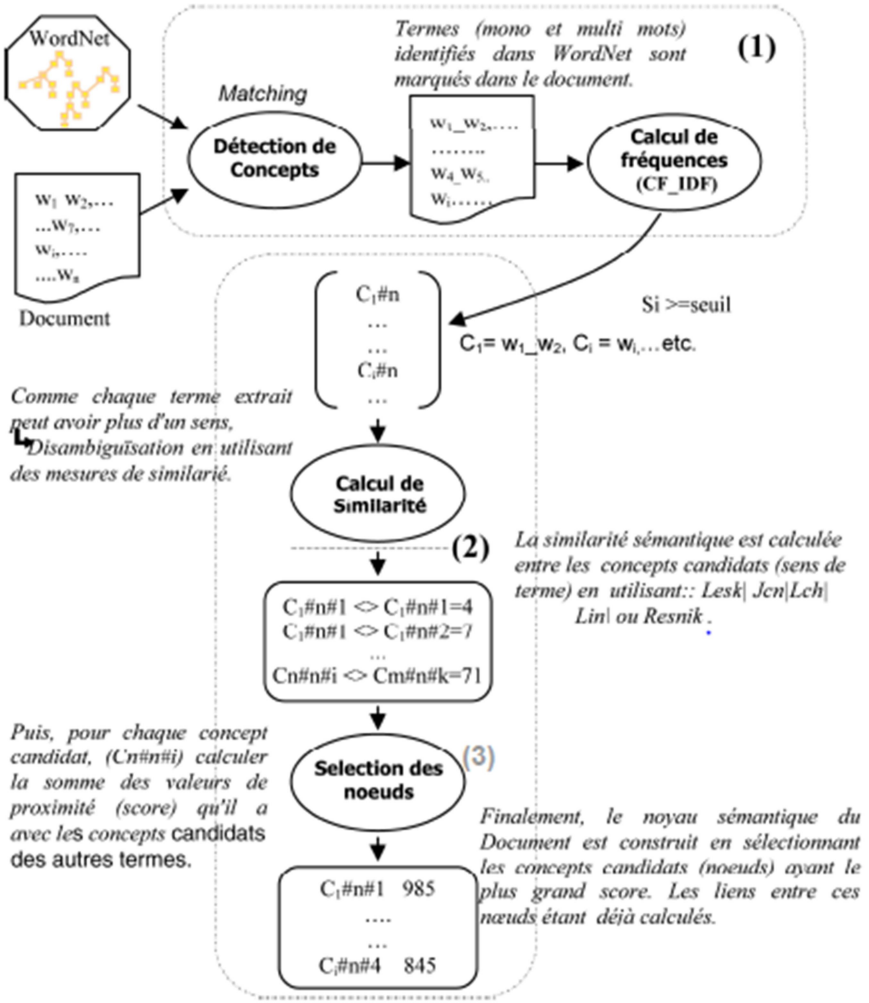


Figure . Schéma synoptique de l'approche de projection d'un document sur l'ontologie.[Baaziz,2005]

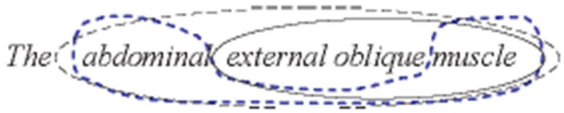


Figure . Exemple de texte avec différents concepts

Références Bibliographiques

[1]. Mustapha Baziz, Indexation conceptuelle guidée par ontologie pour la Recherche d'Information, Thèse de doctorat de l'Université Paul Sabatier, 2005.

[2] Rami HARRATHI, Recherche d'information conceptuelle dans les documents semi-structurés, Institut National des Sciences Appliquées de Lyon, 2010.

[3] Duy Dinh, Lynda Tamine, Recherche d'information sémantique dans les documents biomédicaux : approche basée sur le sens précis des concepts, 2011.

[4] Farah HARRATHI, Vers une approche statistique pour l'indexation sémantique des documents multilingues, *Université de Lyon, 2011.*

[5] Wassila AZZOUG, Contribution a la définition d'une approche d'indexation sémantique de documents textuels, Université M'hamed BOUGARA-Boumerdes, 2013.

[6] Leacock et al. (1998). Leacock C., and Chodorow M. Combining local context and WordNet similarity for word sense identification. In Fellbaum, C., editor, WordNet : an electronic lexical database, volume 11 of Language, Speech and Communication, pages 265–283. The MIT Pr, , Cambridge, Massachusetts.

[7] Khan et al. (2004). Khan L., McLeod D., Hovy E.,. Retrieval effectiveness of an ontology-based model for information selection. TheVLDB Journal (2004)

[8] Ba-Duy DINH (2012), Accès à l'information biomédicale : vers une approche d'indexation et de recherche d'information conceptuelle basée sur la fusion de ressources termino-ontologiques, Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)(2012).

[9]Boubekeur. (2008). Boubekeur-Amirouche F. Contribution à la définition de modèles de recherche d'information flexibles basés sur les CP-Nets, thèse en informatique , Université Toulouse III - Paul Sabatier.

[10] Fellbaum. (1998). Fellbaum M C. WordNet, an Electronic Lexical Database. The MIT Press .

[11] **www.fr.wiképédia.**
[12] **https://netbeans.org/downloads/**
[13] **WordNet : *https://wordnet.princeton.edu/wordnet/download***
[14] **SemCor : *http://web.eecs.umich.edu/~mihalcea/downloads/semcor/semcor2.1.tar.gz***

[15] Zargayouna et al. (2004). Zargayouna, H., Salotti, S. Mesure de similarité dans une ontologie pour l'indexation sémantique de documents XML. Actes de la conférence IC'2004 .