N° Ordre ……../FS/UMBB/2013

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE M'HAMED BOUGARA-BOUMERDES

Faculté des **S**ciences

Mémoire de Magister

Présenté par

M^{elle} Wassila AZZOUG

Filière : Systèmes Informatiques et Ingénierie des logiciels

Option : Spécification de Logiciels et Traitement de l'Information

(École Doctorale)

Contribution à la définition d'une approche d'indexation sémantique de documents textuels

---

Devant le jury :

Mr. Mohamed MEZGHICHE	Professeur (UMBB)	Président
Mme. Farida BOUARAB	MCA (UMMTO)	Examinateur
Mr. Ali BERRICHI	MCB (UMBB)	Examinateur
Mr. Mohand BOUGHANEM	Professeur (IRIT)	Directeur de Mémoire
Mme. Fatiha AMIROUCHE	MCA (UMMTO)	Co-directrice de Mémoire

Année Universitaire : 2012 / 2013.

Remerciements

Je tiens à exprimer mes remerciements et ma très grande reconnaissance à Monsieur **Mohand BOUGHANEM**, Professeur à l'université Paul Sabatier de Toulouse, pour m'avoir honorée en acceptant d'être mon encadreur tout au long des années de préparation de ce mémoire. Je le remercie encore pour m'avoir donné une merveilleuse chance de découvrir le monde de la recherche d'information.

Mes plus vifs remerciements vont également à Madame **Fatiha AMIROUCHE**, Maître de conférences à l'université Mouloud Mammeri de Tizi-Ouzou, pour m'avoir accueillie au sein de son laboratoire où elle m'a assuré un cadre favorable de travail. Je la remercie pour son écoute très attentive, ses remarques et conseils pertinents et sa rigueur scientifique qui m'ont été d'une aide précieuse pour mener à bien ce travail de recherche. Je la remercie encore pour sa disponibilité, son soutien durant mes périodes de doute et pour ses encouragements répétés.

J'adresse mes plus grands remerciements à ceux qui ont accepté de juger ce travail avec le poids de leurs compétences, Monsieur **Mohamed MEZGHICHE**, professeur à l'université M'hamed Bougara de Boumerdès, pour l'honneur qu'il me fait en acceptant de présider le jury de ma soutenance de magister, Madame **Farida BOUARAB**, Maître de conférences à l'université Mouloud Mammeri de Tizi-Ouzou, et Monsieur **Ali BERRICHI**, Maître de conférences à l'université M'hamed Bougara de Boumerdès, pour avoir bien voulu examiner ce travail.

Mes remerciements sincères pleins de reconnaissance vont également à ma famille et à tous mes amis, particulièrement **Sabrina CHARDIOUI**, **Saïda ISHAK BOUSHAKI**, **Amel Garmia BERRAH et Amel KERFI**, qui m'ont soutenue et encouragée depuis le début de ce travail.

A mes chers parents qui ont été à mes cotés dans tous les moments difficiles, par leur amour et leur encouragement, que dieu vous protège pour nous.

A mes frères
Lamia et Mohammed Amine.

A la mémoire de mes grands parents qui resteront à jamais dans mes pensées et au fond de mon cœur.

Je leur dédie ce mémoire.

Résumé

Les systèmes de recherche d'information classiques reposent sur l'indexation par les mots-clés pour représenter le contenu des documents et requêtes. Dans de tels systèmes, les documents sont sélectionnés par un processus de recherche à partir du nombre de mots-clés qu'ils partagent avec la requête. Ce processus, basé sur l'appariement lexical, peut réduire la précision des résultats de la recherche si les sens des mots communs dans la requête et les documents sont différents. L'indexation sémantique tente de pallier à ce problème en offrant une représentation par les sens des mots. Le but étant de retrouver des documents sémantiquement pertinents à une requête utilisateur.

Dans ce présent travail, nous proposons une approche d'indexation sémantique qui s'appuie sur les sens des mots, ou concepts, dans la représentation des documents et requêtes. Ces concepts sont identifiés par des techniques de désambiguïsation des sens des mots en utilisant conjointement les ressources linguistiques WordNet et son extension aux domaines WordNetDomains comme sources d'évidence. Nous proposons en outre, la pondération des termes des index sémantiques par une nouvelle définition de la centralité d'un concept. Les concepts pondérés sont ensuite représentés dans un modèle de recherche qui repose sur une évaluation sémantique de la pertinence d'un document pour une requête donnée. L'évaluation expérimentale de notre modèle de RI sémantique proposé a montré des résultats très satisfaisants.

Mots-Clés : Recherche d'Information (RI), indexation sémantique, concept, pondération des concepts, centralité, désambiguïsation des sens des mots, WordNet, WordNetDomains.

Abstract

Traditional information retrieval systems are based on indexing by keywords in order to represent the content of the documents and queries. In such systems, the documents are selected by a process of research from the number of keywords shared with the query. This process, based on lexical matching, can reduce research results accuracy if the meaning of common words in the query and the documents are different. The semantic indexing tries to solve this problem by providing a representation by words meanings. The goal is to find the semantically relevant documents to a user query.

In the present work, we propose a semantic indexing approach based on the meanings of words, or concepts, in the representation of documents and queries. These concepts are identified by the Word Sense Disambiguation (WSD) techniques, using in jointly linguistic resources WordNet and its extension in domains WordNetDomains as obvious sources. We also propose the weighting of the semantic index terms using a new definition of the centrality of the concept. The weighted concepts are then represented in a research model based on a semantic evaluation of the document relevance for a given query. The experimental evaluation of our proposed semantic model IR has shown very satisfactory results.

Keywords: Information Retrieval (IR), semantic indexing, concept, concepts weighting, centrality, Word Sense Disambiguation (WSD), WordNet, WordNetDomains.

ملخص

نظم التقليدية للبحث عن المعلومات تعتمد على الفهرسة عن طريق الكلمات الرئيسية لتمثيل محتوى الوثائق والطلبات. في مثل هذه الأنظمة، يتم اختيار الوثائق في عملية البحث من خلال عدد من الكلمات الرئيسية التي تشترك مع الطلبات.هذه العملية، التي تستند على المطابقة المعجمية، يمكنها أن تقلل من دقة نتائج البحث إذا كان معنى الكلمات المشتركة في الطلب والوثائق مختلفا. إن الفهرسة الدلالية تحاول تجاوز المشكلة عن طريق تمثيل الوثائق بمعاني الكلمات. الهدف منها هو العثور على الوثائق ذات صلة بالمعنى الدلالي لألفاظ طلب المستخدم.

في هذا العمل، نقترح منهجية فهرسة الدلالية التي تعتمد على معاني الكلمات، أو المفاهيم، في تمثيل الوثائق والطلبات. يتم تحديد هذه المفاهيم من خلال تقنيات توضيح معاني الكلمات، وذلك بالاعتماد على الموارد اللغوية ووردنات وامتدده ووردنات دومنس كمصادر أدلة. كما نقترح أيضا، ترجيح مصطلحات الفهرس الدلالي بتعريف جديد لمركزية المفهوم. يتم تمثيل المفاهيم المرجحة في نموذج بحث يستند على التقييم الدلالي لأهمية وثيقة ما لطلب معين. أظهر التقييم التجريبي لنموذجنا الدلالي للبحث عن المعلومات المقترح نتائج جد مرضية.

الكلمات الأساسية : البحث عن المعلومات، الفهرسة الدلالية، مفهوم، ترجيح مفهوم، توضيح معاني الكلمات، ووردنات، ووردنات دومنس.

Table des matières

Introduction générale 1
Contexte et problématique 1
Contribution 2
Publications dans le cadre de ce mémoire 3
Organisation du mémoire 3
PARTIE 1 : De la RI classique à la RI sémantique 5
Chapitre 1. Recherche d'information 6
1.1 Introduction 7
1.2 Concepts de base de la RI 7
1.3 Processus de recherche d'information 9
1.3.1 Le processus d'indexation 10
1.3.2 Les modèles de recherche d'information 14
1.3.3 Reformulation de requêtes 21
1.4 Evaluation des SRI 23
1.4.1 Collections de test 24
1.4.2 Protocole d'évaluation d'un SRI 24
1.4.3 Métriques d'évaluation 25
1.5 Conclusion 28
Chapitre 2. Indexation sémantique 29
2.1 Introduction 30
2.2 Problématique 30
2.3 L'indexation sémantique 32
2.3.1 Les ressources linguistiques 33
2.3.2 Les approches de désambiguïsation des sens des mots (WSD) 34
2.3.3 Les approches d'indexation sémantique basée sur la désambiguïsation 46
2.4 Conclusion 54
PARTIE 2 : Contributions 55
Chapitre 3. Approche de RI sémantique 56
3.1 Introduction 57
3.2 Motivations 57
3.3 Approche d'indexation sémantique de documents textuels 59
3.3.1 Préliminaires : définitions et notations 60

3.3.2 Aperçu général de l'approche 60
3.3.3 Description détaillée de l'approche 62
3.3.4 Illustration et discussion 71
3.4 Appariement Document-Requête 82
3.5 Conclusion 83
Chapitre 4. Evaluation expérimentale 85
4.1 Introduction 86
4.2 Environnement technologique 86
4.3 Protocole d'évaluation 87
4.3.1 La collection TIME 89
4.3.2 La collection Muchmore 90
4.4. Evaluation avec la collection TIME 92
4.4.1 Evaluation de l'approche d'indexation sémantique dans TIME 92
4.4.2 Evaluation des approches de pondération des concepts dans TIME 112
4.4.3 Evaluation de la mesure sémantique du score d'appariement documents-requête dans TIME 126
4.5 Evaluation avec la collection médicale Muchmore 130
4.5.1 Evaluation de l'approche d'indexation sémantique dans Muchmore 130
4.5.2 Evaluation des approches de pondération des concepts dans Muchmore 131
4.6 Conclusion 134
Conclusion et perspectives 136
Références Bibliographiques 140
Annexe……..149

Liste des figures

Figure 1.1 : SRI en réponse à une requête utilisateur..... 7

Figure 1.2 : Processus en U de la recherche d'information. 9

Figure 1.3 : Distribution des documents dans une collection face à une requête..... 25

Figure 1.4 : Forme général de la courbe Précision-Rappel..... 26

Figure 2.1 : Définitions des mots *pen* et *page* dans le dictionnaire informatisé *Collins English Dictionnary* (*CED*)..... 35

Figure 2.2 : Extrait de l'arbre de listes décision hiérarchiques construite pour le mot *promise* dans le corpus annoté SENSEVAL [Yarowsky et al., 00]..... 45

Figure 2.3 : Exemple de voisinage du premier synset *house* dans WordNet..... 48

Figure 3.1 : Processus d'indexation sémantique..... 62

Figure 3.2 : Extrait d'un document de Muchmore avec ses différents termes descriptifs. ... 71

Figure 3.3: Sens et domaines associés aux différents termes de l'exemple. 74

Figure 3.4 : Résultats de la désambiguïsation du domaine d'usage de chaque terme de $\xi_{Simples}$ dans son contexte local. 75

Figure 3.5 : Résultats de la désambiguïsation contextuelle locale de l'exemple de la figure 3.2, basée sur la similarité de Resnik [Resnik, 99]. 76

Figure 3.6 : Résultats de la désambiguïsation contextuelle locale de l'exemple de la figure 3.2 basée sur la similarité de Lesk [Lesk, 86]. 77

Figure 3.7 : Résultats de la désambiguïsation contextuelle globale de l'exemple, basée sur la mesure de Resnik [Resnik, 99]. 78

Figure 3.8 : Notre désambiguïsation locale *vs* désambiguïsation [Baziz et al., 05a]..... 81

Figure 3.9: Notre désambiguïsation locale *vs* désambiguïsation de [Kolte et al., 09]. 82

Figure 4.1: Indexation classique basée mots clés *vs* Indexation basée {mots simples + collocations}. 93

Figure 4.2: Indexation classique basée mots clés *vs* Indexation sémantique basée (mots simples+collocations+mots simples des collocations). 94

Figure 4.3 : Indexation classique basée mots clés simples *vs* Indexation sémantique basée concepts pondérés par *tf*idf*. 96

Figure 4.4 : Indexation classique basée mots clés simples *vs* Indexation sémantique basée concepts pondérés par *Okapi-BM25*. 96

Figure 4.5 : Impact des concepts noms (issus de la désambiguïsation locale) sur les résultats de la recherche. 98

Figure 4.6 : Impact des concepts-noms (issus de la désambiguïsation globale) sur les résultats de la recherche. 99

Figure 4.7 : Impact des concepts noms (issus de la désambiguïsation mixte) sur les résultats de la recherche. 100

Figure 4.8 : Résultats des Comparaisons entre nos index sémantiques basés concepts-noms (identifiés par nos différentes techniques de désambiguïsation). 100

Figure 4.9 : Apport de l'indexation par les concepts (identifiés par notre approche de désambiguïsation locale) combinés à des mots clés simples. 102

Figure 4.10 : Apport de l'indexation par les concepts (identifiés par notre approche de désambiguïsation globale) combinés à des mots clés simples. 103

Figure 4.11 : Apport de l'indexation par les concepts (identifiés par notre approche de désambiguïsation mixte) combinés à des mots clés simples. 104

Figure 4.12: Résultats des comparaisons entre les index sémantiques basés mots clés simples combinés à des concepts identifiés par nos différentes approches de désambiguïsation (locale, globale ou mixte). 105

Figure 4.13: Impact des collocations sur notre approche d'indexation sémantique basée sur les concepts issus de la désambiguïsation globale. 106

Figure 4.14: Impact des collocations sur notre approche d'indexation sémantique basée sur les concepts issus de la désambiguïsation mixte. 107

Figure 4.15: Apport de notre approche de désambiguïsation des domaines des mots dans la technique de désambiguïsation contextuelle globale. 109

Figure 4.16: Apport de notre approche de désambiguïsation des domaines des mots dans la technique de désambiguïsation contextuelle mixte..... 109

Figure 4.17: Résultats des comparaisons entre notre approche de désambiguïsation des domaines et l'approche de désambiguïsation des domaines de Kolte [Kolte, 09] dans l'indexation sémantique à base de concepts-sens issus de la désambiguïsation globale. 111

Figure 4.18: Résultats des comparaisons entre notre approche de désambiguïsation des domaines et l'approche de désambiguïsation des domaines de [Kolte, 09] dans l'indexation sémantique à base de concepts-sens issus de la désambiguïsation mixte. 112

Figure 4.19: Apport de notre pondération sémantique des concepts (issus de la désambiguïsation locale) avec le schéma *Ct-Ict*. 114

Figure 4.20: Apport de notre pondération sémantique des concepts (issus de la désambiguïsation globale) avec le schéma *Ct-Ict*. 116

Figure 4.21: Apport de notre pondération sémantique des concepts (issus de la désambiguïsation mixte) avec le schéma *Ct-Ict*. 117

Figure 4.22: Résultats des comparaisons entre les index sémantiques à base de concepts, issus des différentes méthodes de désambiguïsation (locale, globale, mixte), pondérés par le schéma *Ct-Ict*. 118

Figure 4.23: Apport de notre pondération sémantique des concepts (issus de la désambiguïsation locale) avec le schéma *Tidf*. 120

Figure 4.24: Apport de notre pondération sémantique des concepts (issus de la désambiguïsation globale) avec le schéma *Tidf*. 120

Figure 4.25: Apport de notre pondération sémantique des concepts (issus de la désambiguïsation mixte) avec le schéma *Tidf*. 121

Figure 4.26: Résultats des comparaisons entre les index sémantiques à base de concepts, issus des différentes méthodes de désambiguïsation (locale, globale, mixte), pondérés par le schéma *Tidf* 123

Figure 4.27: Résultats des comparaisons entre l'approche de pondération sémantique *Ct-Ict* et l'approche de pondération sémantique *Tidf*. 125

Figure 4.28: Apport de la mesure sémantique de l'évaluation des requêtes..... 127

Figure 4.29: Résultats des comparaisons de notre modèle RI basé sur la pondération *Ct-Ict* d'une part et sur la pondération *Tidf* d'autre part. 128

Figure 4.30 : Vue globale sur l'ensemble des résultats des approches évaluées..... 129

Figure 4.31: Impact de l'indexation sémantique de la collection Muchmore, basée sur les concepts de la désambiguïsation globale. 131

Figure 4.32: Apport de notre pondération sémantique des concepts, issus de la désambiguïsation globale, avec le schéma *Ct-Ict*. 132

Figure 4.33: Apport de notre pondération sémantique des concepts, issus de la désambiguïsation globale, avec le schéma *Tidf*. 133

Figure A.1 : Extrait de la hiérarchie *is-a* des noms dans WordNet correspondant au synset *dog*. 150

Figure A.2: Les différents domaines du thésaurus MeSh. 153

Figure A.3: Le descripteur *Pain* (*C10.597.617*) dans l'arborescence du domaine *Diseases* du thésaurus MeSh. 154

Liste des tableaux

Tableau 3.1 : Algorithme de détection des termes descriptifs (document/requête). .. 63

Tableau 4.1 : Nombre de termes identifiés dans TIME couverts par Wordnet. 90

Tableau 4.2 : Nombre de termes identifiés dans Muchmore couverts par WordNet. 92

Tableau 4.3 : Résultats d'évaluation de l'index sémantique *Sem_L_Ct-Ict* pour les différentes valeurs données à α. 113

Tableau 4.4 : Résultats d'évaluation de l'index sémantique *Sem_G_Ct-Ict* pour les différentes valeurs données à α. 115

Tableau 4.5 : Résultats d'évaluation de l'index sémantique *Sem_M_Ct-Ict* pour les différentes valeurs données à α. 117

Tableau 4.6 : Résultats d'évaluation de l'index sémantique *Sem_G_Ct-Ict* pour les différentes valeurs données à α. 132

Tableau A.1 : Les synsets (concepts) de WordNet correspondants au mot *dog*. 150

Tableau A.2 : Extrait de la hiérarchie généralisation/spécialisation de WordNetDomains. 151

Tableau A.3 : Domaines de WordNetDomains associés aux synsets du mot *bank*.152

Introduction générale

Contexte et problématique

La Recherche d'Information (RI) s'intéresse principalement à sélectionner à partir d'un ensemble de documents existants, ceux qui sont pertinents à une requête utilisateur. Afin d'y parvenir, l'une des tâches principales d'un Système de Recherche d'Information (SRI) est l'indexation. L'indexation consiste à construire des représentations simplifiées décrivant le contenu informationnel des documents et requêtes en vue de faciliter la recherche. Ces représentations sont ensuite interprétées par un modèle de recherche dans un formalisme unifié, puis comparées dans le but d'évaluer les degrés de pertinence des documents pour les requêtes.

Dans les SRI classiques, les documents et les requêtes sont représentés (ou indexés) par des mots-clés, manuellement ou automatiquement extraits à partir de leurs textes. Dans de tels systèmes, l'appariement (ou mise en correspondance) document-requête est *lexical* basé sur la présence ou l'absence des mots de la requête dans le document. Un document est alors considéré d'autant plus pertinent pour la requête qu'il a de mots clés en commun avec cette requête. Or, les mots de la langue sont par nature ambigus. Un même mot utilisé dans le document et la requête peut définir des sens différents (cas de polysémie et d'homonymie), et plusieurs mots lexicalement différents utilisés dans le document et la requête peuvent refléter un même sens (cas de synonymie). De ce fait, des documents pourtant non pertinents, contenant des mots de la requête, sont retrouvés, tandis que des documents sémantiquement pertinents, ne contenant aucun mot de la requête, ne sont pas retrouvés. Pour pallier les problèmes de l'indexation basée mots-clés, l'indexation sémantique s'appuie sur la représentation des documents et requêtes par des concepts (ou sens des mots). Ces concepts, sont extraits, à partir du contenu des documents et requêtes, par des techniques de *mapping* sur des ressources linguistiques (dictionnaires, thésaurus, …), ou identifiés par des méthodes de désambiguïsation des sens des mots (*WSD -Word Sense Disambiguation-*) permettant de retrouver le sens adéquat d'un mot ambigu dans son contexte d'utilisation dans le document ou la requête. L'indexation sémantique permet, à l'issue de la recherche, de retrouver des documents sémantiquement pertinents à une requête utilisateur, bien que ne partageant pas de mots en commun avec cette dernière. La qualité de l'indexation sémantique dépend de la précision des techniques de mapping et de WSD utilisées pour sélectionner les concepts représentatifs des documents et requêtes. La qualité d'une recherche d'information sémantique dépend outre de la qualité de l'indexation sémantique, de la qualité de la fonction d'appariement utilisée pour comparer les représentations sémantiques des documents et requêtes et calculer le degré de correspondance entre leurs représentations respectives.

Notre travail s'inscrit dans le contexte de la RI sémantique. En particulier, nous nous intéressons à (1) l'étude des approches d'indexation sémantique existantes et à la définition de nouvelles approches, et (2) à l'étude des approches existantes d'évaluation de la pertinence document-requête (qui sont principalement héritées de l'appariement lexical) et à la proposition de nouvelles approches orientées sémantique.

Contribution

Notre contribution consiste en la proposition d'un nouveau modèle de RI sémantique, qui s'appuie sur les sens des mots dans la recherche de documents textuels pour une requête utilisateur. Ce modèle repose sur deux approches innovantes :

1. la première est *une approche d'indexation sémantique* [Azzoug et al., 11], qui a pour objectif d'améliorer les représentations des documents et requêtes en se basant sur les sens des mots qu'ils contiennent. Cette approche est fondée sur trois étapes :
   - la première étape a pour objet d'extraire à partir des textes de documents (ou requêtes) leurs termes descriptifs (mots simples ou collocations de mots). Pour cette étape, nous avons proposé dans [Azzoug et al., 12] une approche d'identification des concepts (collocations) par mapping du texte sur la ressource terminologique WordNet [Miller, 95].
   - la seconde étape permet de retrouver les sens corrects des mots ambigus par une désambiguïsation des sens des mots (*WSD -Word Sense Disambiguation-*). Pour cette étape, nous avons proposé dans [Azzoug et al., 13c] différentes techniques linguistiques de désambiguïsation sémantique contextuelle se basant sur les ressources linguistiques WordNet et son extension aux domaines WordNetDomains [Magnini et al., 00].
   - la troisième étape consiste à pondérer chaque concept ou sens par un poids traduisant son degré d'importance dans le texte où il apparaît. Pour cette étape, nous avons proposé deux schémas de pondération sémantique basés sur la notion de centralité d'un concept [Azzoug et al., 13b] que nous avons définie.
2. la seconde est *une approche d'évaluation sémantique des requêtes*, qui a pour objectif de calculer le score de pertinence document-requête en se basant sur les proximités sémantiques entre leurs représentations conceptuelles (ie. à base de concepts) respectives [Azzoug et al., 13a].

Ces approches ont été implémentées et testées sur deux collections de test, selon le protocole d'évaluation en rigueur en RI. Les résultats obtenus ont montré l'intérêt de notre modèle de RI sémantique par rapport à un modèle de RI classique basé sur les mots-clés.

Publications dans le cadre de ce mémoire

1. Article de revue internationale

[Azzoug et al., 13a] Fatiha Boubekeur, **Wassila Azzoug**. *Concept-Based Indexing in Text Information Retrieval*. In: *International Journal of Computer Science & Information Technology* (IJCSIT), Vol. 5, No 1, February 2013, pp. 119-136. DOI : 10.5121/ijcsit.2013.5110.

2. Conférences internationales avec actes/proceedings édités et comités de sélection

[Azzoug et al., 11] Fatiha Boubekeur, **Wassila Azzoug**, Mohand Boughanem. *Indexation Sémantique de documents textuels*. Dans : 14 ème *Conférence Internationale sur le Document Electronique (CIDE 2011), Rabat, 07/12/2011-O8/12/2011*, EUROPIA, Décembre 2011, Rabat, Maroc.

[Azzoug et al., 12] Fatiha Boubekeur, **Wassila Azzoug**, Mohand Boughanem. *Les concepts sont-ils de bons candidats à l'indexation ?* Dans: *9ème édition du colloque sur l'optimisation et les systèmes d'information (COSI 2012), Tlemcen, Algérie, -12/05/2012- 15/05/2012*, Université Abou Bekr Belkaid, Mai 2012, Tlemcen, Algérie.

[Azzoug et al., 13b] Fatiha Boubekeur, **Wassila Azzoug**. *Pondération des Concepts en Indexation Sémantique*. Dans: Dixième édition de la COnférence en Recherche d'Information et Applications, CORIA 2013, 3-5 Avril, Neuchâtel, Suisse.

3. Conférences nationales avec actes édités et comités de sélection

[Azzoug et al., 13c] **Wassila Azzoug,** Fatiha Boubekeur. *Désambiguïsation des sens des mots -application en recherche d'information-*. Dans: 7ème Journées scientifiques pour la présentation des travaux de recherches des domaines de l'informatique, INFODays' 2013, Université Hassiba BenBouali, 15-16 Mai 2013, Chlef, Algérie.

Organisation du mémoire

Ce mémoire est organisé en quatre chapitres regroupés en deux parties.

– La première partie présente le contexte et les problématiques qui ont motivé nos travaux. Elle est composée de deux chapitres:

-Le chapitre 1 introduit les concepts et notions de base de la RI classique (indexation, recherche et appariement, reformulation des requêtes). Les principaux modèles de recherche (booléens, vectoriels et probabilistes) y seront aussi présentés, ainsi que le protocole et mesures d'évaluation des SRI.

- Le chapitre 2 présente l'état de l'art en indexation sémantique. Les principaux travaux en désambiguïsation des sens des mots ainsi que les approches d'indexation sémantique basées sur ses travaux sont détaillées.

– La seconde partie présente nos contributions. Elle est composée des deux chapitres:

-Le chapitre 3 présente les fondements théoriques de notre modèle de RI sémantique, et détaille nos différentes propositions (indexation, pondération et évaluation des requêtes) pour ce modèle.

-Le chapitre 4 présente les expérimentations que nous avons menées en vue d'évaluer notre modèle de RI sémantique. Les résultats obtenus y sont aussi détaillés.

Nous terminons par une conclusion et des perspectives.

PARTIE 1

De la RI classique à la RI sémantique

Chapitre 1

Recherche d'information

Plan du chapitre

1.1 Introduction 7

1.2 Concepts de base de la RI 7

1.3 Processus de recherche d'information.....9

1.3.1 Le processus d'indexation.....10

1.3.2 Les modèles de recherche d'information 14

1.3.3 Reformulation de requêtes 21

1.4 Evaluation des SRI 23

1.4.1 Collections de test 24

1.4.2 Protocole d'évaluation d'un SRI.....24

1.4.3 Métriques d'évaluation.....25

1.5 Conclusion 28

1.1 Introduction

Face à l'accroissement rapide du volume documentaire stocké sous format numérique, est née la nécessité de mettre en place des systèmes et mécanismes facilitant l'accès aux informations contenues dans de tels volumes documentaires. Les systèmes mis en œuvre dans le cadre de la recherche d'information (RI), encore appelés systèmes de recherche d'information (SRI), offrent des mécanismes et des techniques qui facilitent le stockage, l'organisation et l'accès aux informations souhaitées, contenues dans des collections de documents. Leur objectif principal étant de retrouver les documents pertinents susceptibles de répondre au mieux à un besoin en information d'un utilisateur, exprimé sous forme de requête.

Notre but à travers ce présent chapitre, est de présenter les fondements de base de la RI et des SRI. En particulier, nous définissons en section 1.2, les concepts de base de la RI classique. Nous décrivons en section 1.3, les différentes étapes d'un processus de RI (indexation, appariement, reformulation de requête). La section 1.4 est dédiée à la présentation des principales techniques d'évaluation d'un SRI.

1.2 Concepts de base de la RI

La *Recherche d'Information* (RI ou *Information Retrieval* en anglais) est une discipline de l'informatique qui s'intéresse à l'acquisition, l'organisation, le stockage, la recherche et la sélection d'informations [Salton, 68]. La RI est mise en œuvre à travers des systèmes de recherche d'information (SRI).

Un *Système de Recherche d'Information* (*SRI*) est un système informatique constitué d'un ensemble de programmes, dont l'objectif principal est de sélectionner, dans une *collection de documents* préalablement enregistrée, les informations (*documents) pertinentes* répondant à un *besoin en information* formellement exprimé par un utilisateur sous forme de *requête*. La Figure 1.1 illustre ce fonctionnement.

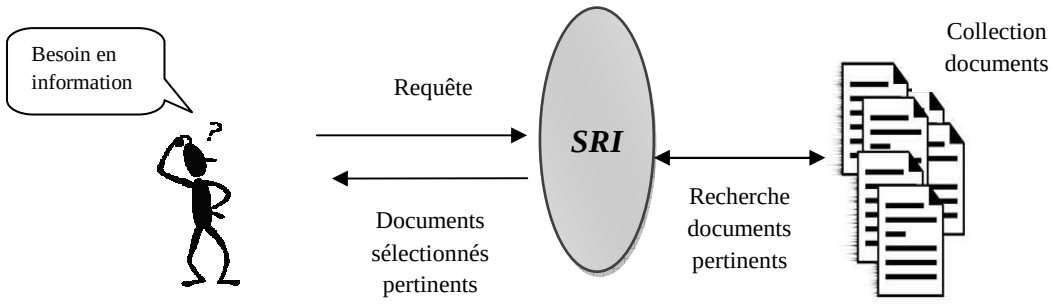


Figure 1.1 : SRI en réponse à une requête utilisateur.

Cette définition d'un SRI fait ressortir les quatre concepts clés suivants : *document*, *besoin en information*, *requête* et *pertinence*.

- Le *document* représente l'information élémentaire recherchée par un SRI. Cette information structurée (HTML, XML, …) ou non structurée (textuelle), peut apparaitre sous plusieurs formes (texte, image, vidéo, son, …) et dans différents langages (français, anglais, arabe, …). L'ensemble des documents sur lequel porte une recherche forme une *collection de documents*. (Nous focalisons dans la suite de ce mémoire sur les documents textuels non structurés).

- Le *besoin en information* est l'expression mentale de ce que l'utilisateur recherche. Trois types de besoins ont été définis dans [Ingwersen, 92] :

- le *besoin vérificatif* : l'utilisateur cherche à vérifier une information particulière dont il sait comment y accéder (ex : la recherche d'un article à partir d'une adresse web connue). Ce type de besoin est stable et ne change pas au cours de la recherche ;
- le *besoin thématique connu* : l'utilisateur cherche à clarifier, à revoir ou à trouver de nouvelles informations dans un domaine connu (ex : l'utilisateur cherche les remèdes d'une maladie qu'il connait déja). Ce type de besoin s'affine généralement au cours de la recherche ;
- le *besoin thématique inconnu* : l'utilisateur cherche de nouvelles connaissances dans un domaine ou un sujet qu'il ne connait pas. Un besoin de ce type est généralement exprimé de manière incomplète et imprécise.

- Une *requête* est la représentation structurée du besoin en information d'un utilisateur. Elle est formulée en langage naturel ou dans un langage graphique ou booléen. La requête représente l'interface entre l'utilisateur et le SRI.

- La *pertinence* est une notion fondamentale en RI. On la classifie généralement en deux types [Saracevic, 75 ; Mizzaro, 97 ; Borlund et al., 98]:

- la *pertinence système* ou *algorithmique* : représente la relation entre la requête et l'information portée par les documents restitués par le SRI. Elle définit « *une correspondance entre un document et une requête, ou encore une mesure d'informativité du document à la requête* » [Boughanem et al, 08]. Pratiquement, la pertinence système se traduit par un score de pertinence basé sur le degré de similitude entre un document et une requête donnée.
- la *pertinence utilisateur* : c'est une mesure subjective qui représente la satisfaction de l'utilisateur vis-à-vis des documents retournés par le système.

L'objectif de tout système de recherche d'information (SRI) est de rapprocher la *pertinence système* de la *pertinence utilisateur*.

1.3 Processus de recherche d'information

Le fonctionnement général d'un SRI est résumé au travers du processus de recherche, couramment appelé *processus en U de la recherche d'information* [Belkin et al, 92], illustré en figure1.2.

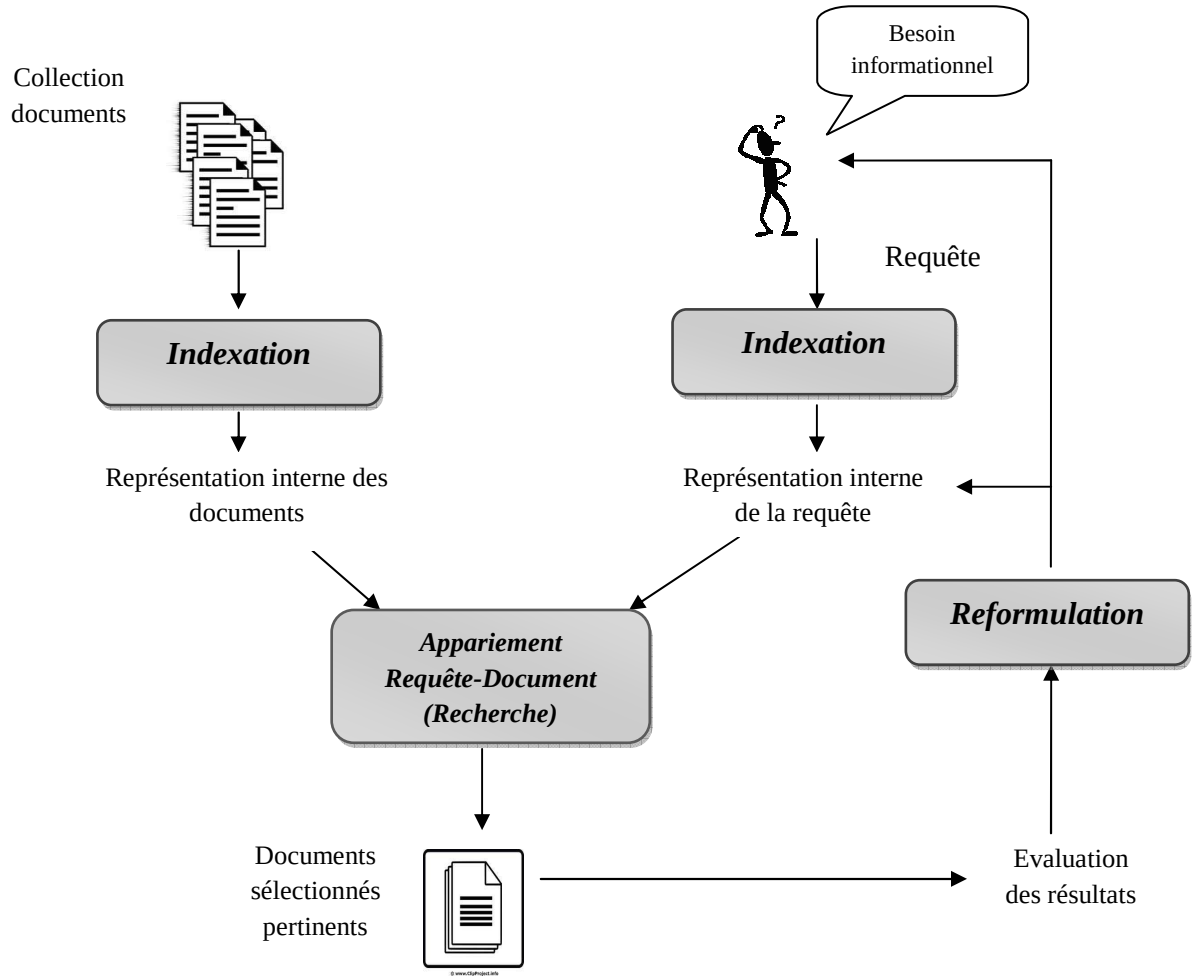


Figure 1.2 : Processus en U de la recherche d'information.

En RI, l'utilisateur interroge le SRI à travers une requête. Ce dernier lui renvoie l'ensemble des documents sensés correspondre au mieux à la requête. Pour cela, le SRI met en œuvre deux principales fonctions de base : un *processus d'indexation* et un *processus d'appariement document-requête*

- Le *processus d'indexation :* consiste à construire pour chaque document de la collection (respectivement pour la requête) une représentation interne appelée *descripteur* ou *index*. Ce dernier est composé d'un ensemble de mots-clés permettant de décrire au mieux le contenu informationnel du document (respectivement de la requête).

- Le *processus d'appariement document-requête :* a pour objet de sélectionner les documents pertinents à une requête utilisateur. Pour ce faire, ce processus compare et calcule le degré de ressemblance entre la représentation interne de la requête et les représentations internes des documents de la collection. Les documents qui correspondent au mieux à la requête, ou documents pertinents, sont retournés à l'utilisateur dans une liste triée selon l'ordre décroissant de leur degré de pertinence lorsque le système le permet.

En plus de ces deux modules de base, le système peut être doté d'un mécanisme d'amélioration et de raffinement de la requête utilisateur par un *processus de reformulation*.

1.3.1 Le processus d'indexation

L'indexation est une étape primordiale dans le processus de recherche d'information. De sa qualité dépend en partie la qualité des réponses du système. En effet, les documents et les requêtes dans leur forme texte libre, sont difficiles à exploiter par la machine lors la recherche. Un traitement préalable permettant leur représentation simplifiée est nécessaire : c'est l'indexation. L'indexation consiste à analyser les documents et les requêtes dans le but d'en définir un ensemble de descripteurs (*termes d'index*) permettant d'exploiter plus facilement leur contenu lors du processus de recherche. Dans une indexation classique, les termes d'index sont des mots-clés simples ou composés. Ils sont organisés dans une liste de descripteurs, *l'index*, caractérisant le contenu informationnel d'un document (ou d'une requête). L'ensemble de tous les termes d'index constitue le *langage d'indexation*. Ce langage peut être libre ou contrôlé.

- Le *langage d'indexation libre* est construit à partir des termes extraits directement du texte analysé par le système. Cette représentation est facile à construire et permet une large couverture du contenu informationnel du document. Néanmoins, elle présente des problèmes du fait de l'ambiguïté des mots de la langue naturelle.

- Le *langage d'indexation contrôlé* est construit à partir d'un ensemble de concepts associés au contenu informationnel du texte analysé. Ces concepts sont à priori connus et généralement organisés dans un thésaurus[1] (ex : le thésaurus Roget [Roget, 95]). Ce type de langage offre une représentation plus riche et sémantiquement plus précise du contenu informationnel des documents (ou requêtes), évitant ainsi les problèmes liés à l'ambiguïté des mots de la langue naturelle.

1.3.1.1 Techniques d'indexation

L'indexation des documents et requêtes peut être : manuelle, automatique ou semi-structurée.

[1] Thésaurus est un outil linguistique dynamique de termes obéissant à des règles terminologiques propres et reliés entre eux par des relations sémantiques (relation d'équivalence, relation d'association, la synonymie, l'hyperonymie …etc.).

- *L'indexation manuelle :* est un processus d'indexation où le document (ou la requête) est analysé par un expert du domaine ou un documentaliste qui se charge d'en représenter le contenu informationnel en utilisant un vocabulaire (ou un langage) contrôlé qui dépend de son savoir propre. L'indexation manuelle assure une meilleure précision de recherche en réponse à une requête utilisateur [Nie et al., 99]. Cependant, elle présente les inconvénients d'être :

  - subjective puisque le choix des termes d'indexation dépend des connaissances des indexeurs dans le domaine (par exemple des termes différents peuvent être affectés à un même document par des indexeurs différents, ou par un même indexeur à des instants différents).
  - très couteuse à réaliser (en temps et en nombre de personnes impliquées).
  - difficile à maintenir du fait de l'évolution de la terminologie.

- *L'indexation automatique* : est un processus d'indexation entièrement automatisé. Il met en œuvre un ensemble de techniques informatisées issues de Traitements Automatiques de la Langue Naturelle (TALN). Ce processus est le plus utilisé en RI, nous le détaillons dans la section suivante.

- *L'indexation semi-automatique* ou *indexation supervisée*: est une combinaison des deux approches d'indexations précédentes [Jacquemin et al, 02]. En indexation supervisée, les résultats issus d'une indexation automatique préalable sont exposés à un documentaliste pour enrichissement et/ou validation de la représentation (index) ainsi obtenue.

1.3.1.2 Indexation automatique

L'indexation automatique classique est basée sur la construction des descripteurs de manière automatique et rapide. Ces descripteurs représentent des mots clés significatifs décrivant le contenu informatif des documents et requêtes. L'indexation automatique repose sur les étapes suivantes :

1) L'analyse lexicale (Tokenisation/Segmentation): consiste à découper le texte d'un document (ou d'une requête) en plusieurs unités lexicales (mots simples ou composés) représentant les termes d'index.

2) L'élimination des mots vides: vise à réduire les termes d'index en éliminant les mots non porteurs de sens, ou mots vides (pronoms personnels, conjonctions, prépositions, …) du langage d'indexation, et de garder uniquement ceux qui sont importants et nécessaires pour une meilleure représentation des documents et requêtes [Luhn, 58]. L'élimination des mots peut se faire en utilisant une liste prédéfinie de mots vides (dite *stoplist* ou *anti-dictionnaire*), ou/et en écartant les mots trop fréquents ou trop rares dans la collection. Bien que ce traitement présente l'avantage d'améliorer la représentation des documents en éliminant des mots non significatifs, il peut cependant induire des effets de silence (par exemple, en éliminant le mot *a* de *vitamine a*, les documents pertinents qui contiennent ce dernier terme ne sont pas retournés par le SRI).

3) La normalisation des termes d'index: est un processus qui permet de regrouper les variantes morphologiques d'un mot sous une forme de base unique. Son objectif est de ne garder, dans le langage d'indexation, que les formes normalisées des mots représentatifs, ce qui offre un gain d'espace mémoire considérable et une recherche efficace. La normalisation se base sur l'une des deux procédures : la *racinisation* ou la *lemmatisation*.

- *la racinisation* (ou *stemming*): est un procédé qui vise à supprimer les suffixes qui différencient les flexions des mots significatifs du texte indexé (ex : la racine des mots anglais : *retrieve*, *retrieving* et *retrieval* est *retriev*). Cette technique est réalisée par l'utilisation des règles de transformation de type condition action (ex: l'algorithme de Porter [Porter, 80] pour l'anglais) ou par une troncature des mots à x caractères (ex : la troncature à 7 caractères pour le français).
- *la lemmatisation* : permet de regrouper les mots de la même catégorie grammaticale et les transformer à leur forme canonique appelée *lemme* (ex : les différentes formes d'un verbe sont transformés à son infinitif). Cette technique est basée sur l'utilisation des patrons syntaxiques et des dictionnaires (ex : TreeTagger[2]).

Des expériences ont montré que la racinisation et la lemmatisation augmentent significativement les performances de la recherche pour les langues morphologiquement riches telles que : le français, l'italien, …etc. [Gaussier et al., 97 ; Gaussier et al., 00].

4) La pondération des termes d'indexation : est une fonction fondamentale en RI. Elle consiste à mesurer l'importance d'un terme t_j dans un document d_i en lui affectant un poids w_{ij} qui exprime son degré de représentativité.

Les méthodes de pondération proposées en RI sont généralement basées sur la combinaison de deux facteurs : une *pondération locale* quantifiant l'importance locale du terme dans le document et une *pondération globale* quantifiant son importance globale vis-à-vis de la collection des documents.

- la *pondération locale* : est mesurée généralement par la fréquence du terme t_j (*term frequency*, notée tf_{ij}) dans le document considéré d_i.
- la *pondération globale :* est fondée sur l'idée qu'un terme ne permet pas de distinguer les documents les uns des autres, lors de la recherche, s'il est distribué d'une manière uniforme dans tous les documents de la collection. Ainsi, ce terme ne possède pas de pouvoir de discrimination. De ce fait, les termes qui apparaissent dans peu de documents sont discriminants et une pondération leur est attribuée. Cette pondération est exprimée par la fréquence documentaire inverse idf_j (*inverse document frequency)* du terme t_j dans la collection. Elle est définie généralement par la formule suivante:

[2] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

$$idf_j = \log\left(\frac{N}{n_j}\right) \qquad [1.1]$$

Où :

- N est le nombre de documents dans la collection.
- n_j est le nombre de documents indexés par le terme t_j.

La plupart des techniques de pondération utilisées en recherche d'information, sont fondées sur la combinaison des deux pondérations locale et globale. A titre d'exemple, la formule de pondération *tf*idf* de Salton [Salton et al., 73], est définie par :

$$w_{ij} = tf_{ij} \times idf_j = tf_{ij} \times \log\left(\frac{N}{n_j}\right) \qquad [1.2]$$

Cette mesure est une bonne approximation de l'importance d'un terme dans un document d'une collection donnée, particulièrement pour les collections composées de documents de tailles homogènes. Cependant, pour les collections contenant des documents de tailles variables, les termes dans les documents les plus longs apparaissent très fréquemment avec des poids très élevés que ceux des documents courts. Ainsi, les documents longs auront plus de chance d'être sélectionnés [Buckley et al., 95]. De ce fait, une normalisation par la taille du document a été proposée par plusieurs chercheurs dans la définition de la pondération des termes dont:

- La normalisation de Robertson [Robertson et al., 97] dans sa célèbre formule dite BM25 définie comme suit:

$$w_{ij} = \frac{tf_{ij} \times (k_1 + 1)}{k_1\left((1-b) + b \times \frac{dl_i}{\Delta l}\right) + tf_{ij}} \times \log\left(\frac{N\text{-}n + 0.5}{n + 0.5}\right) \qquad [1.3]$$

Où :

- k_1 est une constante pour contrôler l'influence de la fréquence du terme t_j dans le document d_i. Sa valeur dépend de la longueur des documents dans la collection. Le plus souvent sa valeur est fixée à 1,2.

- b est une constante qui permet de contrôler l'effet de la longueur du document. Sa valeur la plus souvent utilisée est 0,75.

- dl_i est la longueur du document d_i.

- Δl est la longueur moyenne des documents dans la collection entière.

- N est le nombre de documents dans la collection.

- n est le nombre de documents pertinents dans la collection.

1.3.2 Les modèles de recherche d'information

Un modèle de RI offre un formalisme de représentation des index issus de l'étape d'indexation, et une approche de modélisation de la mesure de pertinence d'un document d_i vis-à-vis d'une requête Q. Cette mesure notée par $RSV(d_i,Q)$ (*Retrieval Status Value*), détermine, lors du processus d'appariement, le degré de ressemblance entre les représentations respectives de d_i et de Q.

Plusieurs modèles de recherche d'information ont été proposés, qui sont classés en trois grandes catégories [Baeza-Yates et al., 99]:

- les *modèles booléens* (ou *ensemblistes*) : qui regroupent le modèle *booléen de base*, le *modèle des ensembles flous* et le modèle *booléen étendu.* Ces modèles ont été les premiers à être utilisés en RI. Ils sont inspirés de la logique booléenne et de la théorie des ensembles.

- les *modèles vectoriels* : incluent le *modèle vectoriel de base*, le *modèle vectoriel généralisé*, le *modèle LSI* (*Latent Semantic Indexing*) et le *modèle connexionniste*. Dans ces modèles la représentation des documents et requêtes est réalisée par des vecteurs de termes pondérés dans l'espace vectoriel multidimensionnel des termes d'index.

- les *modèles probabilistes* : regroupent le *modèle BIR (Binary Independance retrievel)*, le *modèle inférentiel Bayesien* et le *modèle de langue*. Ces modèles sont basés sur les probabilités d'appartenance des termes de la requête aux documents de la collection.

Dans ce qui suit, nous décrivons pour chacune de ces classes son modèle de base et un des modèles qui lui sont associés.

1.3.2.1 Les modèles booléens

1.3.2.1.1 Le modèle booléen de base (ou standard)

Le modèle booléen standard est basé sur la théorie des ensembles et l'algèbre de Boole. Dans ce modèle, un document d_i est représenté par l'ensemble de ses termes descriptifs (non pondérés). Une requête Q est une expression booléenne composée des mots clés reliés par des opérateurs logiques (AND, OR et NOT).

La mesure de pertinence document-requête est calculée selon la fonction booléenne suivante :

$$RSV(d_i, Q) = \begin{cases} 1 & \text{si } d_i \text{ contient l'ensemble de l'expréssion booléenne décrite par } Q \\ 0 & \text{sinon} \end{cases} \qquad [1.4]$$

Par exemple, pour la requête Q exprimée par : « *information AND retrieval* », le SRI sélectionnera l'ensemble des documents qui sont indexés à la fois avec les deux termes : *information* et *retrieval.*

Ce modèle est simple et facile à mettre en œuvre, néanmoins il présente les inconvénients suivants :

- les documents restitués par le SRI ne sont pas ordonnés. En effet, la mesure de correspondance document-requête est binaire (soit 1 ou 0), ce qui rend impossible de différencier le degré de pertinence d'un document par rapport aux autres documents pertinents retournés par le système de recherche.
- la formulation de la requête nécessite une connaissance des opérateurs booléens, ce qui n'est pas une tâche évidente pour tous les utilisateurs.

Pour remédier à ces problèmes, des extensions ont été proposées (telles que : le modèle basé sur les ensembles flous [Zadeh, 65] et le modèle booléen étendu proposé par Salton [Salton et al., 83]), qui permettent d'intégrer dans le modèle de recherche, la représentativité des termes dans les documents.

1.3.2.1.2 Le modèle booléen étendu

Dans ce modèle, la représentation de la requête reste une expression booléenne classique. Tandis que les termes représentant un document sont pondérés [Salton et al., 83]. L'appariement document-requête est déterminé par les relations introduites par le modèle p-norm basé sur les p-distances avec $p \in [1, \infty[$. Si m est le nombre de termes dans la requête, la fonction de similarité est définie comme suit :

$$\begin{cases} RSV(d_i, Q_{OR}) = \left(\dfrac{w_{i1}^p + w_{i2}^p + ... + w_{ij}^p + ... + w_{im}^p}{m} \right)^{\frac{1}{p}} \\ RSV(d_i, Q_{AND}) = 1 - \left(\dfrac{(1-w_{i1})^p + (1-w_{i2})^p + ... + (1-w_{ij})^p + ... + (1-w_{im})^p}{m} \right)^{\frac{1}{p}} \end{cases} \quad [1.5]$$

Où : w_{ij} est le poids du terme t_j de la requête Q dans le document d_i.

1.3.2.2 Les modèles vectoriels (algébriques)

1.3.2.2.1 Le modèle vectoriel de base

Le modèle vectoriel est un modèle algébrique où l'on représente les documents et les requêtes par des vecteurs de poids dans l'espace vectoriel des termes d'index [Salton, 71 ; Salton et al., 83]. Formellement :

- Un document d_i est représenté par un vecteur de dimension n :

$$d_i = (w_{i1}, w_{i2}, w_{i3}, \ldots, w_{ij}, \ldots, w_{in})$$

Où :

- w_{ij} est le poids d'un terme t_j dans le document d_i,
- n est le nombre de termes d'index distincts appartenant aux documents de la collection.

- Une requête Q est aussi représentée par un vecteur de poids dans le même espace vectoriel que le document d_i.

$$Q = (w_{Q1}, w_{Q2}, w_{Q3}, \ldots, w_{Qj}) \text{ pour } j = 1, 2, \ldots, n.$$

Où :

- w_{Qj} est le poids du terme t_j dans la requête Q.

La pertinence du document d_i vis-à-vis de la requête Q est mesurée par le degré de corrélation de leurs vecteurs correspondants. Cette corrélation peut être exprimée par l'une des mesures suivantes :

- Le produit scalaire :

$$RSV(d_i, Q) = \sum_{j=1}^{n} w_{Qj} * w_{ij} \qquad [1.6]$$

- La mesure du cosinus:

$$RSV(d_i, Q) = \frac{\sum_{j=1}^{n} w_{Qj} * w_{ij}}{\sqrt{\sum_{j=1}^{n} w_{Qj}} * \sqrt{\sum_{j=1}^{n} w_{ij}}} \qquad [1.7]$$

- La mesure de Dice :

$$RSV(d_i, Q) = \frac{2 * \sum_{j=1}^{n} w_{Qj} * w_{ij}}{\sum_{j=1}^{n} w_{Qj}^2 + \sum_{j=1}^{n} w_{ij}^2} \qquad [1.8]$$

- La mesure de Jacard :

$$RSV(d_i, Q) = \frac{2 * \sum_{j=1}^{n} w_{Qj} * w_{ij}}{\sum_{j=1}^{n} w_{Qj}^2 + \sum_{j=1}^{n} w_{ij}^2 - \sum_{j=1}^{n} w_{Qj} * w_{ij}} \qquad [1.9]$$

- Le coefficient de superposition :

$$RSV(d_i, Q) = \frac{\sum_{j=1}^{n} w_{Qj} * w_{ij}}{\min_i \left(\sum_{j=1}^{n} w_{Qj}^2 + \sum_{j=1}^{n} w_{ij}^2 \right)} \qquad [1.10]$$

Le modèle vectoriel de base est l'un des modèles les plus utilisés en RI. Son avantage, comparé au modèle booléen de base, réside dans sa capacité à ordonner les résultats de la recherche selon leur degré de pertinence pour la requête utilisateur. Cependant, ce modèle suppose que les termes d'index sont indépendants, et ne tient pas compte des relations sémantiques qui peuvent exister entre ces termes dans le même document ou la même requête. Par conséquent, plusieurs extensions permettant de remédier à cette insuffisance ont été proposées. On peut distinguer en particulier le modèle vectoriel généralisé (*Generalized Vector Space Model*) proposé par [Wong et al, 85], qui ne considère pas l'hypothèse d'indépendance des termes d'index, ou encore le modèle basé sur l'analyse sémantique latente *LSI* (*Latent Semantic Indexing*) [Deerwester et al., 90] qui tient en compte de certaines relations sémantiques latentes entre les termes d'indexation.

1.3.2.2.2 Le modèle d'indexation sémantique latente (LSI)

Le modèle *LSI* (ou *Latent Semantic Indexing*) proposé par [Deerwester et al., 90], est basé sur la décomposition en valeurs singuliers (*SVD* ou *Singulair Value Decomposition)* de la matrice termes-documents, qui représente l'espace d'indexation du modèle vectoriel. Cette décomposition permet de projeter la matrice termes-documents dans un espace de dimension réduit permettant de faire ressortir les relations sémantiques latentes entre mots des documents. Ces relations sont basées sur la notion de cooccurrence où deux mots peuvent être considérés sémantiquement proches s'ils apparaissent dans des contextes (ou documents) similaires. Ainsi, dans ce modèle, les documents qui partagent des termes co-occurrents proches sont groupés (ou clustérisés) dans une seule représentation. Formellement :

Si X est la matrice termes-documents de dimension n x d (où n est le nombre de termes distincts de la collection, et d est le nombre de documents dans cette collection), alors *SVD* la décompose en :

$$X_{n\times d} = T_{n\times m} \times S_{m\times m} \times D'_{m\times d} \qquad [1.11]$$

Où :

- $X_{n\,x\,d}$: matrice termes-documents.
- $T_{n\,x\,m}$ est la matrice orthogonale des vecteurs singuliers de gauche.
- m est le rang de M, tel que ($m \leq min(n, d)$).
- $S_{m\,x\,m}$ est la matrice diagonale triée des valeurs singulières.
- $D_{d\,x\,m}$ est la matrice contenant les colonnes orthogonales des vecteurs singuliers de droite. ($D_{d\,x\,m}$ x $D'_{m\,x\,d} = I$, tel que : $D'_{m\,x\,d}$ est la matrice transposée de $D_{d\,x\,m}$).

Une fois que la SVD de la matrice X est calculée, il s'agit de réduire $X_{n\,x\,d}$ par la matrice $Y_{n\,x\,d}$ contenant uniquement les k termes ayants les plus grandes valeurs singulières de $S_{m\,x\,m}$.

La matrice réduite $Y_{n\,x\,d}$ dans l'espace de dimension k est calculée par la formule suivante :

$$Y_{n\times d} = T_{n\times k} \times S_{k\times k} \times D'_{k\times d} \qquad [1.12]$$

D'autre part, la requête *Q* est aussi transformée dans ce nouvel espace en un pseudo-document D_Q comme suit :

$$D_Q = X'_Q \times T_{n\times k} \times S^{-1}_{k\times d} \qquad [1.13]$$

Où : X_Q est le vecteur contenant les mots-clés de la requête *Q*. X'_Q est son transposé.

La requête, ou pseudo-document, est ajoutée dans la matrice $D_{k\ x\ d}$ comme un nouveau document. Lors de la recherche, le système calcule la similarité entre chaque paire de documents en vérifiant la formule [1.14], puis les documents qui sont proches sémantiquement sont comparés au pseudo-document D_Q suivant le modèle vectoriel de base afin de calculer le degré de pertinence entre la requête *Q* et ces documents.

$$Y_{t\times d} \times Y'_{d\times t} = D_{k\times d} \times S^2_{k\times k} \times D'_{k\times d} \qquad [1.14]$$

L'avantage principal du modèle LSI est son pouvoir de retrouver les documents pertinents pour une requête utilisateur même s'ils ne partagent aucun mot avec elle. Il permet en outre de résoudre partiellement les problèmes liés à la polysémie et la synonymie des mots dans la représentation des documents. Néanmoins, ce modèle perd son efficacité comparativement au modèle vectoriel de base, lorsque le nombre de documents est faible. En effet, une collection de petite taille donne une approximation erronée de la matrice documents-termes dans l'espace réduit.

1.3.2.3Les modèles probabilistes

1.3.2.3.1 Le modèle probabiliste de base

Le premier modèle probabiliste a été proposé par Maron et Kuhns [Marons et al., 60]. L'idée de base de ce modèle est de sélectionner les documents ayant à la fois une forte probabilité d'être pertinents et une faible probabilité d'être non pertinents à une requête utilisateur. Robertson [Robertson, 77] définit son modèle PRP (*Probability Ranking Principle*), sur un principe similaire. Dans ce modèle, la similarité entre un document d_i et une requête *Q*, est estimée par le rapport entre sa probabilité qu'il soit pertinent à *Q* (notée $P(R/d_i)$) et sa probabilité qu'il ne soit pas pertinent à *Q* (notée $P(NR/d_i)$). Ainsi, le score d'appariement entre le document d_i et la requête *Q* est donné par la formule suivante :

$$RSV(d_i, Q) = \frac{P(R/d_i)}{P(NR/d_i)} \qquad [1.15]$$

En appliquant le théorème de Bayes et après simplification, on obtient :

$$RSV\ (d_i, Q) = \frac{P(R/d_i)}{P(NR/d_i)} \approx \frac{P(d_i/R)}{P(d_i/NR)} \qquad [1.16]$$

Où :

- $P(d_i/R)$ est la probabilité que d_i appartienne à l'ensemble de documents pertinents.
- $P(d_i/NR)$: est la probabilité que d_i appartienne à l'ensemble de documents non pertinents.

Plusieurs méthodes ont été proposées pour estimer ces différentes probabilités. Pour calculer cette mesure, le modèle d'indépendance binaire, connu sous le modèle *BIR* (*Binary Independance Retreival)*, suppose l'indépendance des termes dans les documents. Ainsi, chaque document d_i de la collection est représenté par un ensemble d'événements indépendants qui dénotent l'absence ou la présence d'un terme dans ce document.

Formellement :

$$d_i = \{t_1 = x_1, t_2 = x_2, ..., t_j = x_j, ..., t_n = x_n\} \qquad où : x_j = \begin{cases} 1 \ si \ t_j \text{ est présent dans } d_i \\ 0 \text{ sinon} \end{cases}$$

L'application de la distribution de la loi de Bernouli sur les probabilités $P(d_i/R)$ et $P(d_i/NR)$, permet d'obtenir les résultats suivants:

$$P(d_i / R) = \prod_{j=1}^{n} P(t_j = x_j / R) = \prod_{j=1}^{n} P(t_j = 1/R)^{x_j} P(t_j = 0/R)^{(1-x_j)} = \prod_{j=1}^{n} p_j^{x_j} (1 - p_j)^{(1-x_j)}$$

$$P(d_i / NR) = \prod_{j=1}^{n} P(t_j = x_j / NR) = \prod_{j=1}^{n} P(t_j = 1/NR)^{x_j} P(t_j = 0/NR)^{(1-x_j)} = \prod_{j=1}^{n} q_j^{x_j} (1 - q_j)^{(1-x_j)}$$

Ainsi :

$$RSV(d_i, Q) = \frac{\prod_{j=1}^{n} p_j^{x_j} (1 - p_j)^{(1-x_j)}}{\prod_{j=1}^{n} q_j^{x_j} (1 - q_j)^{(1-x_j)}}$$

Où :

- p_j est la probabilité qu'un terme t_j soit présent dans l'ensemble des documents pertinents.
- q_j est la probabilité qu'un terme t_j soit présent dans l'ensemble des documents non pertinents.

Après quelques transformations, le score de pertinence du document d_i pour la requête Q dans le modèle *BIR*, est donné selon la formule suivante :

$$RSV(d_i, Q) = \sum_j \log \frac{p_j (1 - q_j)}{q_j (1 - p_j)} \qquad [1.17]$$

En supposant connus l'ensemble R des documents pertinents et l'ensemble NR des documents non pertinents, il est possible d'estimer les probabilités de pertinence p_j et q_j comme suit :

$$p_j = \frac{r_j}{n} \qquad \text{et} \qquad q_j = \frac{R_j - r_j}{N - n}$$

Où :

- N est le nombre de documents dans la collection.
- n est le nombre de documents pertinents dans la collection.
- R_j est le nombre de documents contenant le terme t_j.
- r_j est le nombre de documents pertinents contenant le terme t_j.

En remplaçant les valeurs de p_j et q_j dans la formule [1.17], la pertinence du document d_i pour la requête Q sera calculée par la formule suivante:

$$RSV\ (d_i, Q) = \sum_j \log \frac{r_j (N - R - n - r_j)}{(n - r_j)(R_j - r_j)} \qquad [1.18]$$

1.3.2.3.2 Le modèle de langue

Un modèle de langue tente de modéliser l'agencement de mots dans une langue donnée en estimant la probabilité de distribution d'une séquence de mots dans cette langue. Les travaux de Ponte et Croft [Ponte et al, 98] ont été les premiers à proposer l'utilisation des modèles de langue en RI. L'idée de base admet qu'un document d_i de la collection, est vu comme une succession de mots générée par son propre modèle de langue du document M_{di}. La pertinence du document d_i vis-à-vis d'une requête utilisateur Q est alors estimée par la probabilité que Q soit inférée par le modèle du document M_{di}.

Formellement :

$$RSV(d_i, Q) = P\left(Q / M_{di}\right) = P(Q = (t_1 t_2 .. t_j .. t_n) / d_i) = \prod_{t_j \in Q} P(t_j / d_i) \qquad [1.19]$$

La probabilité $P(t_j / d_i)$ est mesurée par l'estimation maximale de vraisemblance (*maximum likehood esitmation)*. Elle est donnée par :

$$P(t_j / d_i) = \prod_{t_j \in Q} \frac{tf\ (t_j / d_i)}{N_d} \qquad [1.20]$$

Où :

- $tf\ (t_j/d_i)$ est la fréquence d'occurrence du terme t_j dans le document d_i.
- N_d est le nombre total de termes dans le document d_i.

Dans cette estimation si l'un des termes de la requête *Q* est absent du document d_i alors le score de pertinence $RSV(d_i,Q)$ sera nul. Afin de pallier à ce problème, des techniques de lissage[3] (ex : le lissage de Laplace, le lissage de Backoff, le lissage par interpolation) ont été utilisées permettant d'assigner une probabilité non nulle aux termes de la requête qui n'apparaissent pas dans le document.

1.3.3 Reformulation de requêtes

En RI, l'utilisateur formule son besoin en information par le biais d'une requête dans l'espoir de trouver des réponses pertinentes à ce qu'il recherche. La qualité des réponses dépend d'une part des termes utilisés par l'utilisateur pour formuler son besoin, et d'autre part des termes utilisés dans l'indexation des documents. Du fait de l'ambiguïté/imprécision de la langue naturelle, l'utilisateur peut formuler sa requête dans un vocabulaire différent de celui utilisé par les auteurs/indexeurs des documents, ce qui a pour conséquence d'influer négativement sur la qualité des résultats de la recherche. Pour résoudre ces problèmes, Van Riejsbergen [Van Riejsbergen, 79] introduit le mécanisme de reformulation de requête dans les SRI.

La reformulation de la requête est alors considérée comme un processus ayant pour objectif de générer une nouvelle requête plus ciblée permettant d'obtenir des résultats de recherche plus pertinents que ceux obtenus par la requête initialement formulée par l'utilisateur. Le processus de reformulation de requête se base sur les deux étapes suivantes : la première consiste à étendre la requête initiale par des termes jugés pertinents par l'utilisateur ou par le système de recherche, et la seconde consiste à réajuster les poids des termes de la nouvelle requête.

Deux techniques de reformulation sont à distinguer : la *reformulation interactive* et la *reformulation automatique.*

1.3.3.1 Reformulation interactive de la requête

Cette technique est la plus utilisée dans le domaine de la RI [Rocchio, 71 ; Buckley et al., 94 ; Boughanem et al., 99 ; Smyth et al., 05]. On la désigne communément par *réinjection de pertinence* (ou *relevance feedback* en anglais). L'enrichissement de la requête, dans cette méthode, se fait interactivement entre le SRI et l'utilisateur, de manière *explicite* (*réinjection de pertinence explicite*) ou de manière *implicite* (*réinjection de pertinence implicite*).

1 - dans la *réinjection de pertinence explicite* (*explicit relevance feedback*) : l'approche la plus reconnue pour ce type de réinjection est celle proposée par Rocchio [Rocchio, 71] adaptée au modèle vectoriel. Dans cette approche, le *feedback* se base principalement sur les étapes suivantes :

[3] Un état de l'art détaillé sur les techniques de lissage est donné dans [Boughanem et al., 04].

- la première est l'échantillonnage, qui consiste à construire, à partir des documents retrouvés par le SRI en réponse à une requête utilisateur initiale Q_i, un échantillon de documents jugés pertinents par l'utilisateur.
- la seconde est l'extraction des termes les plus significatifs de l'échantillon de documents considéré.
- La troisième est l'expansion de requête. Elle consiste à générer une nouvelle requête enrichie Q_n par adjonction des termes issus de l'étape précédente à la requête initiale Q_i. La requête Q_n ainsi obtenue est ensuite repondérée. Les poids des termes de Q_n serviront à la discrimination entre le vecteur des documents pertinents et celui des documents non-pertinents. Rocchio [Rocchio, 71] pondère un terme dans la nouvelle requête Q_n par la formule suivante :

$$w_{Qnj} = \alpha\, w_{Qij} + \beta \frac{1}{|\mathrm{R}|} \sum_{k=1}^{|R|} w_{kj} - \delta \frac{1}{|NR|} \sum_{p=1}^{|NR|} w_{pj} \qquad [1.21]$$

Où :

- $w_{Q_{nj}}$ est le poids du terme t_j dans la nouvelle requête Q_n.
- $w_{Q_{ij}}$ est le poids du terme t_j dans la requête initiale Q_i.
- R est l'échantillon de documents restitués par le SRI et jugés pertinents par l'utilisateur. $|R|$ représente le nombre de documents de R.
- NR est l'ensemble de documents restitués par le SRI et jugés non pertinents par l'utilisateur. $|NR|$ représente le nombre de documents de R.
- w_{kj} est le poids du terme t_j dans le document d_k de R.
- w_{pj} est le poids du terme t_j dans le document d_p de NR.
- α, β et δ sont des constantes choisies en fonction de l'importance que l'on souhaite donner à la requête Q_i. .

2 - dans la *réinjection de pertinence implicite* (*implicit relevance feedback*) : les informations sur le profil d'un utilisateur et ses différents comportements durant la recherche (sauvegarde des documents dans le marque-page [Oard et al., 98], le nombre de défilement sur la page [Claypool et al., 01], la durée de consultation des documents [Kelly et al., 01 ; White et al., 02], le nombre de cliques de souris sur les documents [Smyth et al., 05]…etc), révèlent de manière silencieuse les documents que l'utilisateur considère comme pertinents parmi ceux restitués par le SRI. Ces informations sont exploitées comme sources d'évidence pour définir les termes importants des documents, jugés implicitement pertinents par l'utilisateur, et qui seront impliqués dans la réinjection de pertinence.

1.3.3.2 Reformulation automatique de la requête

Dans cette technique [Croft et Harper, 88 ; Robertson, 91 ; Clinchant et al., 10 ; Hammache et al., 13], couramment appelée pseudo-réinjection de pertinence (*Pseudo Relevance Feedback* en anglais) ou réinjection de pertinence aveugle (*Blind feedback*), l'expansion de la requête s'effectue de manière automatique sans intervention de l'utilisateur dans le processus de reformulation. En pratique, il s'agit de construire la requête enrichie Q_n par adjonction des termes les plus significatifs des *k* premiers documents retrouvés par le système (documents pseudo pertinents). Les termes de la requête Q_n sont ensuite repondérés. Dans le modèle probabiliste *DFR* (*Divergence From Randomnes*) [Amati et al., 02], la repondération de ces termes est réalisée par l'utilisation de la mesure de *Kullback-Leilber* (*KL*) ou l'une des mesures de *Bose-Einstein* (*Bo*) [Amati et al., 02].

Formellement :

$$w_{Q_{nj}} = \alpha\, w_{Q_{ij}} + \beta \times a_j \qquad [1.22]$$

Avec :

- $w_{Q_{nj}}$ est le poids du terme t_j dans la nouvelle requête Q_n.
- $w_{Q_{ij}}$ est le poids du terme t_j dans la requête initiale Q_i.
- a_j est le poids normalisé du terme t_j en appliquant l'une des mesures de *Kullback-Leilber* ou de *Bose-Einstein*.
- β est une constante choisie en fonction de l'importance que l'on souhaite donner à la requête Q_i.

La reformulation par pseudo réinjection n'est efficace que si les *k* premiers documents considérés dans cette technique, sont pertinents pour la requête initiale Q_i. Dans le cas contraire, elle peut engendrer une dégradation des performances de la RI.

1.4 Evaluation des SRI

L'évaluation des performances d'un système de recherche d'information peut porter sur plusieurs critères. Selon Cleverdon [Cleverdon, 70], les principaux critères pour mesurer la qualité d'un SRI se résument par :

- le temps de réponse du système,
- la présentation des résultats,
- l'univers du discours de la collection,
- l'effort requis de l'utilisateur pour récupérer l'information pertinente,
- le taux de rappel du système et de précision.

Parmi ces critères, les facteurs les plus importants sont ceux qui permettent de mesurer la capacité du système à satisfaire le besoin de l'utilisateur en information. Cette satisfaction se traduit par le degré d'adéquation entre la requête émise par l'utilisateur et les documents restitués par le système de recherche. C'est dans ce contexte que plusieurs campagnes d'évaluation (dont les campagnes TREC[4] - *Text Retrieval Conference*- [Harman, 92] qui constituent la référence pour l'évaluation des SRI,) ont été mises en place dont l'objectif est de fournir aux chercheurs en RI un cadre unifié (collections de test, protocole d'évaluation et mesures d'évaluation) qui leur permette d'évaluer les performances de leurs systèmes de recherche et de juger de leur qualité.

1.4.1 Collections de test

Une collection de test définit l'environnement de recherche unifié qui assure l'objectivité de l'évaluation d'un SRI. Les collections de test sont obtenues gratuitement à partir des sites web (cas des collections standards : CACM[5], CISI[6], TIME[7], …etc.), ou fournies par des campagnes d'évaluation des SRI, telles que les campagnes TREC, les campagnes CLEF[8] (*Cross-Language Forum*) pour l'évaluation des systèmes multilingues, les campagnes INEX[9] (*INitiative for the Evaluation of XML retrieval*) pour l'évaluation des systèmes de recherche dans des collections semi-structurées XML, …etc.

Une collection de test est composée d'un ensemble de documents (ou corpus de documents), d'une liste de requêtes prédéfinies (également appelées *topics*), ainsi que des jugements de pertinence (liste de documents jugés pertinents pour chaque requête de la collection) établis manuellement par des assesseurs humains.

1.4.2 Protocole d'évaluation d'un SRI

Un protocole d'évaluation a pour objectif de décrire la stratégie à appliquer pour évaluer les performances d'un système de recherche. Le protocole d'évaluation le plus connu et le plus utilisé en RI est le protocole TREC adopté dans les campagnes TREC. Dans ce protocole, les participants à la campagne TREC disposent d'une collection de test sur laquelle ils testent leurs systèmes. Les 1000 premiers documents pertinents restitués par un système participant, pour chaque requête de la collection de test, sont examinés. Puis, différentes métriques sont utilisées pour calculer, à partir de ces résultats de recherche, différentes mesures de performance du système.

4 http://trec.nist.gov

5 http : //www.dcs.gla.ac.uk/idom/ir_resources/test_collections/cacm/

6 http://ir.dcs.gla.ac.uk/resources/test_collections/cisi/

7 https ://isserver11.princeton.edu/

8 http://clef.iei.pi.cnr.it

9 http://www.inex.otago.ac.nz

1.4.3 Métriques d'évaluation

Etant donnée une requête *Q*, les documents de la collection sont classés, selon leur rapport à la requête, en document pertinents et documents non pertinents. Les documents retrouvés par le système de recherche en réponse à la requête *Q* peuvent être pertinents ou non.

Dans ce qui suit, on notera:

- l'ensemble des documents pertinents de la collection par *DP*.
- l'ensemble des documents non pertinents de la collection par *DNP*.
- l'ensemble des documents pertinents retrouvés par *DPS*.
- l'ensemble des documents non pertinents retrouvés par *DNPS*.
- l'ensemble des documents pertinents non retrouvés par *DPNS*.
- l'ensemble des documents non pertinents non retrouvés par *DNPS*.

La figure 1.3 résume cette distribution de documents dans la collection. En supposant connue cette distribution, plusieurs métriques d'évaluation des SRI ont été proposés, dont les principales sont résumées dans ce qui suit.

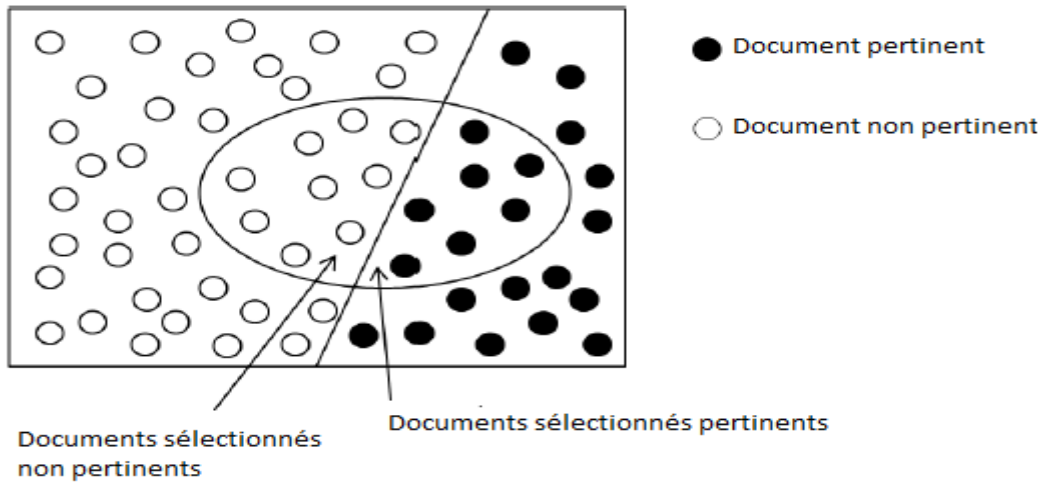


Figure 1.3 : Distribution des documents dans une collection face à une requête.

a) Rappel et Précision

Le *rappel* et la *précision* sont deux mesures de base pour évaluer la capacité d'un système à retrouver tous les documents pertinents de la collection et de rejeter tous les documents non pertinents pour la requête.

- Le *rappel* (noté *R*) représente le taux de documents pertinents sélectionnés parmi l'ensemble des documents pertinents. Il est défini par :

$$R = \frac{|DPS|}{|DP|} \qquad [1.23]$$

Un rappel R=1 indique la capacité du système *à sélectionner tous les documents pertinents.*

- La *précision* (notée *P*) représente le taux de documents pertinents sélectionnés.

$$P = \frac{|DPS|}{|DPS \cup DNPS|} \qquad [1.24]$$

Une précision P=1 indique la capacité du système *à ne sélectionner que les documents pertinents*.

L'idéal serait d'avoir une précision et un rappel égaux à 1, signifiant que tous les documents pertinents ont été retrouvés et qu'aucun document non pertinent n'a été sélectionné. En pratique, ceci n'est jamais atteint. En effet, comme indiqué dans la figure 1.4, ces deux mesures évoluent en sens inverse. Intuitivement, si nous augmentons le rappel pour retrouver plus de documents pertinents, la précision du système diminuera engendrant ainsi trop de documents non pertinents sélectionnés. Inversement, une plus grande précision risque de rejeter des documents pertinents diminuant ainsi le rappel.

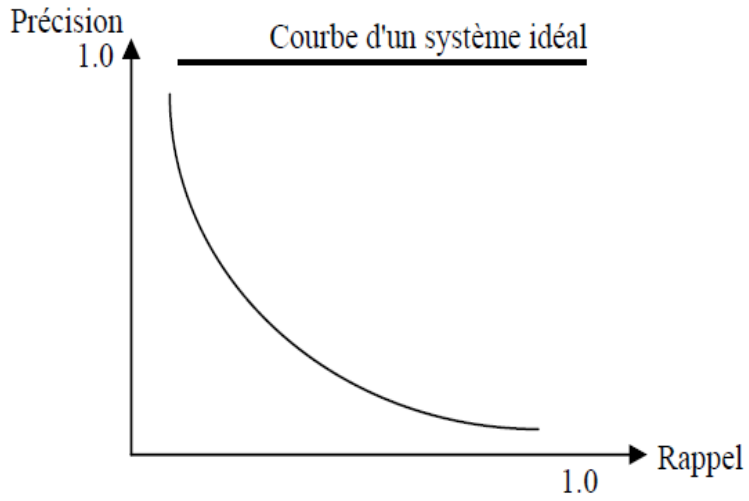


Figure 1.4 : Forme général de la courbe Précision-Rappel.

Les mesures complémentaires du rappel et de la précision sont respectivement : le *silence* et le *bruit.*

On parle du *silence* lorsque les documents pertinents ne sont pas sélectionnés par le système de recherche, alors qu'ils existent dans la collection. Les causes du silence peuvent être multiples. On peut citer par exemple l'imprécision du langage de la requête formulée par l'utilisateur ou une indexation classique basée sur les mots et non par leur sens, …etc. Le silence est calculé selon la formule suivante :

$$Silence \quad = 1 - R = \frac{|DPNS|}{|DP|} \qquad [1.25]$$

On parle de *bruit* lorsque des documents non pertinents sont proposés à l'utilisateur par le système de recherche. Ceci est causé par l'ambiguïté des termes de la requête. Le bruit est défini par la formule suivante :

$$Bruit \quad = 1 - P = \frac{|DNPS|}{|DPS \cup DNPS|} \qquad [1.26]$$

b) Précision à *x* documents

L'efficacité d'un système de recherche d'information est évaluée non seulement par le nombre de documents pertinents retrouvés mais aussi par le nombre de *documents pertinents sélectionnés dans les premiers rangs*. Pour cela, la précision au rang *x* (notée *P@x*) est utilisée. P@x consiste à mesurer la proportion des documents pertinents sélectionnés parmi les *x* premiers documents restitués par le SRI. Elle est définie par :

$$P@x = \frac{|DPS|}{|x|} \qquad [1.27]$$

En général, cette mesure est utilisée dans le protocole TREC, où *x* prend ses valeurs dans l'ensemble {1, 2, 3, 4, 5, 10, 15, 20, 30, 50, 100, 200, 500, 1000}.

c) Précision moyenne (*Mean Average Precision : MAP*)

L'évaluation des performances d'un modèle de RI ou un système de recherche SRI est réalisée sur la base des jugements des résultats de la recherche pour un ensemble de requêtes et non par ceux obtenus pour une seule requête. Pour ce faire, des précisions sont calculées à chaque document pertinent retrouvé en réponse à chaque requête donnée *q*. La moyenne de ces précisions représente la précision moyenne (*AP* : *Average precision*) de la requête *q*, qui désigne sa performance. La *MAP (Mean Average Precision*) ou moyenne arithmétique des précisions moyennes AP_q de toutes les requêtes *q*, permet de mesurer les performances du modèle de RI utilisé. Cette mesure est formellemnt définie par :

$$MAP \quad = \frac{1}{|Q|} \sum_{q=1}^{Q} AP_q \qquad [1.28]$$

Où :

- Q est l'ensemble des requêtes utilisées dans la recherche,
- $|Q|$ est le nombre de requêtes,

- AP_q est la moyenne des précisions calculées aux différents rangs k où un document pertinent est restitué par le système pour une requête q.

d) R-Precision

Etant donnée une requête Q. Et soit R, le nombre de documents de la collection pertinents pour Q. La R-Precision mesure la précision au rang R ($P@R$). Elle désigne la proportion des documents pertinents retrouvés parmi les R premiers documents sélectionnés par le système. Elle est donnée par :

$$\text{R - Precision} = \frac{|DPS|}{R} = \frac{|DPS|}{|DP|} \qquad [1.29]$$

D'autres mesures sont utilisées dans l'évaluation des performances d'un SRI, dont un état de l'art est dressé dans [Boughanem et al., 08 ; Croft et al., 09].

1.5 Conclusion

Dans ce chapitre, nous avons présenté les notions de base de la RI classique. En particulier, nous avons exposé les différentes étapes d'un processus de recherche d'information, et explicité les bases de l'indexation, de l'appariement et des modèles de recherche, ainsi que la reformulation de requête et les approches d'évaluation des SRI. La RI classique est basée sur l'indexation par des mots-clés et l'appariement lexical. Ces techniques présentent des insuffisances du fait de l'utilisation d'entités lexicales, les mots-clés, qui sont par nature ambigües. Pour pallier ces insuffisances, plusieurs approches d'indexation ont été proposés dans l'objectif principal est de représenter les documents et les requêtes par des concepts (sens des mots) qui sont par nature non ambigus. C'est l'objet de l'indexation sémantique que nous présentons dans le chapitre suivant.

Chapitre 2

Indexation sémantique

Plan du chapitre

2.1 Introduction 30

2.2 Problématique 30

2.3 L'indexation sémantique 32

2.3.1 Les ressources linguistiques 33

2.3.2 Les approches de désambiguïsation des sens des mots (WSD) 34

2.3.3 Les approches d'indexation sémantique basée sur la désambiguïsation 46

2.4 Conclusion 54

2.1 Introduction

L'indexation est une étape primordiale dans tout processus de recherche d'information. Sa qualité dépend de sa capacité à mieux représenter l'information portée par le contenu d'un corpus textuel. Dans les systèmes de recherche d'information classiques, l'indexation d'un document (ou d'une requête donnée), est réalisée par l'ensemble des mots-clés qu'il contient. Cette représentation, appelée aussi indexation classique, ne prend pas en considération la sémantique des mots pour décrire avec précision la thématique abordée dans le document (ou respectivement dans la requête), ce qui implique des résultats non pertinents lors de la recherche. C'est ainsi que les travaux récents en recherche d'information proposent l'indexation sémantique, en se basant sur la représentation des documents (et requêtes) par les sens des mots (ou concepts), plutôt que par les mots eux-mêmes.

Dans ce chapitre, nous présentons l'état de l'art sur l'indexation sémantique en RI, en s'inspirant de celui présenté dans [Boubekeur, 08], auquel nous ajoutons d'autres travaux les plus récents en indexation sémantique. En section 2.2, nous abordons la problématique de l'indexation classique basée mots-clés, en introduisant comme solution à ce problème l'indexation sémantique basée sur la désambiguïsation. Enfin, la section 2.3 sera dédiée à la présentation des principales techniques de désambiguïsation sémantique et aux différentes approches d'indexation sémantique en RI, basée sur la désambiguïsation.

2.2 Problématique

En indexation classique, la représentation du contenu d'un document, ou d'une requête utilisateur, par des mots-clés est généralement imprécise. Cette imprécision est causée par deux principaux problèmes [Boubekeur et al., 10a ; 10b] : l'ambiguïté des mots de la langue et leur disparité lors de la recherche.

- *L'ambiguïté des mots*, dite *ambiguïté lexicale*, se rapporte à des mots de la langue qui possèdent plusieurs sens. Elle est généralement divisée en deux types [Krovetz et al., 92 ; 97] : ambiguïté syntaxique et ambiguïté sémantique.

  - *L'ambiguïté syntaxique* se traduit par un mot qui apparait dans plusieurs formes grammaticales. A titre d'exemple, le mot *watch* dans la phrase 'I *watch* my *watch*' apparait d'abord comme verbe, puis comme nom. Il est clair que ces deux occurrences n'ont pas le même sens.

  - *L'ambiguïté sémantique* se rapporte à l'*homonymie* des mots ou la *polysémie* des mots. L'homonymie caractérise des concepts distincts qui sont lexicalement représentés par un même mot. La polysémie caractérise un mot qui désigne plusieurs sens. « 'Playing the *bass* guitar' *vs* 'Enjoy *bass* fishing' » est un exemple d'homonymie, et « '*bright* student' *vs* '*bright* room'» est un exemple de la polysémie.

Dans les SRI classiques, l'*ambiguïté lexicale* entraine un bruit documentaire, réduisant ainsi la précision des résultats de la recherche [Egozi et al., 11]. En effet, des documents

pourtant non pertinents contenant les mêmes mots de la requête (avec des sens différents), seront présentés à l'utilisateur par le système de recherche.

➢ *La disparité des mots (word mismatch problem)* se réfère à des mots qui sont différents lexicalement mais portant des sens liés (par exemple : cas de synonymie ou la relation hyperonymie/hyponymie). Dans un SRI classique, la disparité des termes entraine un silence documentaire. En effet, des documents pourtant pertinents, ne partageant aucun mot avec la requête, ne seront pas restitués par le processus de recherche réduisant ainsi le taux de rappel du système [Egozi et al., 11].

Pour pallier à ces problèmes, plusieurs travaux en recherche d'information ont tenté d'intégrer la sémantique des mots dans la représentation des documents et requêtes, suivant deux orientations :

- La première consiste à utiliser des expressions complexes (ou termes composés) en plus des mots-clés simples dans l'indexation des documents et requêtes [Alvarez et al., 04 ; Jiang et al.,04 ; Hammache, 09]. L'intérêt d'exploiter ces expressions, sémantiquement plus riches et plus précises, comme unités d'indexation, permet de réduire l'ambiguïté des mots et d'améliorer la précision des représentations des documents et requêtes par rapport à leurs représentations basées uniquement sur des mots-clés simples. A titre d'exemple, le terme composé '*fish oil*' est plus précis et plus riche sémantiquement que les mots simples '*fish*' et 'oil' pris isolément. Néanmoins, il est difficile de repérer tous les groupes de mots dans un document (ou une requête), de plus ces expressions ne correspondent pas forcément au langage utilisé par l'utilisateur dans sa requête.

- La seconde consiste à représenter les documents et les requêtes par les sens des mots (ou concepts) plutôt que par les mots eux-mêmes. On parle alors d'une indexation sémantique (ou conceptuelle). Les approches de cette orientation utilisent des index sémantiques qui sont construits à partir : (1) de la sémantique latente du texte donné (indexation sémantique latente), (2) des sens associés aux mots dans leurs contextes d'apparition (indexation sémantique), ou (3) des concepts qui sont extraits du contenu textuel des documents et requêtes (indexation conceptuelle).

  - L'indexation par la sémantique latente (*LSI -Latent Semantic Indexing-*) [Deerwester et al, 90] repose sur le *clustering* des mots co-occurrents, sémantiquement proches, en réduisant l'espace de la matrice termes-documents par une décomposition en valeur singulière. Cette technique permet de sélectionner des documents pertinents même s'ils ne contiennent aucun terme de la requête.

  - L'indexation sémantique se base sur la représentation des documents et requêtes par les sens des mots qu'ils contiennent. Le sens d'un mot est identifié par une approche de désambiguïsation des sens des mots (*WSD -Word Sense Disambiguation-*). Les approches de WSD sont principalement classées en deux catégories: les approches basées sur des corpus d'apprentissage, et les approches basées sur les ressources linguistiques externes (telles que les dictionnaires informatisés, les thésaurus et les ontologies). A ces approches de désambiguïsation, on associe deux approches d'indexation sémantique: l'indexation

sémantique basée sur des corpus d'apprentissage [Schütze et al., 95] et l'indexation sémantique basée sur des ressources linguistiques [Voorhees, 93 ; Uzener et al., 98 ; Baziz et al., 05a ; Boubekeur et al., 10a ; 10b].

- L'indexation conceptuelle est née des travaux de Woods [Woods, 97] qui a été le premier à proposer ce concept qui se réfère alors à la construction de taxonomies à partir de textes. Dans des travaux plus récents, le terme « indexation conceptuelle » a été utilisé pour désigner, de manière plus générale, une indexation à base de concepts issus de terminologies. A titre d'exemple, dans le domaine biomédical, des outils d'extraction de concepts, tels que MetaMap [Aronson, 01] ou MaxMatcher [Zhou et al., 06], ont été proposés s'appuyant sur les thesaurus médicaux MeSh[10] ou UMLS[11]. Les concepts identifiés à partir de ces outils sont exploités pour indexer des documents biomédicaux [Maisonnasse et al., 09, Dinh, 12].

Dans le cadre de notre travail, nous nous intéressons plus particulièrement à l'indexation sémantique basée sur la désambiguïsation, pour proposer une solution aux problèmes posés par l'indexation classique. Notre objectif principal est d'apporter une amélioration dans la représentation des documents et requêtes en se basant sur les sens des mots définissant des concepts dans une ressource linguistique.

2.3 L'indexation sémantique

L'indexation sémantique en RI est née du problème de l'ambiguïté des mots de la langue naturelle utilisés classiquement pour représenter les documents et requêtes. L'indexation sémantique a pour objet de représenter les documents et requêtes par les sens des mots permettant ainsi de lever toute ambiguïté, ce qui a pour conséquence d'améliorer les résultats de la recherche.

Pour retrouver les sens corrects des mots dans un document (ou dans une requête donnée), l'indexation sémantique requiert des techniques de désambiguïsation des sens des mots. Ces techniques se basent sur l'utilisation de ressources, telles que les corpus d'apprentissage ou les ressources terminologiques.

Nous décrivons, dans ce qui suit, les ressources linguistiques externes les plus exploitées par la désambiguïsation des sens des mots, puis nous présentons les principes de base de la désambiguïsation des sens des mots, ainsi que les travaux les plus significatifs dans le domaine. Les approches d'indexation qui sont basées sur les techniques de désambiguïsation linguistique des mots seront aussi présentées.

[10] http://www.nlm.nih.gov/mesh/
[11] http://www.nlm.nih.gov/research/umls/

2.3.1 Les ressources linguistiques

Les ressources externes exploitées en WSD sont classées en deux catégories [Navigli, 09] : *ressources structurées* et *ressources non structurées.*

- *Les ressources structurées* sont principalement les ressources linguistiques suivantes :
  - les dictionnaires informatisés (*Machine Readable dictionaries -MRDs-*) : qui représentaient des sources de connaissances très populaires dans les années 80 pour les différentes disciplines du domaine du traitement automatique de la langue. Dans un dictionnaire informatisé, un mot de la langue possède un ou plusieurs sens qui sont définis par leurs glossaires (*gloss*). Le glossaire d'un sens décrit le sens du mot par une définition, des commentaires et/ou des exemples d'utilisation courante. Comme exemples de dictionnaires informatisés, on peut citer : le *Collins English Dictionary* (*CED*) [Sinclair, 95], le *Oxford Dictionary of English* (*ODE*) [Soanes et al., 03] et le *Longman Dictionary of Contemporary English* (*LDOCE*) [Proctor, 78].
  - les thésaurus : sont des vocabulaires contrôlés permettant de définir les termes et les relations sémantiques entre termes (*Hyperonymie/Hyponymie, Synonymie, Antonymie, …etc.*). Les termes dans un thésaurus sont organisés dans des catégories de domaines décrivant leurs domaines d'usage. Parmi les thésaurus existants on cite : le thésaurus linguistique le *Roget (Roget's International Thesaurus*) [Roget, 95], le thésaurus médical *MeSh (Medical Subject Headings*) et le méta-thésaurus médical *UMLS (Unified Medical Language System*).
  - les ontologies : sont des spécifications conceptuelles d'un domaine. L'ontologie définit et organise l'ensemble des concepts du domaine et les relations sémantiques entre eux (taxonomie, synonymie, …). A titre d'exemple, *WordNet* [Miller, 95] et ses extensions (*WordNetDomains* [Magnini et al.,00], *EuroWordNet* [Vossen, 98]), l'ontologie *Sensus* [Knight et al., 94], la *Gene Ontology* (*GO*)[12] …etc.

Des exemples de ressources structurées les plus exploitées en RI sont données en annexe.

- *Les ressources non structurées* représentent :
  - les corpus d'apprentissage : Ce sont de longs textes utilisés dans les techniques d'apprentissage pour construire la connaissance nécessaire pour la désambiguïsation des sens des mots (WSD). Ces corpus peuvent être étiquetés manuellement avec les sens des mots. A titre d'exemple, le corpus *SemCor* [Miller et al. 93] est la version étiquetée du corpus *Brown* [Kucera et al., 67] avec des sens issus de WordNet.
  - les corpus de collocations : Ce sont des ensembles de collocations de mots qui ont une tendance de se produire ensemble régulièrement. Parmi ces ressources, nous

[12] http://www.geneontology.org/

citons : *The British National Corpus collocations*[13] et le *Collins Cobuild Corpus Concordance*[14].

2.3.2 Les approches de désambiguïsation des sens des mots (WSD)

L'objectif principal de la WSD est de retrouver les sens corrects des mots dans leur contexte d'utilisation. De nombreuses approches de désambiguïsation sémantique des mots existent. Ces approches peuvent être divisées en : *approches basées sur les ressources linguistiques externes* et *approches basées sur les corpus d'apprentissage*. Les approches basées sur les ressources linguistiques utilisent les ressources terminologiques structurées comme sources d'évidence pour la définition des sens des mots. Les approches de désambiguïsation basées sur les corpus d'apprentissage s'appuient plutôt sur de gros textes pour construire la connaissance nécessaire pour cela.

2.3.2.1 Les approches basées sur les ressources linguistiques

Ces approches se basent sur les dictionnaires informatisés, les thésaurus ou les ontologies pour désambiguïser un mot ambigu.

2.3.2.1.2 Les approches basées sur les dictionnaires informatisés

Lesk [Lesk, 86] a développé l'un des premiers systèmes de désambiguïsation basé sur un MRD. Dans ce système, les sens possibles d'un mot cible sont d'abord identifiés à partir d'un dictionnaire. Un score est ensuite associé à chacun de ces sens. Le score associé à un sens possible S_i correspond au nombre de mots communs entre la définition (ou *gloss*) de S_i dans le dictionnaire, et les définitions des sens S_j des autres mots co-occurrents dans son contexte (le contexte est ici défini comme une fenêtre de dix mots autour du mot cible). Le sens qui maximise ce score est retenu comme sens correct du mot cible. Formellement :

$$Score_{Lesk}(S_i) = \arg\max_i \sum_{i \neq j} \left| gloss(S_i) \cap gloss(S_j) \right| \qquad [2.1]$$

Où:

- $gloss(S_i)$ est l'ensemble des mots significatifs non vides, appartenant au glossaire du sens S_i associé au mot cible m_i.
- $gloss(S_j)$ est l'ensemble des mots significatifs du glossaire de S_j associé à un mot m_j appartenant au contexte de m_i.

Ce système a été testé par Lesk sur de petits échantillons du roman *Pride and Prejudice* et des annonces d'*Associated Press*, en s'appuyant sur quatre dictionnaires (*MRDs*[15]) différents. Sa précision de désambiguïsation rapportée était de 50% à 70% des cas. Cependant, tel que

[13] http://www.natcorp.ox.ac.uk.
[14] http://www.collins.co.uk/Corpus/CorpusSearch.aspx
[15] Lesk a testé son désambiguïseur avec quatre MRDs: Oxford Advanced Leaner's Dictionnary of current English (OALDCE), Merriam-Webster 7th (W7), Collins English Dictionnary (CED) et Oxford English Dictionnary (OED).

cité dans [Véronis et al., 90], ce système dépend fortement des mots utilisés dans les définitions et le rend très sensible à la présence ou l'absence d'un mot au sein des gloss, produisant ainsi des erreurs de désambiguïsation dans certains cas. En effet, si plusieurs sens d'un mot ambigu possèdent dans leur définition le même nombre de mots communs avec ceux des mots co-occurrents dans son contexte, il devient difficile de choisir parmi ces sens le sens qui le correspond. A titre d'exemple, dans le dictionnaire informatisé *Collins English Dictionnary* (CED), il a été rapporté dans [Véronis et al., 90], que le mot '*pen*' possède six sens ayant tous un seul mot en commun (le mot *write*) avec les sens du mot *page* (figure 2.1). De ce fait, la méthode de Lesk serait incapable de sélectionner parmi les sens de '*pen*' celui qui est proche au contexte '*page*'.

- **Pen 1.** an implement for *writing* or drawing using ink, formerly consisting of a sharpened and split quill, and now of a metal nib attached to a holder. **2.** the *writing* end of such an implement; nib. **3.** style of *writing*. **4. the pen. a).** *writing* as an occupation, **b).** the *written* word. **5.** the long horny internal shell of a squid. **6.** to *write* or compose.
- **Page 1.** one side of one of the leaves of a book, newspaper, letter, etc. or the *written* or printed matter it bears. **2.** such a leaf considered as a unit. **3.** an episode, phase, or period **4.** Printing. the type as set up for printing a page. **6.** to look through (a book, report, etc.); leaf through.

Figure 2.1 : Définitions des mots *pen* et *page* dans le dictionnaire informatisé *Collins English Dictionnary* (*CED*).

Par ailleurs, les relations implicites entre mots dans les définitions ne peuvent pas être repérées avec cette technique de désambiguïsation. Ainsi, dans l'exemple ci-dessus, bien que les termes '*ink*'et '*drawing*' du glossaire du premier sens de '*pen*' sont sémantiquement liés aux termes '*book*', '*newspaper*', '*letter*' et '*print*' appartenant au glossaire du premier sens de '*page*', la méthode de Lesk ne les prend pas en considération dans son processus de désambiguïsation et repose sur un simple comptage de mots lexicalement identiques dans ces gloss pour retrouver le sens approprié du mot '*pen*'.

Malgré ces inconvénients, la méthode de Lesk a tout de même servi de base pour la plupart des travaux postérieurs en désambiguïsation basée sur les dictionnaires informatisés.

Wilks et al. [Wilks et al., 90] ont repris le principe de Lesk, en proposant d'étendre le contexte local d'un mot ambigu et ses entrées-sens[16], pour obtenir un maximum de recouvrement de mots entre eux. La technique d'expansion utilisée consiste à enrichir manuellement le contexte d'un mot et les définitions de ses sens avec des mots qui co-occurrent fréquemment avec les mots de leurs textes respectifs (contexte et sens). Le contexte et les différentes entrées-sens du mot cible élargis sont ensuite représentés par des vecteurs. Le sens dont le vecteur associé a le plus grand degré de corrélation avec le vecteur contexte est retenu comme sens correct du mot ambigu.

16 Une entrée-sens d'un mot : représente la définition de son sens dans un dictionnaire.

Wilks et al. ont testé leur désambiguïseur sur le mot '*bank*' qui possède 13 sens dans le dictionnaire LDOCE. Les résultats rapportés ont montré une performance de 45% par rapport à une désambiguïsation manuelle. Cependant, cette méthode ne permet pas de désambiguïser simultanément plusieurs mots ambigus qui se trouvent dans une phrase donnée au risque d'une explosion combinatoire du traitement [Véronis et al., 90 ; Cowie et al., 92].

Pour résoudre ce problème, Véronis et al. [Véronis et al., 90] suggèrent une approche de désambiguïsation lexicale qui étend la méthode de Lesk, en construisant un réseau de neurones à partir des définitions du dictionnaire anglais CED (*Collins English Dictionnary*). Dans ce réseau, chaque entrée d'un mot dans le dictionnaire est représentée par un nœud, le nœud mot. Ce nœud est relié, par des *liens d'activation*, à d'autres nœuds, les nœuds sens, qui représentent les différents sens du mot dans le CED. De même, chaque nœud sens est relié à son tour aux nœuds mots qui apparaissent dans sa définition et qui correspondent à des entrées dans le dictionnaire. Ce processus est répété plusieurs fois, en générant ainsi un réseau interconnecté de plus en plus complexe. Pour lever l'ambiguïté des mots dans une phrase donnée, les nœuds mots du réseau sont activés en premier, puis chaque nœud mot envoie une activation à ses nœuds sens, qui à leur tour activent les nœuds voisins auxquels ils sont connectés, … et ainsi de suite tout au long du réseau, pour un certain nombre de cycles. Finalement, le réseau se stabilise progressivement suivant la stratégie '*winner-take-all*', pour obtenir un état où un seul sens de chaque mot d'entrée est plus activé que les autres. Ces sens correspondent aux sens adéquats des mots dans la phrase analysée. Véronis et al., expérimentent leur approche sur 23 mots ambigus dans six contextes différents. Les résultats montrent un taux de précision de 71,7% des cas.

De nombreux travaux qui ont suivis [Guthrie et al., 91 ; Cowie et al., 92 ; Liddy et al., 93 ; … etc], ont tenté d'apporter une amélioration de la désambiguïsation, en intégrant des informations supplémentaires, issus du dictionnaire anglais LDOCE (*Longman Dictionnary of Contemporary English*), dans leur processus de désambiguïsation.

Guthrie et al. [Guthrie et al., 91] exploitent les catégories de sujets[17] associés aux sens dans le dictionnaire LDOCE, en s'appuyant sur l'approche de désambiguïsation de Wilks et al. [Wilks et al., 90]. A la différence de cette dernière, la définition d'un sens d'un mot caractérisée par une catégorie spécifique, est étendue seulement par l'ensemble des mots co-occurrents dans toutes les définitions assignées dans cette même catégorie dans le LDOCE. Cependant, aucun test n'a été rapporté pour ce désambiguïseur.

Cowie et al. [Cowie et al., 92] proposent d'enrichir chaque définition d'un sens d'un mot ambigu par son code de domaine (ou catégorie de sujet), qui lui est attribué dans le LDOCE. Ce code est traité dans cette méthode comme un mot faisant partie de la définition. La désambiguïsation d'un mot repose alors sur le nombre de mots et de codes de domaine en commun entre les définitions des sens de ce mot et celles des sens des autres mots appartenant à son contexte. Le sens ayant le plus grand nombre est retenu comme sens de ce mot dans son

[17] LDOCE comporte 124 catégories de sujets majeurs, par exemple : *economics*, *engineering*, …*etc.*

contexte d'apparition. Les résultats d'évaluation de cette approche ont présenté une bonne précision dans 47% des cas.

Bien que les dictionnaires informatisés aient été exploités largement dans la désambiguïsation des mots, leur qualité est souvent remise en question. En effet, ces dictionnaires, développés pour usage humain, ne donnent pas une description suffisante des liens sémantiques entre mots qui peuvent être utilisés dans un cadre automatique. Ils présentent en outre, une divergence au niveau de la représentation des sens lexicaux et des relations entretenues entre eux [Apidianaki, 08].

2.3.2.1.2 Les approches basées sur un thésaurus

Ces approches reposent sur le vocabulaire contrôlé d'un thésaurus comme source de connaissances pour déterminer les sens exacts des mots dans leur contexte d'apparition. Le vocabulaire d'un thésaurus fournit une description sémantique des associations entre mots et classent les sens des mots liés sémantiquement dans des catégories sémantiques (catégories de domaines).

Yarowsky [Yarowsky, 92] se base sur les catégories sémantiques[18] du thésaurus Roget [Roget, 95], pour retrouver les sens adéquats des différents mots qui apparaissent dans l'encyclopédie *Grolier multimédia* [Grolier]. Il propose une approche de désambiguïsation à deux étapes : la première consiste à retrouver parmi les catégories sémantiques du Roget, celle qui correspond au mot à désambiguïser dans son contexte d'utilisation. La seconde permet alors d'assigner le sens correct du mot ambigu dans la catégorie ainsi identifiée.

(1) Dans la première étape, chaque catégorie sémantique du thésaurus est associée à un ensemble de mots déterminants (*salient words*) permettent d'identifier la catégorie du mot cible. Ces mots sont repérés selon le principe suivant (extrait de [Boubekeur, 08]):

Pour une catégorie sémantique *C* du thésaurus,

- Sélectionner, à partir de l'encyclopédie, pour chaque mot m_i de *C* toutes ses occurrences ainsi que leurs contextes locaux. Le contexte local d'une occurrence m^j_i de m_i est défini par l'ensemble des 100 mots qui l'entourent : soit 50 mots à gauche et 50 mots à droite,
- Construire ensuite le contexte représentatif ψ_i de la catégorie *C* qui est déterminé par l'ensemble de tous les contextes des mots $m_{i\ (i=1,..)}$ liés à la catégorie *C*. Le contexte d'un mot m_i de *C* est l'union de tous les contextes locaux de ses occurrences $m_i^{j(j=1,...)}$. Formellement, ψ_i est défini par :

$$\psi_i = \bigcup_i \left(\bigcup_j \zeta_i^j \right) \qquad [2.2]$$

18 Le thésaurus Roget comporte 1024 catégories de domaines (telles que : ANIMAL/INSECT, TOOLS/MACHENERY, …etc) , qui couvrent les différents sens des mots.

- Pour chaque mot m_i de C, un score lui est associé basé sur le ratio de sa probabilité dans le contexte ψ_i de la catégorie C par rapport à sa probabilité globale dans l'encyclopédie Grolier. Les mots des scores les plus élevés représentent les mots déterminants pour la catégorie C.

A l'issue de cette étape, chaque catégorie dans le thésaurus Roget possède un ensemble de mots déterminants. La présence des mots déterminants d'une catégorie dans le contexte d'un mot ambigu révèle l'évidence de son appartenance à cette catégorie. Cependant, si plusieurs mots déterminants associés à plusieurs catégories apparaissent dans le contexte du mot cible, sa catégorie est celle qui maximise la somme des scores de ses mots déterminants.

(2) Dans la seconde étape, il s'agit de sélectionner le sens du mot dans la catégorie identifiée dans l'étape précédente.

Yarowsky a testé son approche sur 12 mots ambigus. Les résultats rapportés ont montré une précision de 92%.

Dans une approche différente, Mohammad et Hirst [Mohammad et al., 06] utilisent le thésaurus *Macquarie*[19] pour la désambiguïsation des sens des mots. L'idée de base de leur approche est que la majorité des occurrences d'un mot dans un corpus textuel, portent un même sens qui représente le sens prédominant du mot. Ce sens désigne son sens approprié dans le corpus.

Le principe de cette approche consiste, dans un premier temps, à construire une matrice M de cooccurrence mots/catégories en se basant sur les catégories sémantiques du thésaurus *Macquarie*. Un élément w_{ij} de la matrice M indique le nombre de fois où un mot m_i d'une catégorie donnée C_j, co-occurre avec les autres mots de la même catégorie issus des contextes locaux de ses occurrences. Le contexte local d'une occurrence de m_i dans le corpus textuel est défini par une fenêtre de 10 mots autour d'elle, soit 5 mots à gauche et 5 mots à droite. En appliquant des mesures statistiques (plusieurs mesures ont été expérimentées telles que : le coefficient de Dice, le cosinus, …) sur la matrice de co-occurrence M, un poids est calculé pour chaque mot traduisant son degré d'association dans le corpus avec les mots d'une catégorie sémantique donnée.

Mohammad et Hirst proposent de désambiguïser un mot en attribuant un score à chaque catégorie qui contient un de ses sens possibles. La catégorie qui maximise ce score est alors considérée comme catégorie dominante, ou sens prédominant, du mot cible dans le corpus. Pour calculer ce score, quatre mesures différentes ont été proposées :

- La première mesure repose sur l'hypothèse que la *catégorie dominante d'un mot cible est celle qui s'associe fortement (ou possède un plus grand degré d'association) avec les mots co-occurrents de ce mot dans l'ensemble de ses contextes locaux. Cette catégorie contient le sens prédominant du mot qui représente son sens adéquat dans le corpus.* Le score proposé est alors calculé sur la base de la somme des poids des mots co-occurrents dans la catégorie C du mot t, selon la formule suivante *:*

[19] http://www.macquariedictionary.com.au/anonymous@9c9B329512906/-/p/dict/index.html

$$S_{(t,C)} = \arg\max_{C} \frac{\sum_{w \in T} A(w,C)}{\sum_{C' \subset senses\ (t)} \sum_{w \in T} A(w,C')} \qquad [2.3]$$

Où :

- *S(t,C)* représente le sens prédominant du mot cible *t* appartenant à sa catégorie *C ;*
- *T* est l'ensemble des mots co-occurrents du mot *t* appartenant aux contextes locaux de ses occurrences dans le corpus *;*
- *A(w,C)* est le poids d'un mot co-occurrent *w* dans la catégorie *C* exprimant son degré d'association à cette catégorie, avec $w \in T$.

- La seconde mesure repose sur l'hypothèse que *la catégorie dominante d'un mot cible est celle qui possède un plus grand nombre de mots co-occurrents, appartenant à l'ensemble des contextes locaux de ses occurrences, qui s'associent fortement à elle (i.e à la catégorie*). Le score de désambiguïsation est alors défini par:

$$S_{(t,C)} = \arg\max_{C} \frac{\left| \left\{ w \in T : C = \arg\max_{C' \in senses\ (t)} A(w,C') \right\} \right|}{|T|} \qquad [2.4]$$

Où :

- $|T|$ est le nombre de mots co-occurrents du mot *t* dans ses contextes locaux ;
- $\{w \in T : C = \text{argmax } A(w,C')\}$ est l'ensemble des mots co-occurrents *w* appartenant à *T* qui possèdent les plus grand poids dans la catégorie *C*.

- La troisième mesure est une combinaison des deux mesures précédentes. Elle s'appuie sur l'idée que *le sens approprié de chaque occurrence d'un mot cible dans son contexte local appartient à la catégorie qui possède le plus grand cumul des degrés d'associations des mots concurrents, du contexte local de l'occurrence, avec cette catégorie.* Ainsi, chaque occurrence du mot est associée à une catégorie qui contient son sens approprié dans son contexte local. La catégorie de l'une des occurrences qui maximise le nombre d'apparitions dans l'ensemble des contextes locaux du mot cible, est considérée comme une catégorie dominante de ce mot dans le corpus définissant son sens prédominant. Le score de désambiguïsation est alors calculé selon la formule suivante :

$$S_{(t,C)} = \arg\max_{C} \frac{\left| \left\{ T' \in \lambda_t : C = \arg\max_{C' \in senses\ (t)} \sum_{w \in T'} A(w,C') \right\} \right|}{|\lambda_t|} \qquad [2.5]$$

Où :

- T' est le contexte local d'une occurrence du mot ambigu t, avec $|T'|$ est le nombre de mots co-occurrents de l'occurrence dans son contexte local ;
- λ_t est l'ensemble des contextes locaux de toutes les occurrences d'un mot t, avec $|\lambda_t|$ est le nombre de ses contextes qui est égal au nombre d'occurrences de t dans le corpus.

- La quatrième mesure proposée se base sur l'idée que *le sens correspondant à une occurrence d'un mot dans son contexte local est celui dont sa catégorie possède un plus grand nombre de mots co-occurrents qui s'associent fortement à elle.* Par conséquent, la catégorie dominante d'un mot cible dans le corpus est la catégorie de l'une de ses occurrences qui apparaît, dans l'ensemble de ses contextes locaux, plus souvent que les catégories associées aux autres occurrences. Le score est calculé selon la formule suivante :

$$S_{(t,C)} = \arg\max_{C} \frac{\left|\left\{T' \in \lambda_t : \arg\max_{C' \in senses\ (t)} \left|\{w \in T' : A(w,C)\}\right|\right\}\right|}{|\lambda_t|} \qquad [2.6]$$

Ces quatre mesures ont été évaluées dans la désambiguïsation d'un petit échantillon du corpus *British National Corpus World Edition (BNC)* [Burnard, 00]. Les résultats rapportés de ont montré que ces quatre scores de désambiguisation ont apporté de meilleures précisions (plus de 50%) comparativement à une désambiguïsation manuelle de l'échantillon. En outre, la mesure de désambiguïsation de la formule [2.5] est la plus performante que les autres mesures proposées.

Quoique l'utilisation des thésaurus dans les travaux de désambiguïsation des mots ait permis de réduire le problème de l'ambiguïté de la langue, en offrant une catégorisation sémantique des mots représentant leurs degrés d'associations, ils restent cependant une ressource de connaissances peu exploitée par les chercheurs à cause de leur formalisation qui s'appuie sur la notion de termes plutôt que celle de concepts [Mizoguchi, 04].

2.3.2.1.3 Les approches basées sur une ontologie

Ces approches se basent sur les concepts des mots dans l'ontologie ainsi que leurs relations sémantiques conceptuelles, telles que l'équivalence, la synonymie, la relation d'hyponymie/ hyperonymie, et … autres, pour définir le sens correspondant du mot dans son contexte d'utilisation. Parmi les ressources ontologiques les plus exploitées dans la désambiguïsation des mots de la langue on retrouve WordNet [Miller, 95 ; Fellbaum, 98] (une présentation de l'ontologie WordNet est donnée en annexe).

Sussna [Sussna, 93] défini une approche de désambiguïsation basée sur les liens sémantiques entre synsets[20] de noms dans l'ontologie WordNet. Il propose dans un premier temps d'assigner un poids à chaque relation sémantique entre deux synsets dans WordNet, exprimant la proximité sémantique entre eux. Il associe le poids le plus fort à la synonymie et le poids le plus faible à la relation d'antonymie. Pour désambiguïser un mot (nom) dans son contexte d'utilisation, cette approche affecte à chacun de ses synsets dans l'ontologie WordNet un score égal à la somme de ses distances sémantiques minimales avec les synsets des mots co-occurrents de son contexte. La distance sémantique minimale entre deux synsets est calculée par la somme des poids des relations sur le chemin le plus court entre eux. Le synset qui maximise ce score est alors retenu comme sens adéquat du mot dans son contexte. Pour évaluer la précision de cette approche, Sussna compare les résultats de son désambiguïseur sur la collection TIME par rapport à ceux obtenus par une désambiguïsation manuelle. Il obtient une précision de recherche de 56%.

Avec un principe similaire, d'autres approches [Resnik, 95 ; Voorhees, 93 ; Baziz, 05b ; Boubekeur et al., 10a ; 10b] ont proposé de désambiguïser un mot en s'appuyant sur des mesures de similarités sémantiques entres les synsets dans la taxonomie *is-a* des noms et verbes de WordNet. L'idée de base de ces approches est que le sens approprié d'un mot est défini par le synset le plus proche sémantiquement, dans la hiérarchie *is-a* de WordNet, aux synsets des mots co-occurrents de son contexte. Pour calculer la similarité sémantique entre deux synsets, exprimant leur proximité sémantique, plusieurs mesures ont été développées en se basant sur : (1) la notion de distance sémantique du plus court chemin, exprimée en fonction du nombre d'arcs [Rada et al., 89, Wu-Palmer, 94 ; Ehrig et al, 04 ; Slimani et al., 07] , (2) ou la notion du contenu informatif qui détermine le degré de l'information partagée entre les deux synsets [Lin, 98 ; Resnik, 99, Seco et al., 04], (3) ou par une combinaison entre les deux notions (hybride) : leur distance sémantique et leur contenu informatif [Jiang-Conrath, 97 ; Pirró et al, 10]. Un état de l'art sur les différentes mesures de similarités est donné par Budanistky [Budanitsky et al., 06] et Tchechmedjiev [Tchechmedjiev, 12].

Banerjee et Pedersen [Banerjee et Pedersen, 03] ont aussi proposé une approche de désambiguïsation basée sur une mesure de similarité sémantique différente, en adaptant le principe de désambiguïsation de Lesk [Lesk, 89]. Cette approche étend la mesure de Lesk [Lesk, 89] pour supporter les relations disponibles dans l'ontologie WordNet, en définissant pour chaque sens (synset) possible d'un mot ambigu un score basé sur le degré de recouvrement entre les mots de son contexte d'une part, et la définition du synset ainsi que les définitions des sens (synsets) issus des différentes relations (hyperonymes *has-kind*, hyponymes *kind-of*, meronymes *part-of*, …) avec ce synset d'autre part. Le synset qui maximise ce score est considéré comme le concept, ou sens, adéquat du mot dans son contexte. Ce score est calculé selon la formule suivante :

$$Score_{\ Lesk\ étendu}(S_w) = \sum_{S':S_w \longrightarrow S' ou\ S_w \equiv S'} \left| contexte\ (w) \cap gloss\ (S') \right| \qquad [2.7]$$

[20] Un synset est un ensemble de termes synonymes qui définit une entrée (ou concept) deWordNet.

Où :

- S_w est un synset du mot ambigu *w* et *S'*est le sens lié à S_w par l'une des relations définies dans WordNet (synonymie, hyperonymie, …etc).

D'autres travaux [Magnini et al., 02 ; Gliozzo et al. 04; Vázquez et al. 04, Buitelaar et al. 07 ; Kolte et al., 08] ont suggéré d'exploiter l'information du domaine pour désambiguïser un mot dans son contexte d'utilisation. Leur but est de pouvoir identifier avec précision le sens du mot qui le correspond dans son contexte. En s'appuyant sur le lexique *WordNet* et son extension aux domaines *WordNetDomains* [Magnini et al., 00], Gliozzo et al., [Gliozzo et al., 04] représentent l'ensemble des domaines associés à chaque synset d'un mot ambigu par un vecteur ***S*** appelé *Synset Vector*, défini par : ***S***= (R(D_1, *S*), R(D_2, *S*), R(D_3, *S*), …, R(D_d, *S*),) où $D_{i/i=\{1,..,d\}}$ représentent les domaines existants dans WordNetDomains, et R(D_i, *S*) est la probabilité que le syset S soit assigné au domaine D_i. Cette probabilité est déterminée par :

$$R(D_i, S) = \begin{cases} \dfrac{1}{|Dom\ (S)|} & \text{si } D_i \in Dom\ (S) \\ \dfrac{1}{d} & \text{si } Dom\ (S) = \{Factotom\} \\ 0 & \text{sinon} \end{cases} \qquad [2.8]$$

Où :

- *Dom(S)* est l'ensemble des domaines du synset *S* dans WordNetDomains.

Par analogie, un autre vecteur, nommé *Text Vector,* est calculé pour représenter les domaines des synsets des autres mots co-occurrents (du mot cible) dans son contexte. Pour retrouver le sens adéquat du mot, cette approche propose de retenir parmi ces synsets celui dont le vecteur synset (*Synset Vector*) maximise sa similarité avec le vecteur *texte (text Vector)*. Gliozzo et al., ont évalué la précision de leur désambiguïsation sur la collection Senseval-2[21]. Les résultats rapportés montrent que cette approche présente de meilleures précisions (79% en désambiguïsant uniquement les noms et verbes, et 75% en désambiguïsant tous les mots de différentes catégories syntaxiques), et de faibles rappels (40% pour les noms et verbes et 35% pour les tous les mots).

Vázquez et al. [Vázquez et al., 04] se basant sur le même principe que [Gliozzo, 04], proposent d'exploiter les domaines des mots appartenant aux définitions (ou glosses) de WordNet, dans le processus de désambiguïsation. L'approche consiste à représenter dans un *vecteur contexte* les domaines les plus représentatifs des mots appartenant au contexte du mot cible, et dans un autre vecteur, appelé *vecteur sens*, les domaines les plus représentatifs des mots appartenant à la définition de l'un de ces synsets dans WordNet. Ainsi, chaque synset du mot à désambiguïser est associé à un vecteur sens. Les domaines représentatifs dans le contexte (ou respectivement dans la définition d'un synset), sont retrouvés sur la base d'une

[21] http://www.senseval.org/data.html

mesure d'information mutuelle entre le domaine de chaque mot dans le contexte (ou dans la définition) avec les domaines associés aux mots qui co-occurrent avec lui. Pour sélectionner le bon synset d'un mot dans son contexte, Vázquez et al. proposent un score pour chacun de ses vecteurs sens, sur la base de sa similarité avec le vecteur contexte. Le synset ayant le plus grand score est considéré comme sens correct du mot ambigu. L'évaluation de ce désambiguïseur a apporté une précision de 47% dans la désambiguïsation de la collection Senseval-2.

Dans une autre approche intéressante, Kolte et al., [Kolte et al., 08 ; Kolte et al., 09] proposent une désambiguïsation à deux niveaux : d'abord désambiguïser le domaine correct d'un mot dans son contexte local (i.e la phrase où il apparait), puis désambiguïser ce mot dans le domaine ainsi identifié. Le domaine correct d'un mot est celui qui maximise ses occurrences dans le contexte du mot. Pour identifier le sens du mot dans le domaine choisi, Kolte et al. proposent d'exploiter les relations sémantiques entre synsets dans les taxonomies de WordNet (*Hypernym, Meronymy/Holonymy*, …). Ce désambiguïseur a été testé sur le corpus annoté *SemCor*. Les résultats rapportés présentent une précision de 63,92%.

2.3.2.2 Les approches basées sur les corpus d'apprentissage

Ces approches se basent sur l'utilisation d'un corpus composé d'un grand nombre de contextes de mots polysémiques, dans le but d'apprendre les connaissances utiles sur le sens d'usage des mots. Cette phase d'identification automatique des connaissances est appelée apprentissage. A l'issue de cette phase, l'algorithme de désambiguïsation est capable d'assigner le sens adéquat aux mots apparaissant dans une nouvelle phrase en s'appuyant sur les connaissances acquises durant la phase d'apprentissage.

Les approches de désambiguïsation basées sur les corpus d'apprentissage se distinguent en approches supervisées et approches non supervisées.

2.3.2.2.1 Les approches supervisées

Dans ces approches, les connaissances nécessaires pour la désambiguïsation sont construites à partir d'un texte manuellement annoté avec les sens des mots.

Weiss [Weiss ,73] est l'un des premiers à s'intéresser à l'apprentissage des règles de désambiguïsation à partir d'un corpus étiqueté. En analysant les 20 phrases contenant des occurrences de cinq mots ambigus dans le corpus étiqueté ADI[22], il construit manuellement deux types de règles : des règles générales de contexte et des règles de modèle.

- Une règle générale de contexte établit la relation structurelle entre une occurrence d'un mot ambigu avec les mots co-occurrents dans son contexte, permettant ainsi de définir son sens. Par exemple, si le mot *print* apparait au voisinage du mot '*type*' alors son sens est probablement lié à l'impression.
- Une règle de modèle établit la relation de position d'un mot co-occurrent par rapport une occurrence d'un mot ambigu, afin de pouvoir définir son sens. Par exemple, si le

[22] http://ir.dcs.gla.ac.uk/resources/test_collections/adi/

mot *of* apparait juste après le mot *type* alors le sens de cette occurrence est probablement *variety of*.

Ces règles ainsi construites ont été testées par un désambiguïseur automatique sur les 30 phrases restantes associées aux occurrences des cinq mots ambigus dans le corpus ADI. La précision résultante est de l'ordre de 90%.

Dans une approche similaire, Kelly et al., [Kelly et al., 75] ont manuellement créé un ensemble de règles pour 1815 mots possédant une fréquence de plus de 20 occurrences dans un corpus étiqueté. Ces règles sont constituées à partir des règles contextuelles identiques à celles proposées par Weiss et d'autres règles grammaticales construites en analysant la morphologie de chaque occurrence d'un mot ambigu et sa catégorie syntaxique dans un contexte donné du corpus. Les règles ainsi crées présentent de meilleurs indicateurs pour identifier le sens d'un mot se trouvant dans un contexte. A la différence du système de Weiss, ce désambiguïseur a été conçu pour traiter une phrase entière en même temps. Le système n'a cependant pas eu de succès, et Kelly et al. ont rapporté que cette technique ne peut pas réussir à échelle réelle.

Dans une autre approche de désambiguïsation, Yarowsky [Yarowsky, 00] repose sur des connaissances basées sur un arbre de listes de décision hiérarchiques pour identifier le sens adéquat d'une occurrence d'un mot ambigu. Chaque liste est établie à partir d'une classification de succession de règles conditionnelles ordonnées de type *if….then*, en observant les différentes caractéristiques des occurrences d'un mot dans une partie du corpus annoté SENSEVAL. Ces caractéristiques sont traduites par des attributs (sa localisation, son type (lemme, catégorie syntaxique,…) et la valeur du token) de patterns permettant de déclencher les règles de décision pour obtenir le sens attendu du mot ambigu ou se pointer vers une autre liste de décision. Dans une liste, les règles sont ordonnées par le ratio de log-vraisemblance.

Lors de la désambiguïsation, le sens correct d'une occurrence d'un mot ambigu est celui associé à la première règle de la liste de décision s'appliquant à cette occurrence. Si aucune règle ne s'applique alors le sens le plus fréquent lui est assigné.

A titre d'exemple, une occurrence du mot *promise* est désambiguïsée à partir de l'arbre de décision présenté dans la figure 2.2. De cette figure, on peut voir que, si les caractéristiques de cette occurrence correspondent aux attributs du pattern {Loc = +0, Typ=P, Token=Noun}, exprimant que l'occurrence est de catégorie nom, alors un pointeur pointe sur la liste *promise.LN*, qui elle-même peut pointer sur la liste *promise.L4* si le le token *promise* est un mot littéral. Si le mot qui suit *promise* est *to* alors le sens attendu est son quatrième sens.

Le système de désambiguïsation basée sur cette approche a été classé en tête de tous les systèmes supervisés présentés lors de la campagne d'évaluation SENSEVAL de 1998, avec une précision de 78,9%.

Top-level Decision List for **promise**

Pattern Loc	Typ	Token		Next List	Empirical Sense Distribution 1	3	4	4.1	4.2	4.3	4.4	5	6
+0	P	NOUN	→	$LN_{(\Downarrow)}$	0	0	297	53	5	37	11	22	93
+0	P	VERB	→	LV	440	115	0	0	0	0	0	0	0

⇓

Mid-level Decision List for **promise.LN** (noun)

Pattern Loc	Typ	Token		Next List	Empirical Sense Distribution 4	4.1	4.2	4.3	4.4	5	6
V/obj	L	keep/V	→	**4.3**	0	0	0	31	0	0	0
V/obj	L	break/V	→	**4.4**	0	0	0	0	11	0	0
V/obj	L	make/V	→	L1	2	44	0	0	0	0	2
V/obj	L	give/V	→	L2	0	0	5	1	0	1	2
+0	W	promises	→	L3	115	5	0	0	0	0	1
+0	W	promise	→	$L4_{(\Downarrow)}$	180	3	0	1	0	21	88

⇓

(Abbreviated) Terminal Decision List for **promise.L4** (promise-noun-singular)

Pattern Loc	Typ	Token		Output Sense	LogL	Empirical Sense Distribution 4	4.1	4.2	4.3	4.4	5	6
+1	W	to	→	**4**	9.51	41	0	0	0	0	0	0
-1	W	of	→	**6**	8.16	0	0	0	0	0	0	12
-1	L	early/J	→	**6**	7.38	0	0	0	0	0	0	7
V/obj	L	show/V	→	**6**	7.27	0	0	0	0	0	0	13
+1	W	at	→	**6**	6.16	0	0	0	0	0	0	3
-1	L	firm/J	→	**4**	5.74	6	0	0	0	0	0	0
+1	L	do/V	→	**4**	5.70	3	0	0	0	0	0	0
-1	W	such	→	**6**	5.57	0	0	0	0	0	0	2
-1	W	much	→	**6**	5.57	0	0	0	0	0	0	2
+1	W	when	→	**6**	5.57	0	0	0	0	0	0	2
+1	W	on	→	**6**	5.57	0	0	0	0	0	0	2
+1	W	as	→	**6**	5.57	0	0	0	0	0	0	2
-1	W	your	→	**4**	5.16	2	0	0	0	0	0	0
+1	W	during	→	**4**	5.16	2	0	0	0	0	0	0
$\pm k$	L	free/J	→	**4**	4.74	15	0	0	0	0	0	0
V/obj	L	trust/V	→	**4**	4.74	3	0	0	0	0	0	0
$\pm k$	L	support/N	→	**4**	4.64	14	0	0	0	0	0	0
$\pm k$	L	election/N	→	**4**	4.29	11	0	0	0	0	0	0
subj/V	L	contain/V	→	**4**	4.18	2	0	0	0	0	0	0
V/obj	L	win/V	→	**4**	4.16	2	0	0	0	0	0	0
V/obj	L	repeat/V	→	**4**	4.16	2	0	0	0	0	0	0
V/obj	L	honour/V	→	**4**	4.16	2	0	0	0	0	0	0
-1	L	rhetorical/J	→	**5**	4.09	0	0	0	0	0	1	0
-1	L	increase/V	→	**5**	4.09	0	0	0	0	0	1	0
-1	L	future/J	→	**5**	4.09	0	0	0	0	0	1	0

Figure 2.2 : Extrait de l'arbre de listes décision hiérarchiques construite pour le mot *promise* dans le corpus annoté SENSEVAL [Yarowsky et al., 00].

2.3.2.2.2 Les approches non supervisées

Ces approches se basent sur des corpus non annotés pour construire la connaissance nécessaire à la désambiguïsation. L'apprentissage est basé sur l'idée que les occurrences d'un mot qui ont même sens possèdent souvent des mots co-occurrents similaires. Il s'agit alors de regrouper en clusters, les mots voisins des différentes occurrences similaires d'un mot dans une partie du corpus. Les ensembles des clusters obtenus représentent ainsi les sens possibles de ce mot, qui seront utilisés pour la désambiguïsation des autres occurrences se trouvant dans l'autre partie du corpus.

Schütze [Schütze, 98] propose une technique de désambiguïsation non supervisée basée sur le modèle vectoriel. Dans sa phase d'apprentissage, cette approche représente les contextes des différentes occurrences d'un mot par des vecteurs de mots co-occurrents dérivés à partir de leurs contenus. Le contexte d'une occurrence est défini par une fenêtre de cinquante mots autour de l'occurrence en question. Ces vecteurs contextes sont ensuite regroupés en clusters en fonction de leur degré de similitude calculé sur la base de la mesure de cosinus. Pour chaque cluster obtenu, Schütze calcule le centroïde (ou le barycentre) de l'ensemble de vecteurs dans ce cluster est lui affecte le sens du mot qu'il lui correspond. Par

conséquent, les ensembles de clusters caractérisés par leurs centroïdes décrivent les sens possibles (ou usages possibles) du mot. La désambiguïsation d'une nouvelle occurrence de ce mot est alors basée sur la distance entre le vecteur contexte construit pour cette occurrence et le centroïde de chaque cluster associé à ce mot dans la phase d'apprentissage. Le cluster ayant son centroïde le plus proche du vecteur contexte de l'occurrence est alors considéré comme sens adéquat de cette occurrence dans son contexte d'utilisation.

Dans une autre approche, Pantel et al. [Pantel et al., 02] proposent un algorithme de clusterisation dit *Clustering by Committee*. Cet algorithme consiste dans un premier temps à retrouver pour chaque occurrence m_i d'un mot *m* dans un corpus, ses *k mots voisins les plus proches*. Pour ce faire, les mots co-occurrents de m_i sont représentés par des vecteurs traits représentants leurs caractéristiques syntaxiques ou sémantiques. Par exemple, (John, SUBJ_V, found) est défini comme un vecteur trait du mot *John*, désignant que *John* est le *sujet du verbe found*. Un score est attribué à chacun de ces mots co-occurrents, basé sur une mesure de l'information mutuelle entre les vecteurs traits associés respectivement à l'occurrence m_i et au mot co-occurrent appartenant à son voisinage. Les *k* (valeur prédéfinie) premiers mots voisins ayants les scores les plus élevés sont les plus proches sémantiquement à l'occurrence m_i en question. Ces mots son représentés dans un vecteur définissant le vecteur contexte de m_i. Les vecteurs contextes associés aux différentes occurrences d'un mot dans le corpus d'entrainement, sont ensuite regroupés en clusters (ou comités) en fonction de leur degré de similitude. Chaque comité défini un sens possible du mot cible. Pour désambiguïser une nouvelle occurrence de ce mot, l'algorithme calcule la distance entre son vecteur contexte (construit à partir des *k* premiers mots co-occurrents appartenant à son voisinage) et le centroïde de chaque comité qui correspond à ce mot. Le centroïde du comité le plus proche du vecteur contexte de l'occurrence est considéré comme sens approprié de l'occurrence du mot à désambiguïser.

2.3.3 Les approches d'indexation sémantique basée sur la désambiguïsation

L'indexation sémantique a pour but de représenter les documents et les requêtes par les sens des mots qu'ils contiennent plutôt que par les mots eux-mêmes. Les sens des mots sont identifiés par des techniques de désambiguïsation qui s'appuient soit sur des ressources linguistiques externes telles que les dictionnaires, les thésaurus, les ontologies, …etc , soit sur des corpus d'apprentissage. Par conséquent, deux types d'approches d'indexation sont à distinguer : les approches d'indexation basée sur les ressources linguistiques externes et les approches d'indexation basées sur les corpus d'apprentissage.

2.3.3.1 Les approches d'indexation basée sur les corpus d'apprentissage

Le principe de ces approches est de construire d'abord les connaissances nécessaires pour la désambiguïsation des mots, en utilisant les corpus d'entrainement. Les termes d'indexation sont ensuite extraits à partir des contenus textuels des documents et requêtes, puis désambiguïsés en se basant sur les connaissances appris dans la phase d'apprentissage. Finalement, les documents et les requêtes sont indexés par les sens ainsi retrouvés.

L'approche de Weiss [Weiss, 73] s'appuie sur la connaissance des règles d'agencement et de fonctionnement qui sont générées par apprentissage à partir des contextes associées aux différents mots d'un corpus d'entrainement. Ces règles ont été appliquées dans la désambiguïsation des mots ambigus d'une collection de documents et requêtes. Les sens adéquats des mots issus de ce désambiguïseur ont été utilisés dans l'indexation, puis la recherche d'information en utilisant le système SMART. Les résultats rapportés par Weiss montrent que l'utilisation de sa technique de désambiguïsation ne permettait pas de rapporter une amélioration considérable des performances du SRI (une amélioration seulement de 1%).

L'approche de Schütze et al. [Schütze et al., 95] s'appuie sur la connaissance des usages des mots (*words usage*) qui sont obtenus en regroupant pour chaque mot dans le corpus d'apprentissage les contextes similaires de ses différentes occurrences, définissant ainsi les sens possible (ou usages possibles) de ce mot. La désambiguïsation d'une occurrence d'un mot ambigu dans un document ou une requête consiste alors à calculer le degré de similarité entre le contexte de l'occurrence et les usages possibles ce mot dans le corpus examiné. L'usage de mot qui maximise son degré de similitude est retenu comme sens adéquat de l'occurrence en question. En indexant la collection TREC-1 catégorie B, avec seulement 25 requêtes, Schütze et al. ont rapporté que l'indexation basée sur la combinaison des mots-clés et de leur trois meilleurs usages apportait un gain en précision de 14%.

2.3.3.2 Les approches d'indexation basées sur les ressources linguistiques externes

Le principe de base de ces approches consiste dans un premier temps à extraire, à l'issue d'une indexation classique, à partir du texte d'un document ou d'une requête ses mots-clés descriptifs (simples ou composés). Chaque mot-clé de l'index possède une ou plusieurs entrées dans une ressource linguistique (dictionnaire, thesaurus,…). Par conséquent, ces approches proposent de désambiguïser un mot d'index ambigu en se basant d'une part sur une ressource linguistique externe et d'une autre part sur les mots co-occurrents de son contexte. Pour ce faire, les différents sens possibles (ou sens candidats) d'un mot sont d'abord retrouvés à partir d'un dictionnaire ou autre ontologie. Puis un score est associé à chaque sens de ce mot généralement sur la base de son degré de relations avec les sens des mots de son contexte. Le sens (concept) qui maximise ce score est alors sélectionné comme sens correct du terme dans son contexte. Les sens des termes ainsi désambiguïsés seront ensuite utilisés seuls pour indexer les documents et les requêtes [Voorhees, 93 ; Khan et al., 04 ; Boughanem et al., 10 ; Boubekeur et al., 10b], ou éventuellement combinés aux mots clés de l'index classique [Hersh et al., 92 ; Mihalcea et al., 00 ; Baziz et al., 05a].

Dans l'approche d'indexation de Voorhees [Voorhees, 93], les sens possibles d'un mot représentent les synsets (concepts) qu'ils lui sont associés dans l'ontologie WordNet. Pour assigner le synset adéquat à un mot cible dans son contexte local (défini par la phrase où il apparait), chaque synset de ce mot est classé sur la base du nombre des mots co-occurrents communs entre son voisinage et son contexte local. Le sysnset le mieux classé représente le

sens adéquat de ce mot dans son contexte. Voorhees propose d'indexer les documents et les requêtes uniquement par les synsets des noms pondérés par le schéma classique *tf*idf*.

En s'appuyant sur la hiérarchie *is-a* dans WordNet, reliant les synsets noms par les relations *hyperonymie / hyponymie*, Voorhees définit le voisinage d'un synset nom *s* par « *le plus large sous-graphe connexe contenant s et seulement les descendants d'un ancêtre de s et ne contenant aucun synset ayant un descendant qui inclut une autre instance d'un membre (i.e mot) de s* ».

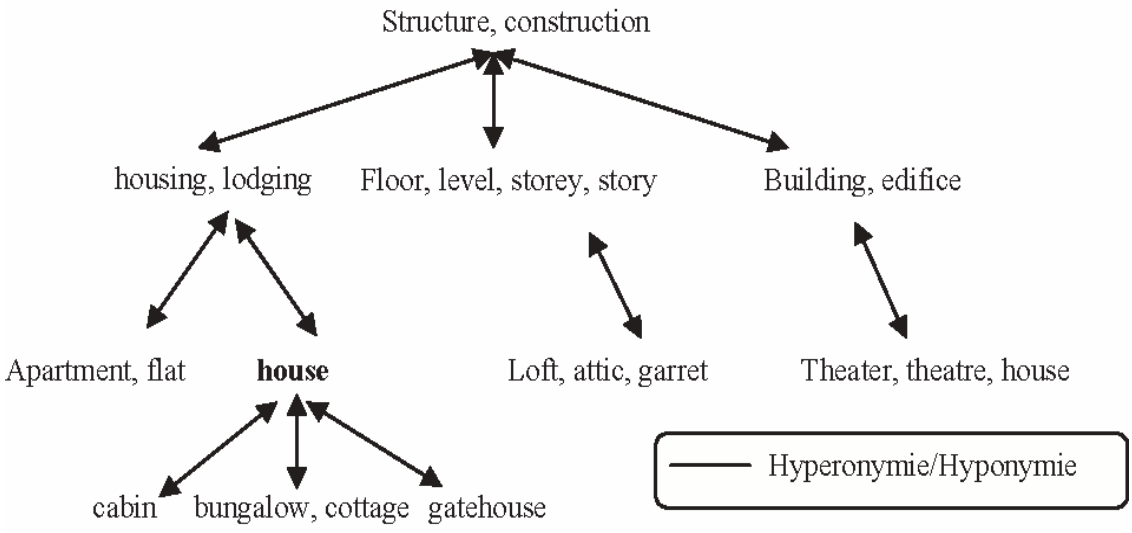


Figure 2.3 : Exemple de voisinage du premier synset *house* dans WordNet.

A titre d'exemple, dans la figure ci-dessus (figure2.3), le voisinage du premier synset de *house* correspond à l'ensemble des synsets {*housing, lodgins, Apartement, flat ; cabin ; bunglow, cottage ; gathehouse*}. Cependant, le synset *Structure, construction* ne sera pas inclus puisque son synset descendant *Theatre,theatre,house* contient le mot *house*. Par conséquent, *Theatre, theatre, house* correspond à un autre sens de *house*.

En utilisant une version modifiée du système SMART [Salton et al., 83] basé sur le modèle vectoriel, Voorhees a expérimenté son approche sur une collection de test désambiguïsée manuellement (les requêtes de la collection de test sont aussi désambiguïsées manuellement) par rapport aux performances du même processus sur la même collection avec une indexation classique basée sur mots-clés sans désambiguïsation. Les tests ont été effectués sur les collections CACM, CISI, Cranfield 1400, MEDLINE, et Time. Les résultats de ses expérimentations ont montré que pour chacune de ces collections, les performances du SRI diminuent sensiblement dans le cas de l'utilisation des collections désambiguïsées. Voorhees n'a pas mesuré la précision de son outil de désambiguïsation, elle a néanmoins évalué ses résultats subjectivement. Sa conclusion sur cette évaluation est que sa technique de désambiguïsation pourrait être imprécise conduisant ainsi à une dégradation des performances. De plus, elle a trouvé que son désambiguïseur était incapable de déterminer de façon exacte le sens attendu des mots dans les petites requêtes.

Dans une approche similaire, Uzener et al. [Uzener et al., 98] indexent les documents et requêtes par les sens correct des mots identifiés à partir de leur contexte local. Le contexte local d'un mot est défini dans cette approche, comme étant la liste ordonnée des mots démarrant du mot utile le plus proche du voisinage gauche ou droit jusqu'au mot cible. Par exemple dans, le texte suivant: ‘‘*the jury had been* ***charged*** *to investigate reports of irregularities in the primary…*”, le contexte local droit de ‘*charged*’ est ‘‘*X to investigate*’’ et son contexte local gauche est ‘‘*the jury had been X*’’.

En se basant sur l'hypothèse que *les mots utilisés dans le même contexte local ont souvent des sens proches*, Uzener et al. proposent d'utiliser l'ensemble des mots du contexte local (ensemble de *Selecteurs*) d'un mot ambigu pour retrouver le bon sens qui le correspond dans ce contexte. Pour ce faire, l'ensemble de *Selecteurs* d'un mot cible est d'abord construit à partir des mots non vides extraits de son contexte local, puis comparé à chaque synset (ensemble de mots synonymes) de ce mot. Le synset ayant le plus grand nombre de mots communs avec l'ensemble de *Selecteurs* représente le sens correct du mot dans son contexte local. Pour évaluer la précision de ce désambiguïseur, Uzener et al. l'ont testé sur le corpus SemCor. Les résultats rapportés présentent une précision de 60%. Néanmoins, en intégrant ce désambiguïseur dans le système de recherche d'information SMART, les résultats obtenus ont montré que leur algorithme de désambiguïsation n'améliore pas les performances du système et produit des erreurs dans l'identification des sens des mots.

Dans une approche différente, Khan et al. [Khan et al., 00 ; 04] proposent une indexation basée sur les concepts de l'ontologie de domaine du sport. Les termes d'indexation sont d'abord identifiés à partir des textes, en utilisant les techniques linguistiques classiques, puis projetés sur l'ontologie de domaine (du sport) pour détecter les concepts qui leur correspondent dans l'ontologie. Pour ce faire, chaque terme est comparé à chaque concept (caractérisé par un ensemble de termes synonymes) dans l'ontologie. L'ensemble des concepts de l'ontologie qui s'accordent à un terme défini alors l'ensemble de ses sens possibles. Pour désambiguïser un terme, Khan et al. se basent sur l'idée que *les concepts qui se trouvent dans une même région dans l'ontologie sont souvent proches sémantiquement*. Ils proposent alors de calculer un score associé à chaque concept du terme sur la base de ses proximités sémantiques dans l'ontologie, avec les concepts des autres mots de son contexte. La proximité sémantique entre deux concepts dans cette approche, est mesurée par l'inverse de leur distance minimale. Le concept ayant le plus haut score est alors retenu comme sens (ou concept) correct du terme ambigu.

Khan et al. expérimentent leur approche sur des paragraphes annotant des passages audio dans le domaine du sport. Les résultats rapportés montrent que leur modèle vectoriel, basé sur l'appariement de vecteurs de concepts issus de leur technique d'indexation et pondérés par le schéma *tf*idf*, produit de meilleures valeurs en précision et rappel (de l'ordre de 90%) par rapport à un modèle vectoriel classique basé sur les mots clés.

Dans une approche similaire, Baziz et al., [Baziz et al, 04, 05a ; Baziz et al, 05b] proposent de représenter un document ou une requête par des concepts et des relations entre concepts. Cette approche consiste dans un premier temps à projeter (mapper) le contenu textuel d'un

document (ou d'une requête) sur la ressource linguistique WordNet. L'objectif est d'extraire ses termes (simples et collocations de mots) représentants des concepts dans WordNet. Les termes extraits sont ensuite pondérés par un nouveau schéma proposé dit *Cf*idf*, qui étend la pondération *tf*idf* pour tenir compte des termes composés (collocations de mots). L'intérêt de la pondération de ces termes est de garder uniquement les termes d'indexation les plus représentatifs du contenu du document ou de la requête. Lorsqu'un de ces termes s'apparie avec plusieurs concepts (dits concepts candidats) dans l'ontologie, il est désambiguïsé suivant le principe que le concept correspondant au sens adéquat du terme est celui qui est fortement lié avec les concepts des autres termes du document. Ainsi, un score est affecté à chaque concept candidat du terme ambigu basé sur le cumul de ses similarités sémantiques avec les concepts candidats des autres termes. Le concept qui maximise ce score définit alors son sens correct dans le document. Finalement, le document (ou la requête) dans cette approche, est représenté par un réseau sémantique (noyau sémantique), où les nœuds sont des concepts représentant les sens de ses termes (concepts) et les arcs pondérés définissent les similarités sémantiques [Leacock et al., 94 ; Lin, 98 ; Resnik, 99 ; Lesk , 86] entre les concepts liés.

L'approche d'indexation ainsi proposée, dite DocCore, a été évaluée d'une part dans le cadre la collection Muchmore et d'autre part dans le cadre de la campagne CLEF 2004, en utilisant un SRI basé sur le modèle connexionniste [Boughanem et al., 98]. Les résultats rapportés montrent que l'indexation sémantique par les concepts seuls diminue les performances de la recherche par rapport à une indexation classique basée sur les mots-clés. Cependant, la combinaison de l'indexation sémantique avec les mots-clés de l'indexation classique apporte une amélioration significative de la précision du SRI.

En outre, Baziz et al. proposent l'approche DocTree, complémentaire à l'approche précédente, qui projette les réseaux sémantiques de la requête et le document construits, à partir de la méthode DocCore, sur l'hiérarchie *is-a* (*hyperonymie/ hyponymie)* de l'ontologie WordNet. Un mécanisme de complétion a été appliqué sur ces représentations permettant de rajouter les nœuds intermédiaires en remontant la hiérarchie de concepts, à partir des feuilles vers la racine. Par conséquent, la requête et le document sont représentés par deux sous arbres formés par les concepts qu'ils contiennent et qui apparient ceux de l'ontologie WordNet. Pour évaluer la pertinence document-requête, leurs sous-arbres respectifs seront comparés, en se basant sur des opérateurs flous permettent d'évaluer jusqu'à quel point le document traite le thème décrit dans la requête, et d'affecter ainsi un degré de pertinence au document.

Dans [Boubekeur et al., 08], les auteurs ont proposé une approche d'indexation conceptuelle sémantique qui utilise le réseau lexical WordNet et les règles d'association pour construire le graphe CP-Net représentatif du contenu du document. Les termes d'indexation sont d'abord identifiés en suivant les étapes de l'indexation classique. Puis chaque mot non vide est projeté sur WordNet dans le but de retrouver toutes les entrées (mot simples ou collocations de mots) de la ressource contenant ce mot. Ces entrées qui correspondent à des termes d'indexation, ont été pondérées par des poids traduisant leur importance dans le document. L'approche propose une pondération classique, par le schéma *tf*idf*, pour les mots clés simples, et une pondération sémantique pour chaque terme composé (ou collocation de mots), en se basant sur une mesure probabiliste des sens possibles de ce terme par rapport aux

sens de ses sous-termes et de ses sur-termes, en tenant compte de leurs fréquences d'occurrences respectives dans le document. Les seuls termes (simples ou composés) dont le poids est supérieur à un seuil minimal fixé, ont été retenus comme termes d'indexation représentatifs du document. Ces termes dénotent un ou plusieurs concepts dans la ressource WordNet définissant leurs sens possibles. Pour désambiguïser un terme ambigu, un score associé à chacun de ses concepts est calculé en fonction de la somme de ses similarités sémantiques avec les concepts (sens) des autres mots dans le document, en tenant compte des poids de leurs termes respectifs. Le concept qui a le plus haut score est retenu comme sens adéquat du terme dans le document. Finalement, le document est représenté par un graphe CP-Net, où les nœuds définissent les concepts attachés aux sens adéquat des termes du document, et les arcs représentent les relations contextuelles latentes entre ces concepts. Ces relations sont retrouvées en moyen des règles d'association sémantiques. De même, la requête est représentée par le formalisme CP-Net, en considérant les préférences qualitatives d'un utilisateur. Pour retrouver les documents pertinents pour une requête donnée, l'appariement repose sur la comparaison de leurs graphes CP-Net respectifs, permettant ainsi de calculer le degré de similarité entre eux.

Cette approche n'a pas été expérimentée pour manque d'un cadre d'évaluation adéquat supportant le formalisme CP-Net. Néanmoins, Boubekeur et al. revisitent leur approche et proposent une autre approche d'indexation sémantique dans [Boubekeur et al., 10b], basée sur un principe similaire. A la différence de la première, cette approche définit un nouveau score de désambiguïsation et un nouveau schéma de pondération des concepts. Le score de désambiguïsation proposé est associé à chaque concept d'un terme ambigu, sur la base de la fréquence de ce terme dans le document, et de ses distances sémantiques avec les concepts des autres termes les plus fréquents dans le document. Le concept qui présente le meilleur score est considéré comme sens correct du terme. Les concepts ainsi identifiés, ont été alors pondérés par une mesure sémantique exprimée par le cumul de ses distances avec les autres concepts dans le document et sa fréquence relative. Au final, l'ensemble des concepts associés aux termes dans le document (ou la requête), ont permis de construire un noyau sémantique représentant le contenu de document (ou respectivement de la requête).

L'approche a été expérimentée sur la collection Muchmore, en utilisant le système Mercure [Boughanem et al, 92], basé sur le modèle connexionniste. Les résultats rapportés ont montré que l'indexation sémantique a apporté un gain de précision de plus de 50%, en pondérant les concepts par *tf***idf*. Néanmoins, elle en demeure moins précise, en pondérant avec le schéma de pondération proposé. Boubekeur et al. expliquent cette diminution par le fait que le score ranking du système Mercure, qui est basé sur *tf***idf* (ou une de ses variantes), remplace *tf* par le poids sémantique du concept proposé combiné à une mesure non corrélée *idf*.

Dans une approche différente, Mallak [Mallak, 11] indexe les documents et les requêtes par des clusters de concepts les plus représentatifs de leurs contenus sémantiques. Pour ce faire, il reprend la même technique de mapping proposé par Baziz [Baziz et al., 05a] pour la détection des termes-clés (simples ou composés) d'un document (ou d'une requête donnée). Chaque terme-clé identifié peut être attaché dans l'ontologie WordNet à un ou plusieurs

concepts (dits concepts-termes). Pour retrouver le meilleur concept qui décrit le sens d'un terme dans son contexte d'apparition, Mallak propose une nouvelle méthode de désambiguïsation basée sur la notion de centralité [Mallak, 11 ; Boughanem et al., 10]. Il définit la centralité d'un concept-terme par le nombre de ses relations sémantiques de WordNet (telles que : la synonymie, l'hyperonymie, …etc), qu'il partage avec les concepts des autres termes dans le document. Le concept choisi comme sens de ce terme est celui qui possède une valeur maximale de sa centralité. Une fois que tous les concepts-termes significatifs d'un document (ou respectivement d'une requête) sont déterminés, ceux qui sont en relation dans WordNet sont groupés en clusters. L'ensemble des clusters de concepts-termes décrit ainsi le contenu d'un document (respectivement d'une requête). Pour calculer la pertinence d'un document vis-à-vis d'une requête donnée, le modèle de recherche proposé se base sur l'appariement entre les graphes associés aux clusters document-requête, en utilisant dans la fonction de matching une mesure qui combine trois facteurs : la centralité d'un concept-terme dans le cluster, sa fréquence relative dans le document et sa spécificité définie par sa profondeur dans l'hiérarchie *is-a* de WordNet. Le modèle de recherche ainsi proposé permet d'obtenir des documents sémantiquement pertinents même s'ils ne contiennent aucun concept-terme de la requête.

Mallak a expérimenté son approche sur deux collections TREC : la collection TREC1 et la collection TREC7. Les résultats rapportés montrent que les clusters concepts utilisés dans l'indexation ont permis d'améliorer les résultats de la recherche comparés à ceux obtenus par une indexation classique basée mots-clés. De plus, la définition de centralité d'un concept appliquée dans la technique de désambiguïsation proposée apporte plus de précision (des gains de précisions >5%), que les liens sémantiques entre concepts utilisés dans la technique de désambiguïsation proposée par Baziz et al. [Baziz et al., 05a]. Par ailleurs, Mallak a testé l'apport de son modèle de recherche, en s'appuyant sur les facteurs proposés : centralité, fréquence et spécificité d'un concept, par rapport à un modèle classique basé sur une indexation par des mots-clés simples pondérés par le schéma *Okapi-BM25*. Il conclut que son modèle de recherche est plus performant que le modèle de recherche classique.

Dans [Harrathi et al., 10], les auteurs proposent une nouvelle approche d'indexation sémantique de documents multilingues. D'une part, les termes d'indexation simples sont extraits par une approche d'indexation classique. D'une autre part, les termes composés représentant des collocations de mots, sont identifiés par une mesure statistique qui repose sur la fréquence des mots simples qui apparaissent mutuellement dans le contenu textuel d'un document (ou d'une requête). L'ensemble des termes simples ou composés sont ensuite pondérés par nouvelle mesure *CTF*Idf* (*CTF* pour *Compound Term Frequency*), qui étend le schéma *tf*idf* pour la prise en compte dans la pondération des termes composés leur longueur ainsi que les fréquences des mots simples qui les composent. Les termes d'indexation dont le poids est supérieur à un seuil minimal, sont alors projetés sur une ontologie sémantique multilingue dans le but d'obtenir les concepts définissant les sens possibles de ces termes. Pour retrouver le sens (ou le concept) adéquat de chaque terme ambigu, deux types de désambiguïsation ont été utilisés : une désambiguïsation langagière et une désambiguïsation sémantique. La désambiguïsation langagière a pour objet d'identifier la langue des termes

ambigus dans leur contexte d'apparition, afin de garder uniquement pour ces termes leurs concepts candidats appartenant à cette langue. A titre d'exemple, le mot *table* possède dans la langue française des concepts différents à ses concepts dans la langue anglaise. Pour résoudre le problème de l'ambiguïté langagière, la langue d'un terme ambigu est définie, dans cette approche, à partir du terme non ambigu le plus proche de lui, qui possède un seul concept dans l'ontologie multilingue. La désambiguïsation sémantique consiste alors à retrouver le sens correct du terme dans sa langue d'utilisation. Pour ce faire, le contexte du terme cible est d'abord construit à partir de l'ensemble de toutes les phrases où il apparait. Puis, un score est attribué à chaque concept candidat de ce terme, basé sur le nombre de ses relations dans l'ontologie multilingue avec les concepts des autres termes non ambigus appartenant à son contexte. Le concept ayant le score le plus élevé est alors retenu comme sens approprié de ce terme dans le document. Finalement, chaque document (ou requête) est représenté par l'ensemble des concepts des termes qui se rattachent à leurs sens adéquats utilisés dans la langue du document (ou respectivement de la requête).

Harrathi et al., ont évalué leur approche en l'incorporant dans un SRI basé sur le modèle de langue proposé par [Maisonnasse et al., 09], en utilisant la ressource médicale UMLS et la collection de test CLEFmed 2007, contenant des documents médicaux écrits en trois langues : l'anglais, le français et l'allemand. Les résultats ont montré que cette approche apporte un gain de la précision moyenne de 5% par rapport à une indexation classique basée sur les mots clés.

Dans le domaine biomédical, Dinh [Dinh, 12 ; Dinh et al., 10] présente une approche d'indexation sémantique qui s'appuie sur les concepts du thesaurus MeSh. Cette approche débute par une extraction de concepts à partir d'un document ou d'une requête, en projetant son contenu textuel sur une liste préétablie de tous les concepts appartenant au thesaurus MeSh. Les concepts de la liste contenant un terme du document (ou respectivement de la requête), sont considérés comme concepts candidats de ce terme. Un score est ensuite affecté à chaque concept candidat du terme, sur la base de sa similarité thématique au texte et sa similarité structurelle définie par la corrélation entre son entrée concept dans le thésaurus et l'expression représentée par une succession de mots du texte en incluant le terme. La similarité structurelle d'un concept est calculée dans ce score selon la mesure de Spearman. Le concept candidat qui maximise ce score est alors considéré comme concept représentatif du document ou de la requête. Cependant, certains concepts sont ambigus. Par conséquent, Dinh a proposé deux techniques de désambiguïsation permettant d'assigner à chaque concept le sens adéquat dans son contexte d'utilisation :

- La première technique de désambiguïsation, dite désambiguïsation de proche en proche avec propagation de sens, consiste à calculer de proche en proche le sens du concept dans le document par la similarité sémantique entre celui-ci et son voisin précédent désambiguïsé, en utilisant la mesure de [Leacock et al., 98]. En se basant sur l'hypothèse *d'unicité de sens d'un concept dans le document*, une fois que le concept est désambiguïsé, son sens est propagé pour toutes ses occurrences dans le document.

- La seconde technique de désambiguïsation, dite désambiguïsation basée sur le clustring, consiste à regrouper tous les concepts du document du même groupe de domaines (les domaines MeSh sont au nombre de 16 dont : A : Anatomy, C : Adiseases, …), en clusters. Par conséquent, il est possible qu'un concept soit classé dans différents groupes de domaines ou clusters, d'où l'ambiguïté de ce concept. Pour déterminer le bon sens du concept dans le document, un score est attribué à chaque sens possible appartenant à chaque cluster possible du concept, basé sur sa similarité [Leacock et al., 98] avec les sens des autres concepts dans le même cluster. Le sens qui maximise son score dans un cluster définit ce dernier comme son domaine d'application dans le document et son sens est celui qui appartient à ce cluster. Ce sens est alors propagé à toutes les occurrences de ce concept dans le document.

Finalement, chaque document (ou requête) est représenté par un index sémantique contenant à la fois les concepts MeSh identifiés et les termes du document (ou respectivement de la requête) qui ne correspondent pas à des entrées dans MeSh. Pour mesurer l'importance de chaque terme d'indexation sémantique, Dinh propose de pondérer un terme qui ne possède pas une entrée dans MeSh par la mesure *tf*idf*, tandis qu'un concept de l'index est pondéré par une mesure sémantique basée sur sa centralité normalisée, exprimée par son degré de relations qu'il partage avec les autres concepts du document (ou la requête), et sa fréquence normalisée, exprimée par son *tf*idf*.

L'approche est évaluée en l'adaptant à l'indexation des journaux médicaux de la collection OSHUMED. Les résultats montrent que le SRI améliore les résultats de la recherche avec les deux techniques de désambiguïsation proposées, par rapport à ceux obtenus par une indexation classique basée mots-clés pondérés par *Okapi-BM25* (un gain de performance de 17,35% en utilisant la désambiguïsation de proche en proche et de 17,06% avec la désambiguïsation basée sur le clustering).

2.4 Conclusion

Dans ce chapitre, nous avons passé en revue les différentes approches d'indexation sémantique proposées en recherche d'information. Ces approches ont apporté la preuve que la représentation des documents et requêtes par les sens (ou concepts) de leurs mots est bénéfique dans un processus de recherche, permettant ainsi de résoudre les problèmes causés par les SRI classiques. Ces sens sont le plus souvent identifiés par des approches de désambiguïsation utilisant des ressources linguistiques externes. En se basant sur ce même principe, nous présentons dans le chapitre suivant nos contributions à la définition d'un nouveau modèle de RI sémantique.

PARTIE 2

Contributions

Chapitre 3

Approche de RI sémantique

Plan du chapitre

3.1 Introduction **57**

3.2 Motivations **57**

3.3 Approche d'indexation sémantique de documents textuels **59**

3.3.1 Préliminaires : définitions et notations 60

3.3.2 Aperçu général de l'approche 60

3.3.3 Description détaillée de l'approche 62

3.3.4 Illustration et discussion 71

3.4 Appariement Document-Requête **82**

3.5 Conclusion **83**

3.1 Introduction

Nous avons présenté dans le chapitre précédent, un état de l'art sur les principaux travaux dans la RI sémantique. Ces approches tentent de pallier les limites de l'indexation classique, basée mots-clés, en offrant le moyen de lever l'ambiguïté de la langue naturelle grâce à la représentation des documents et requêtes par les concepts (sens) plutôt que par les mots qu'ils contiennent. Le but étant de retrouver les documents sémantiquement pertinents à une requête utilisateur.

Dans ce chapitre, nous présentons notre contribution portant sur la définition d'un nouveau modèle de RI sémantique dans les documents textuels [Azzoug et al., 13a]. En particulier, nous proposons une approche d'indexation sémantique des documents et une approche d'évaluation sémantique des requêtes.

1. *L'approche d'indexation sémantique des documents* [Azzoug et al., 11] : repose sur une nouvelle méthode de détection des concepts s'appuyant sur WordNet et WordNetDomains comme sources d'évidence dans le processus de désambiguïsation. Les concepts identifiés sont ensuite pondérés et rangés dans un index représentant le contenu sémantique du document. Nous proposons deux schémas de pondération sémantique basés sur la centralité d'un concept. La centralité d'un concept est traduite d'une part par son importance sémantique (exprimé par ses relations sémantiques avec les autres concepts du document) et d'autre part par sa fréquence d'occurrence dans le document.
2. *L'approche d'évaluation sémantique des requêtes :* vise à attribuer un score de pertinence à chaque document pour une requête donnée, basé sur la proximité sémantique des vecteurs de concepts associés respectivement au document et à la requête. En particulier, nous proposons d'étendre la mesure de cosinus [Salton et al., 83] afin de calculer la pertinence sémantique du document vis-à-vis de la requête.

Le chapitre est structuré comme suit : en section 3.2, nous présentons les motivations qui ont été l'origine de nos propositions. En section 3.3, nous exposons les fondements théoriques de notre approche d'indexation de documents textuels. Enfin, la section3.4 est dédiée à notre approche d'évaluation des requêtes.

3.2 Motivations

En RI sémantique, documents et requêtes sont indexés par des concepts permettant de décrire au mieux leurs contenu informationnel, plutôt que par les mots qu'ils contiennent. L'objectif à travers cette représentation est de résoudre les problèmes d'ambiguïté et de disparité des mots. En indexation sémantique, les concepts sont identifiés à partir d'ontologies ou autres ressources linguistiques, en s'appuyant sur des techniques de désambiguïsation des sens des mots. Les concepts sont ensuite assignés de poids définissant leur degré d'importance dans le document et la requête. Finalement, les représentations conceptuelles du

document et de la requête sont comparées lors de la recherche, pour retrouver les documents sémantiquement pertinents pour la requête.

Néanmoins, il reste à notre sens certains points problématiques sur lesquels nous souhaitons les améliorer à travers notre modèle de RI sémantique proposé dans ce contexte.

1. Le premier point concerne la *désambiguïsation des sens des mots*, qui consiste à identifier le concept (sens) correct d'un mot ambigu selon son contexte d'apparition. Pour cela, les approches de désambiguïsation proposées en RI reposent généralement sur un score associé à chaque sens sur la base de son degré de relations sémantiques avec les sens des autres mots de son contexte. Ces relations sémantiques sont définies à partir d'un dictionnaire informatisé ou un thésaurus ou autres ontologies. Il nous parait qu'une autre dimension sémantique pourrait être aussi une voie intéressante à exploiter dans la désambiguïsation des sens des mots, celle de leurs domaines d'usage dans le document. Cette intuition est fondée sur l'idée que les mots de la langue, utilisés dans un même contexte, portent des sens fortement liés sémantiquement traitant un même domaine ou des domaines similaires qui se rattachent à la thématique abordée dans le document. Par conséquent, l'exploitation de cette dimension renforce la désambiguïsation du mot cible permettant ainsi de retrouver le ou les concepts probables qui peuvent lui être associés, par rapport aux autres concepts qui n'appartiennent pas à son domaine d'usage, puis sélectionner parmi ces concepts celui qui le correspond dans son contexte d'utilisation. Nous proposons alors de désambiguïser un mot en nous basons d'abord sur une désambiguïsation de ses domaines pour trouver le domaine adéquat dans son contexte d'apparition, afin de garder uniquement les sens liés au sujet (topic) du document, puis sur une désambiguïsation sémantique dans le domaine choisi pour identifier le concept correct du mot dans son contexte. Pour ce faire, en s'appuyant les deux ressources linguistiques WordNet et son extension aux domaines WordNetDomains, nous désambiguïsons le domaine d'usage d'un mot par son degré de proximités sémantiques dans WordNetDomains avec les domaines des autres mots de son contexte. Sa désambiguïsation sémantique dans ce domaine repose sur un score basé sur le cumul de ses similarités sémantiques dans WordNet avec les sens des autres mots appartenant à leur domaine d'usage respectif. Le sens qui maximise ce score représente le sens approprié du mot cible dans son contexte.

2. Le second point concerne la *pondération des concepts* qui permet d'associer à chaque concept un poids traduisant son importance dans le document. Le poids d'un concept est calculé généralement par une mesure statistique exprimée par la fréquence relative de son terme associé dans le document. Ceci à notre sens présente une problématique du fait que cette mesure considère uniquement l'importance apparente du concept, traduite par sa fréquence de son terme correspondant, et ne tient pas compte de son importance sémantique latente par rapport aux autres concepts dans le document. De ce fait, un poids important peut être attribué à un concept ayant une forte présence de son terme associé et une faible corrélation avec les autres concepts identifiés dans le document. Tandis qu'un poids de moindre importance peut être attribué à un concept ayant une faible présence de son terme associé et une forte corrélation avec les autres concepts dans le document. Pour résoudre ce problème, nous proposons de pondérer les concepts par une mesure sémantique exprimée par

leur centralité dans le document. Nous définissons la centralité d'un concept par la combinaison de son importance apparente exprimée par sa fréquence, et son importance sémantique latente exprimée par ses proximités sémantiques avec les autres concepts dans le document.

3. Le dernier point concerne le *processus d'évaluation des requêtes*. Dans les modèles de recherche les plus appliqués en RI sémantique, tels que le modèle vectoriel ou le modèle connexionniste, …, la fonction de *matching* dans l'appariement document-requête calcule le degré de pertinence d'un document via une requête en considérant uniquement les termes (concepts éventuellement combinés à des mots-clés) de la requête qui sont présents dans le document. Par conséquent, les documents ne contenant aucun terme de la requête mais qui présentent une pertinence sémantique latente (exprimée à travers les relations sémantiques qui lient leurs termes respectifs), ne seront pas retrouvés par le processus de recherche. A titre d'exemple, dans une recherche avec le concept *AIDS* qui correspond à la maladie du sida, les documents qui ne contiennent pas ce concept mais qui traite le concept *HVI* qui correspond au virus du sida ne seront pas retournés par le système. Pour pallier au problème de disparité de termes, nous proposons d'intégrer dans le score de pertinence les proximités sémantiques entre les termes de la requête et le document.

En résumé, dans notre modèle de RI sémantique nous proposons :

1. Une approche d'extraction des termes des documents et requêtes,
2. Une désambiguïsation sémantique des mots basée sur l'utilisation conjointe de la base lexicographique WordNet et son extension aux domaines WordNetDomains,
3. Une pondération sémantique des concepts fondée sur la notion de centralité d'un concept dans le document,
4. Une approche d'évaluation sémantique des requêtes.

3.3 Approche d'indexation sémantique de documents textuels

Notre approche d'indexation sémantique [Azzoug et al., 11] a pour objectif de représenter le contenu textuel d'un document (ou d'une requête) par un vecteur de concepts (sens). Cette approche s'articule autour des deux caractéristiques suivantes :

1. L'identification des concepts représentatifs du document (ou respectivement *d'une requête*) [Azzoug et al., 12], est basée sur le *mapping* de chaque terme du document (ou de la requête) sur WordNet afin d'en récupérer le ou les sens (- synsets-) possibles. Un terme pouvant être associé à un ou plusieurs concepts (sens) dans WordNet. Pour identifier son sens correct dans son contexte d'utilisation dans le document (ou la requête), nous proposons de nouvelles techniques de désambiguïsation contextuelle [Azzoug et al., 13c], en utilisant comme sources d'évidences WordNet et son extension aux domaines, WordNetDomains. (Notons qu'une brève description de la ressource WordNet et son extension WordNetDomains est donnée en annexe).

2. *La pondération sémantique des concepts* associe à chaque concept identifié un poids numérique représentant son degré de centralité (importance) dans le document [Azzoug et al., 13b].

3.3.1 Préliminaires : définitions et notations

Dans cette partie, nous présentons les définitions et notations qui seront utilisées dans la description théorique de notre approche d'indexation sémantique :

- Une *occurrence* d'un mot *m* est toute instance m_i de *m* dans le document.

- Une *collocation* est un groupe de mots qui apparaissent ensemble dans un texte donné plus souvent que par le simple fait du hasard [Harrathi, 10]. Un mot composé est considéré comme une collocation. La taille d'une collocation est définie comme le nombre de mots qui la composent.

- Un *terme* est une unité linguistique représentée par un mot-clé non vide (ex : *dog*) ou par une collocation de mots (ex : *domestic_dog)*.

- On appelle *expression locale* de l'instance m_i (notée E_i) du mot *m*, la chaîne de caractères concaténant, par le biais du souligné (_), le mot m_i et les mots successifs situés à sa droite jusqu'à la prochaine ponctuation. La taille d'une expression locale est définie comme le nombre de mots qui la composent.

- Le *lemme d'un mot* m_i est la racine canonique de m_i, représentant son infinitif si m_i est un verbe, ou sa forme masculin singulier si m_i est un nom, un adjectif ou un adverbe. (ex : le lemme du nom *ligaments* est *ligament*) .On notera le lemme de m_i par *lemme_*m_i.

- Le *lemme de l'expression locale* E_i d'un mot m_i est l'ensemble des formes de base de tous les mots qui la composent, concaténées par le biais du souligné (_). (ex : lemme de *domestic_dogs* est *domestic_dog*). Le lemme de l'expression locale E_i sera noté par *lemme_*E_i.

- Un concept est une abstraction généralisée de propriétés communes à plusieurs objets, faits ou évènements, … [Chevallet, 09]. Dans un texte, le concept est représenté par un ou plusieurs termes synonymes ayant une entrée dans une ressource terminologique (ontologie, thesaurus, dictionnaire…) [Chevallet, 09]. Ainsi, un *synset* de WordNet est un concept.

3.3.2 Aperçu général de l'approche

Notre processus d'indexation, illustré en figure 3.1, est constitué en quatre étapes :

(1) L'identification des termes descriptifs du document (ou d'une requête) consiste à extraire à partir du contenu textuel du document (ou de la requête), et en s'appuyant sur la liste prédéfinie des collocations de WordNet :

- l'ensemble des collocations de mots,

- l'ensemble des mots simples ayant une entrée dans WordNet,

- l'ensemble des mots orphelins correspondant aux mots simples n'ayant pas d'entrée dans WordNet.

(2) La désambiguïsation des termes identifie les sens adéquats des mots dans leur contexte d'apparition en utilisant conjointement les ressources linguistiques: WordNet et WordNetDomains.

Nous définissons deux types de contextes pour un mot donné :

- un contexte local,
- un contexte global.

Et nous proposons trois différentes techniques de désambiguïsation, selon le contexte considéré :

- une désambiguïsation dans le contexte local,
- une désambiguïsation dans le contexte globale,
- une désambiguïsation mixte (désambiguïsation dans le contexte local suivie d'une désambiguïsation dans le contexte global).

Remarque

Les collocations sont considérées comme des expressions désambiguïsées. Notre approche de désambiguïsation est ainsi appliquée aux seuls mots simples ambigus ayant une entrée dans l'ontologie WordNet.

(3) La pondération des concepts repose sur le calcul d'un score de centralité. La centralité d'un concept est une mesure combinée de sa fréquence d'occurrence et de son importance sémantique dans le document, exprimée par la somme de ses similarités sémantiques aux autres concepts du document. Les termes orphelins sont pondérés par la mesure classique *tf*idf.*

(4) Construction de l'index sémantique en organisant les mots orphelins et les concepts (collocations et sens des mots simples) pondérés dans un index décrivant le contenu sémantique du document.

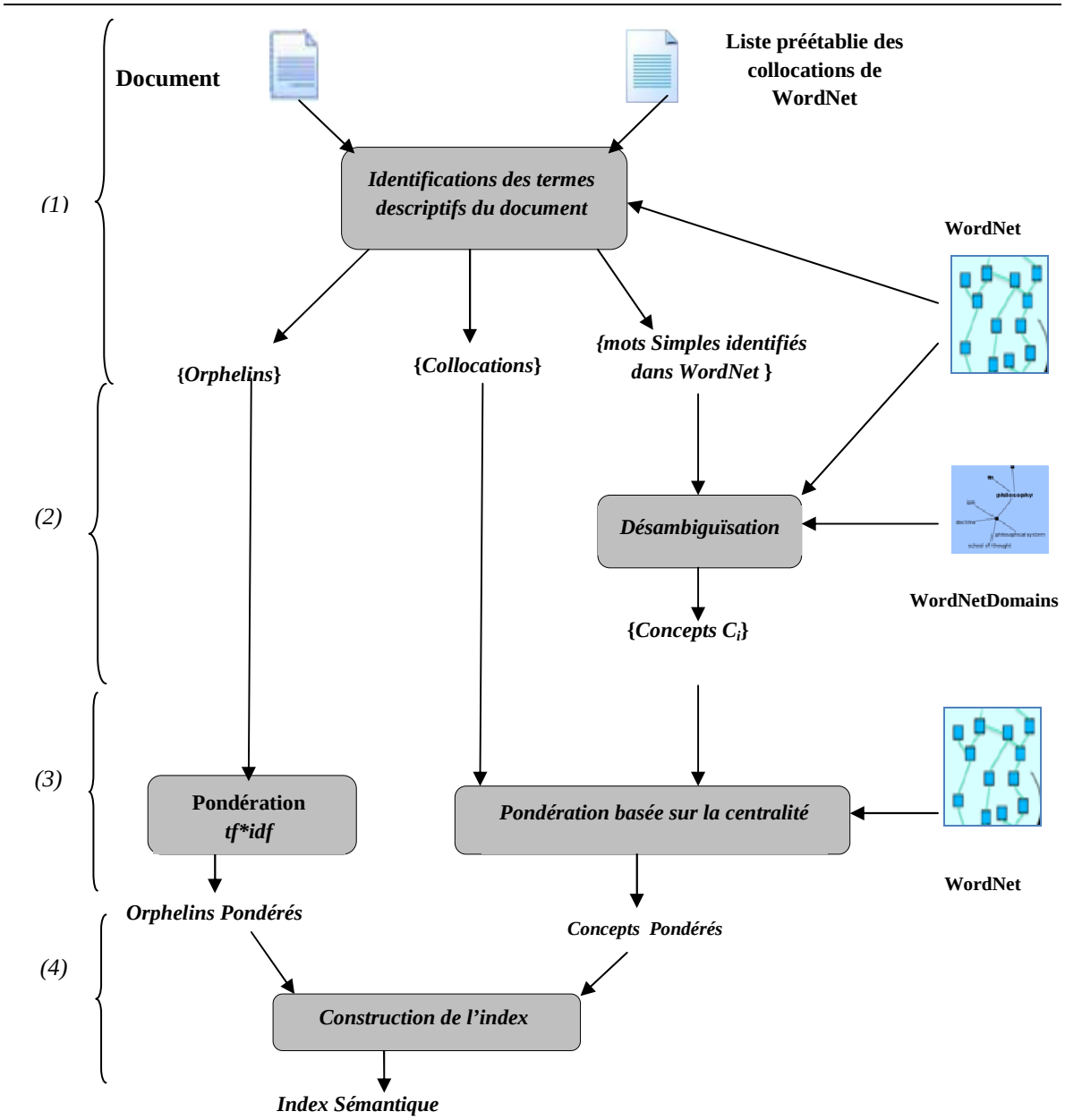


Figure 3.1 : Processus d'indexation sémantique.

3.3.3 Description détaillée de l'approche

3.3.3.1 Identification des termes descriptifs du document (ou de la requête)

Cette étape débute par l'identification des collocations de mots dans un document *d* à partir d'une liste préétablie $\varphi_{Collocs}$ de toutes les collocations existantes dans WordNet. Cette démarche est très importante dans la mesure où les collocations sont considérées comme des termes monosémiques et possèdent généralement un seul sens dans WordNet, ce qui a pour effet de réduire l'ambiguïté des mots [Baziz et al, 05a]. Pour cela, pour chaque mot m_i à analyser, nous récupérons de $\varphi_{Collocs}$, la liste ς_i des collocations commençant par le mot m_i

triées par ordre décroissant de tailles. La plus longue collocation de ς_i qui s'apparie à l'expression locale E_i de m_i ou à sa forme de base *Lemme_*E_i est retenue comme collocation du document *d*. Si aucune collocation de ς_i ne s'apparie avec E_i ou *Lemme_*E_i alors m_i est un mot simple. Il est considéré comme un mot orphelin s'il ne possède pas d'entrée dans WordNet.

Le principe d'identification des termes est décrit dans l'algorithme présenté en tableau 3.1.

Algorithme de détection des termes

Entrée : document *d* et la liste $\varphi_{Collocs}$ de touts les collocations existantes dans WordNet 2.1

Sortie : ξ_{Expres}, $\xi_{Simples}$, ξ_{Orphel}

{On suppose m_i le prochain mot à analyser dans *d*}

Début

1. Extraire de $\varphi_{Collocs}$ l'ensemble $\varsigma_i = \{C^i_1, C^i_2, \ldots C^i_n\}$ des collocations commençant par le mot m_i ;
2. Ordonner ς_i comme suit : $\varsigma_i = \{C^i_{(1)}, C^i_{(2)}, \ldots, C^i_{(n)}\}$ où $(j)_{1..n}$ est une permutation d'indices telle que $|C^i_{(1)}| \geq |C^i_{(2)}| \geq \ldots \geq |C^i_{(n)}|$, où $|C^i_{(j)}|$ est la taille de la collocation $\{C^i_{(j)}\}$;
3. bool<- faux ;
4. Tant qu'il existe $C^i_{(j)}$ dans ς_i et bool=faux

faire :

4.1. Calculer l'expression locale E_i de taille $|C^i_{(j)}|$,;

4.2. Si $E_i = C^i_{(j)}$ alors bool <- vrai ;

Sinon
Début
*lemme_* E_i =*lemmatiser* (E_i) ;
Si $lemme\_E_i = C^i_{(j)}$ alors bool<-vrai ;
Fin

fait ;

5. Si (bool) alors insérer *lemme_*E_i dans ξ_{Expres}
6. Sinon Debut
   *lemme_* m_i =*lemmatiser* (m_i) ;
   Refaire les étapes de 1 à 5 en utilisant *lemme_*m_i à la place de m_i ;
   Si (bool=faux) alors
   Si (m_i mot non vide) alors
   Début
   Si *lemme_*m_i possède une entrée dans WordNet
   alors insérer *lemme_*m_i dans $\xi_{Simples}$
   Sinon insérer m_i dans ξ_{Orphel} ;
   Fin
   Fin

Fin.

Tableau 3.1 : Algorithme de détection des termes descriptifs (document/requête).

A l'issue de cette étape d'indexation, trois ensembles de termes sont identifiés:

(1) L'ensemble des expressions ξ_{Expres} correspondant aux collocations dans WordNet ;

(2) L'ensemble des mots simples $\xi_{Simples}$ ayant une entrée dans WordNet ;

(3) L'ensemble des mots orphelins ξ_{Orphel}.

3.3.3.2 Désambiguïsation des termes

Les collocations sont des expressions quasiment désambiguïsées et possèdent généralement un seul sens dans WordNet. Nous proposons de désambiguïser uniquement les mots simples de l'ensemble $\xi_{Simples}$. Un mot *m* de ξ_{Simple} peut avoir plusieurs sens dans WordNet. Le but de cette étape est de retrouver le sens correct de *m* suivant son contexte d'utilisation.

Nous définissons le contexte d'utilisation d'un mot selon deux hypothèses distinctes :

- La première (H1) est l'hypothèse du « *Multiple senses per discourse* » [Krovetz, 98], qui stipule qu'un mot peut être utilisé avec plusieurs sens différents dans un même document.

- La seconde (H2) est l'hypothèse du « *One sense per discourse* » [Gale et al., 91], qui stipule qu'un mot est généralement utilisé avec un seul sens dans un même document.

D'une part, en adoptant l'hypothèse (H1), nous supposons que les différentes occurrences m_i d'un mot *m* dans un même document peuvent avoir des sens différents. Ainsi, la phrase où l'occurrence m_i apparait, détermine son contexte d'utilisation. Nous définissons alors le contexte local de l'occurrence m_i (noté ζ_{L_i}) par l'ensemble des termes appartenant à la phrase de m_i. Pour retrouver le sens correct de chaque occurrence m_i, nous proposons de désambiguïser m_i dans son contexte local ζ_{L_i} par *une approche de désambiguïsation contextuelle locale*.

D'autre part, en adoptant l'hypothèse (H2), nous supposons que les différentes occurrences m_i d'un mot *m*, ayant une même forme syntaxique dans un document donné, portent toutes un sens identique. Nous définissons alors le contexte global de l'occurrence m_i de *m* (noté ζ_{G_i}) par l'union de tous les contextes locaux dans lesquels les instances m_i apparaissent avec une même catégorie syntaxique dans le document. Pour retrouver le sens correct de l'occurrence m_i dans le document, nous proposons deux approches de désambiguïsation comme suit:

1) *Une approche de désambiguïsation contextuelle globale*, qui consiste à identifier le sens adéquat de l'occurrence m_i de *m* dans son contexte global ζ_{G_i}.

2) *Une approche de désambiguïsation contextuelle mixte*, qui consiste à identifier le sens correct de l'occurrence m_i en appliquant successivement:

- Une désambiguïsation locale qui permet de retrouver les sens corrects de toutes les occurrences m_i du mot m dans leurs contextes locaux respectifs ζ_{L_i} en appliquant l'approche de désambiguïsation contextuelle locale.
- Une désambiguïsation globale qui permet de retrouver, à partir des différents sens locaux, le sens le plus utilisé dans le contexte global.

3.3.3.2.1 Désambiguïsation contextuelle locale

Notre approche de désambiguïsation contextuelle locale consiste à identifier le sens correct d'une occurrence m_i d'un mot m dans son contexte local ζ_{L_i}. Elle s'articule sur trois niveaux de désambiguïsation successifs:

1- le premier est un niveau de désambiguïsation syntaxique, qui consiste à déterminer la forme grammaticale de l'occurrence m_i dans son contexte local ζ_{L_i}.

2- le second est un niveau de désambiguïsation des domaines, qui permet d'identifier pour chaque occurrence m_i son domaine d'usage dans son contexte local ζ_{L_i}.

3- le troisième est un niveau de désambiguïsation des sens des mots. Le but est de sélectionner le sens adéquat de l'occurrence m_i dans son contexte local ζ_{L_i} en tenant compte de son domaine d'usage identifié à l'étape précédente.

a) Désambiguïsation syntaxique des mots

La désambiguïsation syntaxique de l'occurrence m_i d'un mot m est réalisée en appliquant l'étiqueteur syntaxique *Standford Pos Tagger*[23] pour identifier la *partie de discours* (*Part Of Speech* ou *POS*) de m_i dans son contexte local. L'objectif de cette étape, est de sélectionner dans WordNet les seuls sens de m_i liés à son *POS* qui seront examinés lors des prochains niveaux de désambiguïsation.

b) Désambiguïsation des domaines des mots

Une occurrence m_i définie dans son contexte local par une catégorie syntaxique (ou *POS*), possède un ou plusieurs sens dans *WordNet* dans cette catégorie. Ces sens sont étiquetés dans *WordNetDomains* par des labels de domaines (voir annexe). Un sens de m_i peut appartenir à un ou plusieurs domaines. Le domaine d'usage de l'occurrence m_i dans son contexte local est celui qui maximise ses proximités sémantiques, relativement à *WordNetDomains*, avec les autres domaines associés aux autres termes t_k de son contexte local.

[23] http://nlp.stanford.edu/software/tagger.shtml

Formellement :

$$D_j = \arg\max_j \left(\sum_{\substack{t_k \in \zeta_{L_i} \\ i \neq j}} \sum_{\substack{k \mid t_k \in \{\xi_{simples} \cup \xi_{E\,xpres}\} \\ 1 \leq k \leq |N_{D_k}|}} Sim\left(D_j, D_k\right) \right) \qquad [3.1]$$

Où :

- ζ_{L_i} est le contexte local de m_i ;
- D_i est le domaine d'usage de m_i dans ζ_{L_i} ;
- D_k est un domaine associé à un sens d'un terme t_k $\left(t_k \in \{\xi_{simples} \cup \xi E_{xpres}\}\right)$ de ζ_{L_i} ;
- $|N_{Dk}|$ est le nombre de domaines associés à t_k dans *WordNetDomains* ;
- $Sim\left(D_j, D_k\right)$ désigne la similarité (ou proximité) sémantique entre les domaines D_j et D_k.

La similarité entre les domaines D_j et D_k, est calculée à partir de la mesure de Wu-Palmer [Wu-Palmer, 94] adaptée à la hiérarchie *Top-Level* de *WordNetDomains* comme suit:

$$Sim\left(D_j, D_k\right) = \frac{2 * profondeur\ (D^*)}{profondeur\ (D_j) + profondeur\ (D_k)} \qquad [3.2]$$

Où :

- D^* est le domaine le plus spécifique qui subsume D_j et D_k dans la hiérarchie *Top-Level* dans *WordNetDomains* ;
- *profondeur(D*)* : est le nombre d'arcs entre la racine *Top-Level* et le domaine D^* dans WordNetDomains ;
- *profondeur(D_j)* : est le nombre d'arcs entre la racine *Top-Level* et le domaine D_j dans WordNetDomains;

Remarque

Les domaines associés aux synsets de m_i dans la hiérarchie *Factotum* (ex : *time, number, color,…etc*) ne sont pas considérés dans cette étape de désambiguïsation du fait qu'ils ne sont pas informatifs.

c) Désambiguïsation des sens des mots

A l'issue de l'étape précédente, le domaine d'usage D_j du mot m_i est identifié. Le mot m_i peut avoir plusieurs sens dans son domaine d'usage. Il faut alors le désambiguïser. Pour désambiguïser m_i dans son domaine D_j, nous associons à chaque sens (synset de WordNet)

$S_{i(j)}[k]$ associé à m_i dans le domaine D_j, un score basé sur la somme de ses similarités sémantiques avec les autres sens associés aux autres termes t_l de son contexte local dans leurs domaines respectifs. Le sens ayant le plus grand score est alors retenu comme sens adéquat de m_i dans son contexte local ζ_{L_i}. Formellement:

$$S_{i(j)}[k] = Arg\max_{k}\left(\sum_{\substack{l \mid t_l \in \zeta_{L_i} \\ l \neq i}} \sum_{1 \leq n \leq |S_{l(m)}|} Sim\left(S_{i(j)}[k], S_{l(m)}[n]\right)\right) \qquad [3.3]$$

Où :

- $S_{l(m)}[n]$ est le $n^{ième}$ synset de t_l $\left(t_l \in \{\xi_{simples} \cup \xi_{E_{xpres}}\}\right)$ appartenant à son domaine d'usage D_m ;

- $|S_{l(m)}|$ est le nombre de sens associés à t_l dans son domaine d'usage D_j ;

- $Sim\left(S_{i(j)}[k], S_{l(m)}[n]\right)$ est la similarité sémantique entre les concepts $S_{i(j)}[k]$ et $S_{l(m)}[n]$ calculée sur la base de la mesure de Resnik [Resnik, 99] (ou de toute autre mesure de similarité sémantique entre synsets de WordNet [Lesk, 86], [Lin, 1998]…)

3.3.3.2.2 Désambiguïsation contextuelle globale

Notre approche de désambiguïsation contextuelle globale consiste à retrouver le sens correct d'une occurrence m_i d'un mot m de $\xi_{Simples}$ dans son contexte globale ζ_{G_i}. Nous définissons le contexte global (ζ_{G_i}) d'une occurrence m_i d'un mot m comme l'ensemble de toutes les phrases où m_i possède une même *POS*.

Cette approche de désambiguïsation repose sur les mêmes étapes que l'approche de désambiguïsation locale, telle que définie en section précédente, seul le contexte de désambiguïsation considéré diffère (contexte global au lieu du contexte local). En particulier :

a) La désambiguïsation des domaines du mot m_i est basée sur le score suivant :

$$D_i = \arg\max_{j}\left(\sum_{\substack{t_k \in \zeta_{G_i} \\ i \neq j}} \sum_{\substack{k \mid t_k \in \{\xi_{simples} \cup \xi_{E_{xpres}}\} \\ 1 \leq k \leq |N_{D_k}|}} Sim\left(D_j, D_k\right)\right) \qquad [3.4]$$

La similarité entre les domaines D_j et D_k est définie comme précédemment (formule 3.2).

b) La désambiguïsation des sens du mot m_i dans son contexte d'usage est basée sur le score suivant :

$$S_{i(j)}[k] = Arg\max_{k}\left(\sum_{\substack{l \mid t_l \in \zeta_{G_i} \\ l \neq i}} \sum_{1 \leq n \leq |S_{l(m)}|} Sim\left(S_{i(j)}[k], S_{l(m)}[n]\right)\right) \qquad [3.5]$$

Où :

- $S_{l(m)}[n]$ est le $n^{ième}$ synset de t_l $\left(t_l \in \{\xi_{simples} \cup \xi_{E_{xpres}}\}\right)$ appartenant à son domaine d'usage D_m;

- $|S_{l(m)}|$ est le nombre de sens associés à t_l appartenant à son domaine respectif D_j ;

- $Sim\left(S_{i(j)}[k], S_{l(m)}[n]\right)$ est la similarité sémantique entre les concepts $S_{i(j)}[k]$ et $S_{l(m)}[n]$ ([Resnik, 99], [Lesk, 86], [Lin, 1998]…).

Le sens correct de m_i dans son domaine d'usage D_i est celui qui maximise sa proximité sémantique aux autres sens associés aux autres termes t_l de son contexte global ζ_{G_i} dans leurs domaines respectifs.

3.3.3.2.3 Désambiguïsation contextuelle mixte

Cette approche de désambiguïsation repose sur deux étapes successives de désambiguïsation : une première étape de désambiguïsation contextuelle locale et une désambiguïsation globale.

a) Désambiguïsation contextuelle locale

Cette étape consiste à identifier le sens correct de chaque occurrence m_i d'un mot *m*, dans son contexte local ζ_{L_i} (approche définie en section 3.3.3.2.1).

b) Désambiguïsation dans le contexte global

A l'issue de l'étape précédente, chaque occurrence m_i d'un mot *m* est associée à son sens (synset de WordNet) adéquat dans son contexte local. Ainsi, le mot *m* possède différents sens dans son contexte global (chaque contexte local d'une occurrence m_i de *m* définissant un sens éventuellement différent). Pour retrouver le sens correct du mot *m* dans son contexte d'utilisation global, nous attribuons à chaque sens S_j identifié pour une occurrence m_i de *m*, un score $Score\_S_j$ égal au nombre d'occurrences de ce sens dans le contexte global ζ_{G_i}. Le sens qui maximise ce score est alors retenu comme sens correct du mot *m* dans le document.

Formellement :

$$S_i = Arg \max_j \left(Score\_S_j \right) \qquad [3.6]$$

Où : $Score\_S_j$ est le nombre d'occurrences du sens S_j dans le contexte global.

3.3.3.3 Pondération des concepts

Une fois les concepts (collocations et sens des mots simples ; dans ce qui suit, nous choisissons de noter un concept par C^i) identifiés, il s'agit de leur affecter un poids numérique représentant leur degré d'importance, ou centralité, dans le document. Nous proposons deux approches de pondération basées sur une nouvelle définition de la centralité d'un concept.

3.3.3.3.1 Pondération *Ct-Ict*

Dans cette approche, nous définissons la centralité d'un concept C^i par la combinaison de deux facteurs : sa *centralité locale* et sa *centralité globale*.

- La *centralité locale d'un concept* reflète son importance dans le document. Un concept C^i est central localement dans un document *d* s'il est fréquent et pertinent. La fréquence d'occurrence du concept C^i est définie par $tf\ (C_i)$. Nous définissons sa pertinence sur la base de ses similarités sémantiques par rapport aux autres concepts dans le document *d*. On notera la centralité locale d'un concept C^i dans le document *d* par : $Ct(C^i, d)$.

  Formellement, la centralité locale du concept C^i dans *d* est définie par :

$$Ct\left(C^i, d\right) = \alpha \times tf\left(C^i\right) + (1-\alpha) \sum_{i \neq l} Sim\left(C^i, C^l\right) \qquad [3.7]$$

  Où :

  - α est un facteur de pondération qui permet de balancer la fréquence d'un concept par rapport à sa pertinence. Ce facteur sera fixé expérimentalement.
  - $tf(C^i)$ est la fréquence du concept C^i dans le document *d*.
  - $Sim(C^i, C^l)$ est une mesure de la similarité sémantique entre les concepts C^i et C^l, calculée sur la base de la mesure de Resnik [Resnik, 99] (ou toute autre mesure de similarité sémantique entre synsets de WordNet [Lesk, 86], [Lin, 1998]…).

  Définition: Un concept C^i est dit central dans le document *d* si sa centralité locale $Ct(C^i, d)$ est supérieure ou égale à un seuil fixé *s*.

- La *centralité globale d'un concept* reflète son importance globale dans l'ensemble des documents de la collection (ou corpus). On définit la centralité documentaire du concept C^i comme le nombre de documents de la collection dans lesquels C^i est central. Un concept C^i qui est central dans plusieurs documents n'est pas discriminant. Le pouvoir de discrimination de C^i est alors défini comme sa centralité documentaire inverse (notée Ict_i).

Formellement :

$$Ict_i = \log\left(\frac{N}{n_i}\right) \qquad [3.8]$$

Où:

- N est le nombre de documents dans la collection.

- n_i est le nombre de documents de la collection dans lesquels le concept C^i apparaît comme central.

Le poids $W(C^i,d)$ d'un concept C^i dans un document d est alors défini par la combinaison de sa centralité locale et de sa centralité globale comme suit:

$$W\left(C^i,d\right) = Ct\left(C^i,d\right) \times Ict_i \qquad [3.9]$$

3.3.3.3.2 Pondération *Tidf*

Dans cette approche, nous définissons la centralité d'un concept C^i par une mesure basée d'une part sur sa fréquence normalisée dans le document d, et d'autres parts par ses proximités sémantiques aux autres concepts centraux appartenant à d et leurs fréquences normalisées respectives.

Formellement, la centralité d'un concept C^i dans d est représenté par son poids $W(C^i,d)$ défini comme suit :

$$W\left(C^i,d\right) = \left(Tidf_i \times \sum_{i \neq l}\left(Sim\left(C^i,C^l\right) \times Tidf_l\right)\right) \qquad [3.10]$$

Où :

- $Sim(C^i, C^l)$ est une mesure de la distance sémantique entre les concepts C^i et C^l, calculée sur la base de la mesure de Resnik [Resnik, 99] (ou toute autre mesure de similarité sémantique entre synsets de WordNet [Lesk, 86], [Lin, 1998]…)

- $Tidf_i$ est la fréquence normalisée du concept C^i dans le document d. Elle est formalisée par :

$$Tidf_i = tf\left(C^i\right) \times idf_i = tf\left(C^i\right) \times \log\left(\frac{N}{n_i}\right) \qquad [3.11]$$

Remarques

- Les deux approches de pondération (*Ct-Ict* et *Tidf*) ainsi proposées permettent la pondération des concepts (collocations et sens des mots simples). Les termes orphelins sont pondérés par la mesure *tf*idf.*
- Dans le cas où la mesure de similarité utilisée $Sim(C^i, C^l)$ est basée sur la taxonomie (*is-a)* des noms et verbes de WordNet (telle que la mesure de Resnik [Resnik, 99], Lin [Lin, 98] , Leacok [Leacok et al., 98]), les adjectifs et les adverbes sont pondérés par la mesure classique *tf*idf.*

3.3.3.4 Construction de l'index

Notre objectif à travers cette étape est de construire l'index sémantique du document, basé sur le modèle vectoriel. Dans cet index, les concepts (collocations et sens des mots simples) sont représentés par les numéros de leurs synsets dans WordNet concaténés à leurs *POS* dans le document (*n* pour les noms, *v* pour les verbes, *a* pour les adjectifs et *r* pour les adverbes). A titre d'exemple, les concepts *human_knee#n#1* (1er sens de *human_knee* dans WordNet*)* et *develop#v#2* (2ème sens du verbe *develop*) seront respectivement représentés par leurs numéros *05504248n* et *01723296v*. Cette représentation par les numéros des synsets est plus riche qu'une représentation par les concepts [Gozalo et al., 98]. Elle permet d'une part de résoudre le problème de la synonymie des mots et d'autre part d'alléger la disparité des termes dans le processus de la recherche. En effet, deux concepts synonymes (ex : *human_knee#n#1* et *genu#n#1*) représentés dans l'index sémantique par un même numéro de synset, sont vus comme un seul terme lors de la recherche.

3.3.4 Illustration et discussion

Dans cette section, nous illustrons par l'exemple notre approche d'indexation sémantique proposée. Pour ce faire, nous l'appliquons sur un extrait du texte (Figure 3.2) du document *Arthroskopie.0013003.eng.abstr* de la collection Muchmore[24]. Nous focalisons en particulier sur la désambiguïsation puisque de sa précision dépend en grande partie la précision de l'indexation.

"The posterior cruciate ligament (PCL) is the strongest ligament of the human knee joint. Its origin is at the lateral wall of the medial femoral condyle and the insertion is located in the posterior part of the intercondylar area and this posterior cruciate ligament consists of multiple small fiber bundles."

Figure 3.2 : Extrait d'un document de Muchmore avec ses différents termes descriptifs.

[24] http://muchmore.dfki.de/

En appliquant l'algorithme de détection des termes, décrit en section 3.3.3 (Tableau 3.1), et en utilisant la liste préétablie des collocations de la ressource sémantique WordNet, nous identifions, à partir du document, les trois ensembles suivants :

$\xi_{Expres} = \{human\_knee, knee\_joint, fiber\_bundle\}$,

$\xi_{Orphel} = \{pcl, intercondylar\}$,

$$\xi_{Simples} = \left\{ \begin{array}{c} \text{posterior, cruciate, ligament, strong, origin, lateral, wall, medial, femoral,} \\ \text{condyle, insertion, locate, part, area, consist, multiple, small} \end{array} \right\}$$

Les collocations (*human_knee, knee_joint, fiber_bundle)* sont des expressions désambiguïsées et possèdent chacune un seul sens correspondant dans WordNet. L'étape suivante de désambiguïsation concernera uniquement les termes ambigus de $\xi_{Simples}$. La désambiguïsation permet de sélectionner pour chacun de ces termes son sens correct dans le document.

Dans ce qui suit, nous illustrons nos différentes approches de désambiguïsation proposées.

a) Désambiguïsation contextuelle locale

On rappelle que la désambiguïsation locale d'un terme de $\xi_{Simples}$, est réalisée pour chacune de ses occurrences dans son contexte local (ie. la phrase où elle apparait).

A titre exemple, le contexte local de mot *wall* est représenté par l'ensemble des termes non vides suivants :

$$\zeta_{L_{wall}} = \left\{ \begin{array}{l} \text{origin, lateral,medial, femoral, condyle, insertion,locate,part, area,} \\ \text{part, posterior,cruciate,ligament,consist,multiple, small, fiber\_bundes} \end{array} \right\}$$

La désambiguïsation d'une occurrence d'un mot dans son contexte local est basée sur la succession des trois étapes suivantes :

(1) désambiguïsation de la forme grammaticale de l'occurrence,

(2) désambiguïsation de son domaine d'usage,

(3) désambiguïsation des sens dans le domaine identifié.

La désambiguïsation de la forme grammaticale permet d'identifier la *POS* (partie du discours) d'un mot dans son contexte local. Le résultat de cette désambiguïsation est donné comme suit :

$$\xi_{Simples} = \left\{ \begin{array}{c} \text{posterior/JJ, cruciate/JJ, ligament/NN, strong/JJ, origin/NN, lateral/JJ,} \\ \text{wall/NN, medial/JJ, femoral/JJ, condyle/NN, insertionNN/, locate/VB,} \\ \text{part/NN, area/NN, consist/VB, multiple/JJ, small/JJ} \end{array} \right\}$$

La désambiguïsation du domaine d'usage permet d'identifier le domaine adéquat d'un mot de $\xi_{Simples}$ dans son contexte d'utilisation (local dans ce cas). Pour ce faire, nous identifions d'abord pour chaque terme de $(\xi_{Simples} \cup \xi_{Express})$, ses sens associés dans WordNet, puis les domaines relatifs à ces sens dans WordNetDomains. Les résultats obtenus sont présentés à travers les tableaux de la figure 3.3.

Pour trouver les domaines corrects des mots simples ambigus : *ligament, strong, part, medial, area, wall, origin, insertion, locate, consist, lateral* et *small,* dans leurs contextes locaux, nous appliquons les formules définies en section 3.3.3.2.1. Le domaine *factotum* et les domaines de sa hiérarchie (ex : *quality, acoustics*) associés à ces mots, ne sont pas concernés par cette étape.

Les résultats de la désambiguïsation des domaines de chaque mot ambigu de $\xi_{Simples}$ dans son contexte local sont donnés dans la figure 3.4. Dans cette figure, les domaines (en gras) et les synsets (grisés) représentent respectivement les domaines d'usage des termes dans leurs contextes et leurs synsets associés. Ces résultats montrent que la majorité des domaines identifiés représentent effectivement les domaines adéquats des mots dans leurs contextes locaux. C'est ainsi que par exemple les termes *wall, area* et *ligament* se sont vu assigner le domaine *anatomy* qui est le domaine le plus probable du texte indexé.

La désambiguïsation des sens d'un mot permet de sélectionner parmi ses sens associés dans son domaine identifié à l'étape précédente, le sens adéquat du terme dans son contexte local. Nous présentons à travers les tableaux de la figure 3.5 les résultats de la désambiguïsation des sens des mots basée sur la mesure de Resnik [Resnik, 99] (formule [3.3]).

La figure 3.5 montre que tous les *noms* ambigus du texte ont été correctement désambiguïsés. Cependant, nous signalons que les *adjectifs* : *strong, lateral, medial* et *small* n'ont pas été désambiguïsés. Ceci est dû au fait que la mesure de Resnik [Resnik, 99] utilisée dans la désambiguïsation est basée sur la taxonomie *(is-a)* des noms et verbes de WordNet et ne tient pas compte des relations sémantiques qui existent entre les adjectifs dans WordNet. Par ailleurs, les verbes *locate* et *consist* sont restés ambigus puisqu'ils ne sont pas liés sémantiquement dans la hiérarchie (*is*-a) des verbes dans WordNet. L'utilisation d'une mesure de similarité des concepts, autre qu'une mesure basée sur la hiérarchie (*is-a*) des noms et verbes de WordNet, telle que la mesure de Lesk [Lesk, 86] est susceptible de résoudre ce problème. Nous présentons dans la figure 3.6 les résultats de la désambiguïsation contextuelle locale basée sur la mesure de Lesk [Lesk, 86].

posterior	00136931-a : posterior#a#1:animals
cruciate	02290409-a : cruciate#a#1 : factotum
ligament	04990644-n : ligament#n#1 : anatomy
	03528343-n : ligament#n#2 : factotum
strong	02239657-a : strong#a#1 : quality
	02238035-a : strong#a#2 : factotum
	01462766-a : strong#a#3 : factotum
	01766728-a : strong#a#4 : factotum
	01764165-a : strong#a#5 : quality
	02436955-a : strong#a#6 : factotum
	02193170-a : strong#a#7 : factotum
	01901958-a : strong#a#8 : grammar
	01118836-a : strong#a#9 : factotum
	01023736-a : strong#a#10 : factotum
	00807924-a : strong#a#11 : factotum
part	13028617-n : part#n#1 : factotum
	08103697-n : par#n#2 : factotum
	05344775-n : par#n#3 : factotum
	03746082-n : part#n#4 : factotum
	05526011-n : part#n#5 : factotum
	00678112-n : part#n#6 : factotum
	08797461-n : part#n#7 : factotum
	05584351-n : part#n#8 : theatre
	12529266-n : part#n#9 : money
	00738055-n : part#n#10 : factotum
	06600579-n : part#n#11 : music
	04953910-n : part#n#12 : fashion
medial	00746822-a : medial#a#1 : factotum
	00326753-a : medial#a#2 : factotum
femoral	02614670-a : femoral#a#1 : anatomy
Area	07980485-n : area#n#1 : geography
	13687487-n : area#n#2 : factotum
	05646624-n : area#n#3 : factotum
	02641332-n :area#n#4 : factotum
	04921220-n : area#n#5 : anatomy
	04842376-n : area#n#6 : factotm
condyle	05157880-n : condyle#n#1 : anatomy
multiple	02140712-n : multiple#a#1 : factotum

wall	04369872-n : wall#n#1 : buildings
	03899176-n : wall#n#2 : military
	08876684-n : wall#n#3 : factotum
	04370607-n : wall#n#4 : factotum
	05284169-n : wall#n#5 : anatomy
	08876934-n : wall#n#6 : geology
	04370836-n : wall#n#7 : factotum
	13731862-n : wall#n#8 : factotum
origin	07989929-n : origin#n#1 : factotum
	04675241-n : origin#n#2 : biology
	06874664-n : origin#n#3 : factotum
	05654262-n :origin#n#4 : mathematics
	07610417-n : origin#n#5 : factotum
insertion	06308340-n : insertion#n#1 : factotum
	00306609-n : insertion#n#2 : factotum
locate	02220173-v : locate#v#1 : factotum
	02613751-v : locate#v#2 : factotum
	02266215-v : locate#v#3 : factotum
	00401714-v : locate#v#4 : politics
small	01343705-a : small#a#1 : quality
	01366545-a : small#a#2 : factotum
	02259250-a : small#a#3 : factotum
	01597253-a : small#a#4 : factotum
	01419050-a : small#a#5 : factotum
	01481659-a : small#a#6 : factotum
	01406192-a : small#a#7 : acoustics
	01503651-a : small#a#8 : factotum
	00845424-a : small#a#9 : factotum
	01416997-a : small#a#10 : factotum
consist	02574255-v : consist#v#1 : factotum
	02670036-v : consist#v#2 : factotum
	02578919-v : consist#v#3 : factotum
	02554853-v : consist#v#4 : factotum
lateral	02353094-a : lateral#a#1 : factotum
	00746517-a : lateral#a#2 : factotum
knee_joint	05254826-n : knee_joint#n#1 : anatomy
human_knee	05254826-n : human_knee#n#1 : anatomy
fiber_bundle	05161310-n: fiber_bundle#n#1: anatomy

Figure 3.3: Sens et domaines associés aux différents termes de l'exemple.

Contexte local #1 (Phrase 1)		
Termes ∈ $\xi_{Simples}$	**Domaines désambiguïsés+ Domaines des synsets ∈ l'hiérarchie factotum**	**Synsets ∈ (domaines désambiguïsés + domaines de la hiérarchie factotum)**
posterior	***animals***	00149484: posterior#a#1
cruciate	factotum	02452977: cruciate#a#1
ligament	***anatomy*** factotum	05232304 : ligament#n#1
		03624146 : ligament#n#2
strong	***grammar*** quality factotum	02399006: strong#a#1
		01567822: strong#a#2
		01892526: strong#a#3
		01889809: strong#a#4
		02609394: strong#a#5
		02352092: strong#a#6
		02032599: strong#a#7
		01205631 : strong#a#8
		01109960: strong#a#9
		00881015: strong#a#10

Contexte local #2 (Phrase 2)		
Termes ∈ $\xi_{Simples}$	**Domaines désambiguïsés+ Domaines des synsets ∈ l'hiérarchie factotum**	**Synsets ∈ (domaines désambiguïsés + domaines de la hiérarchie factotum)**
origin	factotum ***biology***	08389765: origin#n#1
		04868433: origin#n#2
		07223696: origin#n#3
		07991959: origin#n#5
lateral	factotum	02523581: lateral#a#1
		00816785: lateral#a#2
wall	factotum ***anatomy***	09338002: wall#n#3
		04493140: wall#n#4
		05534758: wall#n#5
		04493369: wall#n#7
		14371846: wall#n#8
medial	factotum	00817090 : medial#a#1
		00353522: medial#a#2
femoral	***anatomy***	02812026: femoral#a#1
condyle	***anatomy***	05404114: condyle#n#1
insertion	factotum	06631667: insertion#n#1
		00317253: insertion#n#2
locate	factotum ***politics***	02264659: locate#v#1
		02668919: locate#v#2
		02311450: locate#v#3
		00409546: locate#v#4
posterior	***animals***	00149484: posterior#a#1
part	factotum *fashion*	13628130: part #n#1
		08509674: part#n#2
		05601888: part#n#3
		03849293: part#n#4
		05793444 : part#n#5
		00709615 : part#n#6
		09251280 : part#n#7
		00775927 : part#n#10
		05193808 : part#n#12
area	factotum ***anatomy***	14321345: area#n#2
		05922776: area#n#3
		02710247 : area#n#4
		05160192: area#n#5
		05068199: area#n#6
ligament	***anatomy*** factotum	05232304 : ligament#n#1
		03624146 : ligament#n#2
consist	factotum	02627955: consist#v#1
		02726166: consist#v#2
		02632727: consist#v#3
		02607974: consist#v#4
multiple	factotum	02292011: multiple#a#1
small	quality ***acoustics*** factotum	01443454: small#a#1
		01467170: small#a#2
		02419704: small#a#3
		01708858: small#a#4
		01588010: small#a#5
		01508124: small#a#6
		01610884: small#a#7
		00921344: small#a#8
		01520244: small#a#9
		2309723: small#a#10

Figure 3.4 : Résultats de la désambiguïsation du domaine d'usage de chaque terme de $\xi_{Simples}$ dans son contexte local.

Contexte local #1 (Phrase 1)	
Terme	**Sens des termes après désambiguïsation**
posterior	00149484 : posterior#a#1
cruciate	02452977 : cruciate#a#1
ligament	05232304 : ligament#n#1
strong	02399006 : strong#a#1
	01567822 : strong#a#2
	01892526 : strong#a#3
	01889809 : strong#a#4
	02609394 : strong#a#5
	02352092 : strong#a#6
	02032599 : strong#a#7
	01205631 : strong#a#8
	01109960 : strong#a#9
	00881015 : strong#a#10
human_knee	05504248 : human_knee#n#1
knee_joint	05504248 : knee_jointe#n#1

Contexte local #2 (Phrase 2)	
Terme	**Sens des termes après désambiguïsation**
origin	08389765 : origin#n#1
lateral	02523581 : lateral#a#1
	00816785 : lateral#a#2
wall	05534758 : wall#n#5
medial	00817090 : medial#a#1
	00353522 : medial#a#2
femoral	02812026 : femoral#a#1
condyle	05404114 : condyle#n#1
insertion	00317253 : insertion#n#2
locate	02264659 : locate#v#1
	02668919 : locate#v#2
	02311450 : locate#v#3
	00409546 : locate#v#4
posterior	00149484 : posterior#a#1
part	09251280 : part#n#7
area	05160192 : area#n#5
ligament	05232304 : ligament#n#1
consist	02627955 : consist#v#1
	02726166 : consist#v#2
	02632727 : consist#v#3
	02607974 : consist#v#4
multiple	02292011 : multiple#a#1
small	01443454 : small#a#1
	01467170 : small#a#2
	02419704 : small#a#3
	01708858 : small#a#4
	01588010 : small#a#5
	01508124 : small#a#6
	01610884 : small#a#7
	00921344 : small#a#8
	01520244 : small#a#9
	02309723 : small#a#10
fiber_bundle	05407769 : fiber_bundle#n#1

Figure 3.5 : Résultats de la désambiguïsation contextuelle locale de l'exemple de la figure 3.2, basée sur la similarité de Resnik [Resnik, 99].

Contexte local #1 (Phrase 1)	
Terme	**Sens des termes après désambiguïsation**
posterior	00149484 : posterior#a#1
cruciate	02452977 : cruciate#a#1
ligament	05232304 : ligament#n#1
strong	02352092 : strong#a#6
human_knee	05504248 : human_knee#n#1
knee_joint	05504248 : knee_jointe#n#1

Contexte local #2 (Phrase 2)	
Terme	**Sens des termes après désambiguïsation**
origin	07223696: origin#n#3
lateral	00816785 : lateral#a#2
wall	09338002 : wall#n#3
medial	00817090 : medial#a#1
femoral	02812026 : femoral#a#1
condyle	05404114 : condyle#n#1
insertion	06631667 : insertion#n#1
locate	02311450 : locate#v#3
posterior	00149484 : posterior#a#1
part	03849293 : part#n#4
area	05160192 : area#n#5
ligament	05232304 : ligament#n#1
consist	02627955 : consist#v#1
multiple	02292011 : multiple#a#1
small	01443454 : small#a#1
fiber_bundle	05407769 : fiber_bundle#n#1

Figure 3.6 : Résultats de la désambiguïsation contextuelle locale de l'exemple de la figure 3.2 basée sur la similarité de Lesk [Lesk, 86].

Nous remarquons à travers ces résultats que la désambiguïsation locale basée sur la mesure de Lesk [Lesk, 99] a permis de désambiguïser tous les termes ambigus de l'exemple. Cependant, contrairement à la mesure de Resnik [Resnik, 99] certains *noms* n'ont pas été correctement désambiguïsés. A titre d'exemple, le mot *wall* est désambiguïsé par le concept *wall#n#3* défini dans WordNet par : **«** *anything that suggests a wall in structure or function or effect; "a wall of water"; "a wall of smoke"; "a wall of prejudice"; "negotiations ran into a brick wall"* **»** en appliquant la mesure de Lesk [Lesk, 86] dans la désambiguïsation locale, alors que l'utilisation de la mesure de Resnik [Resnik, 99] lui associe le concept *wall#n#5*,défini dans WordNet par : *«(anatomy) a layer (a lining or membrane) that encloses a structure; "stomach walls"* **»**. En examinant le contexte local de *wall*, il nous semble qu'il est plus adéquat de l'associer au concept *wall#n#5*. Par conséquent, on peut déduire que la qualité de notre approche de désambiguïsation locale dépend en partie de la mesure de similarité utilisée dans le score désambiguïsation de la formule [3.3].

b) Désambiguïsation contextuelle globale

Après avoir déterminé le POS de chaque occurrence de mot simple dans son contexte local, nous déterminons le contexte global d'un mot simple comme l'ensemble des mots $(\xi_{Simples} \cup \xi_{Expres})$ issus des contextes locaux associés à toutes ses occurrences appartenant à la même *POS*. Par exemple, les contextes globaux des mots « *ligament* » et « *wall*» sont données comme suit:

$$\zeta_{G_{ligament}} = \left\{ \begin{array}{c} \text{posterior, cruciate,strong, origin,lateral,wall,medial,femoral, condyle, insertion,locate,} \\ \text{part, area,consist,multiple,small,human\_knee, knee\_joint, fiber\_bunde} \end{array} \right\}$$

$$\zeta_{G_{wall}} = \left\{ \begin{array}{l} \text{origin, lateral, medial, femoral, condyle, insertion, locate, part, area,} \\ \text{part, posterior, cruciate, ligament, consist, multiple, small, fiber\_bundles} \end{array} \right\}$$

Terme	**Domaines désambiguïsés + Domaines des synsets ∈ l'hiérarchie factotum**	**Les sens des mots de WordNet ∈ (domaines désambiguïsés + domaines de l'hiérarchie factotum)**
posterior	***animals***	00149484: posterior#a#1
cruciate	factotum	02452977: cruciate#a#1
ligament	***anatomy*** factotum	05232304 : ligament#n#1
		03624146 : ligament#n#2
strong	*ùfactotum ***grammar***	02399006: strong#a#1
		01567822: strong#a#2
		01892526: strong#a#3
		01889809: strong#a#4
		02609394: strong#a#5
		02352092: strong#a#6
		02032599: strong#a#7
		01205631 : strong#a#8
		01109960: strong#a#9
		00881015: strong#a#10
part	factotum ***theatre***	13628130: part #n#1
		08509674: part#n#2
		05601888: part#n#3
		03849293: part#n#4
		05793444 : part#n#5
		00709615 : part#n#6
		09251280 : part#n#7
		00775927 : part#n#10
		05193808 : part#n#12
lateral	factotum	02523581: lateral#a#1
		00816785: lateral#a#2
medial	factotum	00817090 : medial#a#1
		00353522: medial#a#2
femoral	***anatomy***	02812026: femoral#a#1
area	factotum ***anatomy***	14321345: area#n#2
		05922776: area#n#3
		02710247 : area#n#4
		05160192: area#n#5
		05068199: area#n#6
condyle	***anatomy***	05404114: condyle#n#1
multiple	factotum	02292011: multiple#a#1

Terme	**Domaines désambiguïsés + Domaines des synsets ∈ l'hiérarchie factotum**	**Les sens des mots de WordNet ∈ (domaines désambiguïsés + domaines de l'hiérarchie factotum)**
wall	factotum ***anatomy***	09338002: wall#n#3
		04493140: wall#n#4
		05534758: wall#n#5
		04493369: wall#n#7
		14371846: wall#n#8
insertion	factotum	06631667: insertion#n#1
		00317253: insertion#n#2
locate	factotum ***politics***	02264659: locate#v#1
		02668919: locate#v#2
		02311450: locate#v#3
		00409546: locate#v#4
origin	factotum ***biology***	08389765: origin#n#1
		04868433: origin#n#2
		07223696: origin#n#3
		07991959: origin#n#5
small	quality ***acoustics*** factotum	01443454: small#a#1
		01467170: small#a#2
		02419704: small#a#3
		01708858: small#a#4
		01588010: small#a#5
		01508124: small#a#6
		01610884: small#a#7
		00921344: small#a#8
		01520244: small#a#9
		2309723: small#a#10
consist	factotum	02627955: consist#v#1
		02726166: consist#v#2
		02632727: consist#v#3
		02607974: consist#v#4
human_knee	anatomy	05504248 : human_knee#n#1
knee_joint	anatomy	05504248 : knee_joint#n#1
fiber_bundle	anatomy	05407769 : fiber_bundle#n#1

Figure 3.7 : Résultats de la désambiguïsation contextuelle globale de l'exemple, basée sur la mesure de Resnik [Resnik, 99].

Suit alors la désambiguïsation des domaines des mots (*ligament, strong, part, medial, area, wall, origin, insertion, locate, consist, lateral* et *small*) dans le contexte global du mot considéré. Finalement, les sens associés aux domaines identifiés, en appliquant la formule [3.5], pour chaque mot sont ensuite désambiguïsés. Les résultats de la désambiguïsation globale pour notre exemple sont donnés à travers les tableaux de la figure 3.7. Les domaines (en gras) représentent les domaines d'usage identifiés lors de la désambiguïsation des domaines des mots et les synsets (grisés) présentent les sens adéquat des mots dans leur contexte global (i.e dans le document).

A travers ces résultats, nous constatons que:

(1) la désambiguïsation au niveau des domaines a permis d'associer pour la majorité des mots leurs domaines corrects dans le document (ex : *wall, ligament, …*).

(2) la désambiguïsation au niveau des sens a permis de retrouver pour tous les *noms* leurs sens adéquats dans le document. Cependant, les *adjectifs* et les deux verbes *consist* et *locate* n'ont pas été désambiguïsés. La cause est probablement liée à la mesure de similarité utilisée (tel que mentionné précédemment).

c) Désambiguïsation mixte

La désambiguïsation mixte s'effectue selon une première désambiguïsation locale des sens des mots, suivie d'une désambiguïsation globale de ces sens (décrite en section 3.3.3.2.3). Les résultats de la désambiguïsation locale (illustrée précédemment), en appliquant la mesure de Resnik [Resnik,99] dans la formule [3.3], sont donnés à travers les tableaux de la figure 3.5.

Nous remarquons dans les résultats de la figure 3.5, que seul le terme *ligament* possède deux contextes locaux. La désambiguïsation sémantique de chacune de ses occurrences respectivement dans son contexte locale, lui a associée le synsets *ligament#n#1*. Par conséquent, on déduit que le sens du mot *ligament* dans son contexte global (i.e l'union de ses contextes locaux) est *ligament#n#1* vu que sa fréquence est majoritaire (2 occurrences) dans ce contexte (global). De ce fait, sa désambiguïsation dans le contexte global lui assigne son premier sens dans WordNet, qui correspond effectivement au sens attendu dans le document. Pour les autres termes, Les résultats de leur désambiguïsation respective dans leurs contextes globaux sont identiques à ceux obtenus par leur désambiguïsation dans leurs contextes locaux, puisque le contexte global et le contexte local de chacun de ces termes sont identiques.

Discussion

a. *Autour des approches de désambiguïsation proposées :* Les différentes approches de désambiguïsation (globale, locale et mixte) illustrées sur l'exemple de la figure 3.2 ont permis d'identifier tous les sens corrects des *noms* dans leur contexte d'utilisation lorsque la mesure de Resnik [Resnik, 99] est utilisée. Cependant, les adjectifs n'ont pas été désambiguïsés du fait que cette mesure ne tient pas compte des relations sémantiques entre les adjectifs dans WordNet. Ce problème a été résolu avec la mesure de Lesk [Lesk 86].

Ainsi, on peut déduire que la qualité des résultats obtenus dépend fortement de la mesure de similarité utilisée dans ces approches.

b. *Autour de notre approche de désambiguïsation des domaines* : A ce niveau, nous tentons de répondre à deux questions :

1. la première : « *qu'apportent les domaines dans le choix des sens corrects des termes dans leur contexte d'utilisation ?* »

2. la seconde question : « *qu'apporte notre approche de désambiguïsation par les domaines ?*»

Pour répondre à ces questions, nous avons procédé à deux comparaisons à travers un exemple illustratif sur la phrase suivante : *'The virus infected all files on the hard disk.'*

1. La première comparaison concerne notre approche de désambiguïsation basée sur les domaines de la phrase (locale) et une approche de désambiguïsation classique (sans les domaines) [Baziz et al., 05a]. Les sens des termes dans l'approche de [Baziz et al., 05a] sont désambiguïsés en se basant uniquement sur un score de similarité sémantique entre un concept candidat du mot cible et l'ensemble des concepts des autres termes appartenant au document. Le concept qui maximise ce score est retenu comme sens adéquat du terme dans le document.

2. La seconde comparaison concerne notre approche de désambiguïsation par les domaines basée sur les domaines de la phrase (locale) et une autre approche de désambiguïsation par les domaines proposée dans [Kolte et al, 09]. C'est cette dernière approche qui nous avait inspiré dans nos travaux sur la désambiguïsation par les domaines. A la différence de la notre, l'approche de [Kolte et al, 09] calcule le domaine correct d'un mot sur la base de sa fréquence d'occurrence dans le contexte local.

- ***Comparaison 1 :*** Notre approche *vs* Approche de Baziz [Baziz et al., 05a] . Les résultats de cette comparaison sont présentés dans les tableaux de la figure 3.8.

Dans cet exemple, il apparaît que les trois mots suivants : *virus*, *infect* et *file*, sont ambigus. Les synsets (grisés) de la figure 3.8 représentent les sens désambiguïsés de ces mots. De ces résultats, il ressort que notre approche de désambiguïsation basée sur les domaines des concepts a permis de trouver les sens corrects des deux mots *virus* et *file*. Néanmoins, aucun de ces trois termes ambigus n'a été correctement désambiguïsé par l'approche de [Baziz et al., 05a]. Ainsi, il semblerait que l'identification des domaines d'usage a permis de supprimer les synsets qui ne sont pas liés au contexte d'utilisation.

A titre d'exemple, le mot *file* possède 4 sens dans WordNet définies par :

1. (17) file, data file -- (a set of related records (either written or electronic) kept together)
2. (1) file, single file, Indian file -- (a line of persons or things ranged one behind the other)
3. (1) file, file cabinet, filing cabinet -- (office furniture consisting of a container for keeping papers in order)
4. (1) file -- (a steel hand tool with small sharp teeth on some or all of its surfaces; used for smoothing wood or metal)

Résultats obtenus par l'approche de désambiguïsation contextuelle locale		
Terme	**Domaines des termes dans WordNetDomains**	**Sens des Termes désambiguïsés**
virus	*factotum*	*01312417 : virus#n#1*
	factotum	*13821730 : virus#n#2*
	computer_science	*06498276 : virus#n#3*
infect	*factotum*	*06422370 : infect#v#1*
	medecine	*08314378 : infect#v#2*
	factotum	*03301857 : infect#v#3*
	psychological_features	*03301556 : infect#v#4*
file	***telecommunication***	*06422370 : file#n#1*
	factotum	*08314378 : file#n#2*
	administration furniture	*03301857 : file#n#3*
	buildings industry	*03301556 : file#n#4*
hard_disk	***computer_science***	*03454622 : hard_disk#n#1*

Résultats obtenus par l'approche de Baziz	
Terme	**Sens des Termes désambiguïsés**
virus	*01312417 : virus#n#1*
	13821730 : virus#n#2
	06498276 : virus#n#3
infect	*06422370 : infect#v#1*
	08314378 : infect#v#2
	03301857 : infect#v#3
	03301556 : infect#v#4
file	*06422370 : file#n#1*
	08314378 : file#n#2
	03301857 : file#n#3
	03301556 : file#n#4
hard_disk	*03454622 : hard_disk#n#1*

Figure 3.8 : Notre désambiguïsation locale *vs* désambiguïsation [Baziz et al., 05a].

Ces sens sont étiquetés dans WordNetDomains par leurs domaines respectifs représentés dans la figure 3.8 (ex : *telecommunication* pour *file#n#1*, *factotum* pour *file#n#2*). La désambiguïsation des domaines du mot *file* et des autres mots de son contexte a permis de garder uniquement les synsets appartenant à ces domaines d'usage (ex : le domaine d'usage de *file* est *telecommunication*). Lors de la désambiguïsation de ces synsets le mot *file* est désambiguïsé par le synset *file#n#1*. En examinant le terme *file* dans son contexte d'utilisation et sa définition dans WordNet, il semble que ce synset représente effectivement le sens correct de *file* dans l'exemple.

- ***Comparaison 2 :*** Notre approche vs Approche de Kolte [Kolt et al., 09]. Les résultats de cette comparaison sont présentés dans les tableaux de la figure 3.9.

Les résultats rapportés en figure 3.9 montrent que notre approche de désambiguïsation des domaines a permis de sélectionner pour les mots *virus* et *file*, leurs domaines d'usage les plus probables dans leur contexte d'utilisation, soit respectivement *computer science* et *telecommunication*. Cependant l'approche de désambiguïsation des domaines proposé dans [Kolte et al., 09] a sélectionné les domaines d'usage *telecommunication, administration, furniture , buildings* et *industry*. La désambiguïsation des sens de *file* associés à ces domaines n'a pas permis d'identifier son sens adéquat. Ainsi, nous concluons que l'utilisation des relations sémantiques entre domaines (dans notre approche de désambiguïsation des domaines) a permis d'améliorer les résultats de la désambiguïsation des sens mots dans cet exemple.

Résultats obtenus par l'approche de désambiguïsation contextuelle locale		
Terme	**Domaines des termes dans WordNetDomains**	**Sens des Termes désambiguïsés**
virus	*factotum*	*01312417 : virus#n#1*
	factotum	*13821730 : virus#n#2*
	computer_science	*06498276 : virus#n#3*
infect	*factotum*	*06422370 : infect#v#1*
	medecine	*08314378 : infect#v#2*
	factotum	*03301857 : infect#v#3*
	psychological_features	*03301556 : infect#v#4*
file	***telecommunication***	*06422370 : file#n#1*
	factotum	*08314378 : file#n#2*
	administration furniture	*03301857 : file#n#3*
	buildings industry	*03301556 : file#n#4*
hard_ disk	***computer_science***	*03454622 : hard_disk#n#1*

Résultats obtenus par l'approche de Kolte		
Terme	**Domaines des termes dans WordNetDomains**	**Sens des Termes désambiguïsés**
virus	*factotum*	*01312417 : virus#n#1*
	factotum	*13821730 : virus#n#2*
	computer_science	*06498276 : virus#n#3*
infect	*factotum*	*06422370 : infect#v#1*
	medecine	*08314378 : infect#v#2*
	factotum	*03301857 : infect#v#3*
	psychological_features	*03301556 : infect#v#4*
file	***telecommunication***	*06422370 : file#n#1*
	factotum	*08314378 : file#n#2*
	administration furniture	*03301857 : file#n#3*
	buildings industry	*03301556 : file#n#4*
hard_ disk	***computer_science***	*03454622 : hard_disk#n#1*

Figure 3.9: Notre désambiguïsation locale *vs* désambiguïsation de [Kolte et al., 09].

Ces différents résultats nous confortent dans l'idée que l'exploitation des domaines et des liens sémantiques entre domaines, permet de renforcer la désambiguïsation sémantique des sens d'un mot dans son contexte.

3.4 Appariement Document-Requête

L'une des étapes clés dans le processus de RI est l'appariement document-requête. Il consiste à mettre en correspondance la représentation de la requête avec les représentations des documents. Cette correspondance est mesurée par un score de pertinence associé au document, qui reflète son degré de ressemblance ou de similarité avec la requête. Dans le modèle de recherche vectoriel (le plus simple et le plus intuitif des modèles de RI), documents et requêtes sont représentés par des vecteurs de termes pondérés. L'appariement est alors basé sur une mesure de similarité entre les vecteurs correspondants, généralement calculée comme le cosinus [Salton et al., 83] entre les deux vecteurs. Cette mesure est lexicale et dépend des termes communs au document et à la requête. Ceci pose un problème crucial : Un document pertinent qui ne partage aucun terme avec la requête n'est pas retourné par le système.

Nous proposons de pallier à ce problème en étendant la mesure du cosinus pour tenir compte des liens sémantiques existants entre les termes des documents et de la requête. Ainsi, le score de pertinence d'un document dépendra non seulement des poids des termes communs mais aussi des similarités sémantiques entre les concepts de la requête et du document.

Formellement :

$$RSV\ (D_i, Q) = \frac{\sum_{j=1}^{n} w_{Qj} \times w_{ij} + \sum_{k=1}^{m}\sum_{l=1}^{g} Sim\,(C_{Qk}, C_{il})}{\sqrt{\sum_{j=1}^{n} w_{Qj}^2} \times \sqrt{\sum_{j=1}^{n} w_{ij}^2}} \qquad [3.12]$$

Où :

- w_{Qj} et w_{ij} représentent respectivement le poids d'un terme t_j (orphelin ou concept) dans la requête Q et son poids dans le document d_i, mesurés par l'un des schémas proposés (*Ct-Ict* ou *Tidf*)
- $Sim(C_{Qk}, C_{ik})$ est la similarité entre le concept C_{Qk} de la requête Q et le concept C_{ik} du document D_i.

3.5 Conclusion

Nous avons présenté au cours de ce chapitre les fondements théoriques d'un nouveau modèle de RI sémantique basé sur l'utilisation conjointe de la ressource lexicographique WordNet et de son extension WordNetDomains. Nos contributions ont porté sur deux aspects : (1) une proposition d'une nouvelle approche d'indexation sémantique des documents et requêtes, (2) et une proposition d'une approche d'évaluation sémantique des requêtes.

(1) L'approche d'indexation sémantique a pour objet de résoudre les problèmes causées par l'indexation classique basée mots-clés, en représentant les contenus des documents (ou respectivement des requêtes) par des descripteurs sémantiques. Ces descripteurs, ou index, sont composés de concepts pondérés qui reflètent les sens adéquats des mots des index classiques dans leur contexte d'utilisation. Les mots simples ou composés de l'indexation classique sont extraits en projetant les textes des documents (ou des requêtes) sur la ressource linguistique WordNet, puis assignés par leurs sens (concepts) en appliquant l'une des techniques de désambiguïsation contextuelle proposées (locale, globale ou mixte). Ces techniques de désambiguïsation s'appuient sur WordNet et son extension aux domaines WordNetDomains comme sources d'évidence. Finalement, les concepts identifiés sont pondérés par leur centralité et organisés dans des index sémantiques.

(2) L'approche d'évaluation sémantique des requêtes consiste à interpréter les index sémantiques, issus de notre approche d'indexation, par des vecteurs de concepts pondérés dans le but de pouvoir calculer la pertinence des documents pour les requêtes. Dans cette approche, nous avons proposé un score de pertinence sémantique qui considère les liens sémantiques latents entre les concepts respectifs d'un document et d'une requête afin que le SRI puisse récupérer les documents sémantiquement pertinents mêmes s'ils ne contiennent aucun terme des requêtes.

Pour mesurer les performances de ces approches, nous proposons dans le chapitre suivant l'évaluation expérimentale de notre modèle de RI sémantique.

Chapitre 4

Evaluation expérimentale

Plan du chapitre

4.1 Introduction 86

4.2 Environnement technologique 86

4.3 Protocole d'évaluation 87

4.3.1 La collection TIME 89

4.3.2 La collection Muchmore 90

4.4. Evaluation avec la collection TIME 92

4.4.1 Evaluation de l'approche d'indexation sémantique dans TIME 92

4.4.2 Evaluation des approches de pondération des concepts dans TIME 112

4.4.3 Evaluation de la mesure sémantique du score d'appariement documents-requête dans TIME 126

4.5 Evaluation avec la collection médicale Muchmore 130

4.5.1 Evaluation de l'approche d'indexation sémantique dans Muchmore 130

4.5.2 Evaluation des approches de pondération des concepts dans Muchmore 131

4.6 Conclusion 134

4.1 Introduction

Nous avons évalué notre approche de RI sémantique dans deux cadres. Le premier cadre, en utilisant une collection de test issue d'un domaine général, la collection TIME[25], et le second cadre, en utilisant une collection du domaine médical, la collection Muchmore[26]. L'objectif est de mesurer la qualité de notre modèle de RI sémantique.

Ce chapitre est organisé comme suit : en section 4.2, nous présentons l'environnement technologique de développement de notre SRI sémantique. En section 4.3, nous décrivons le protocole d'évaluation utilisé. L'évaluation de notre approche de RI sémantique en utilisant la collection de test TIME est présentée en section 4.4. Enfin, la section 4.5 est dédiée à l'évaluation de notre approche avec la collection médicale Muchmore.

4.2 Environnement technologique

Nous avons implémenté en Java (version 1.6) les différents modules de notre approche de RI sémantique correspondants à nos différentes propositions présentées dans le chapitre précédent. Le choix de Java est motivé par la disponibilité et la gratuité de diverses APIs java qui nous ont servis à la réalisation et à l'évaluation de notre approche proposée, en voici la liste :

- ***L'API Stanford POS Tagger[27] :*** est utilisée pour l'analyse syntaxique du contenu textuel des documents (et requêtes,). Cette API nous a permis de récupérer la forme grammaticale de chaque mot dans le texte.

- ***L'API JWNL[28] (Java WordNet Library):*** est utilisée pour accéder à la base lexicographique WordNet. Cette API nous a permis de lemmatiser les mots du texte en se basant sur les relations morphologiques d'un mot dans WordNet d'une part et d'exploiter les principales relations entre concepts de WordNet, telles que la synonymie et la taxonomie (*is-a*) des noms et verbes dans WordNet, d'autre part.

- ***L'API JWS[29] (Java WordNet Similarity) :*** offre la possibilité d'évaluer la similarité sémantique entre deux concepts dans WordNet. Cette API implémente 10 mesures de similarités sémantiques dont les mesures suivantes : la mesure de Resnik [Resnik et al, 99], la mesure de Lin [Lin, 89], la meure de Wu-Palmer [Wu-Palmer, 94], la mesure de Jiang-Conrath [Jiang et al, 97], …etc.

[25] https ://isserver11.princeton.edu/
[26] http://muchmore.dfki.de/
[27] http://nlp.stanford.edu/software/tagger.shtml/
[28] http://jwordnet.sourceforge.net/handbook.html
[29] http://www.sussex.ac.uk/Users/drh21/

- ***L'API d'évaluation de la plateforme de RI Terrier***[30] **:** fournit, dans Terrier [Ounis et al., 06], une implémentation du programme *trec_eval*[31] utilisé pour évaluer les résultats d'un SRI selon les mesures d'évaluation TREC. Cette API prend en entrée deux fichiers : le fichier (.res) contenant les résultats de la recherche pour l'ensemble de requêtes utilisées et le fichier des jugements de pertinence fourni par la campagne d'évaluation associés à ces requêtes.

Le format du fichier (.res) des résultats de la recherche est le suivant :

Num_requête id_document_selectionné rang score

Un exemple de fichier (.res) est donné dans ce qui suit :

```
Num_requête Q0    id_Doc  rang   score
1           Q0    320.txt  0      15.969306142072558
1           Q0    390.txt  1      14.101577090590148
1           Q0    418.txt  2      13.121560879850819
2           Q0    434.txt  0      16.285063082159752
2           Q0    464.txt  1      14.82533730236604
```

A l'issue de l'évaluation des résultats de la recherche, Terrier renvoie en sortie un fichier de valeurs (.val) pour les différentes mesures d'évaluation réalisées telles : les précisions à différents points *x* (*P@x, x=1, 2, 3, 4, 5, 10, 15, 20, 30, …,1000),* la moyenne des précisions moyennes obtenues par l'ensemble des requêtes à chaque fois qu'un document pertinent est retrouvé (*MAP*), le nombre de documents sélectionnés pertinents, …etc.

4.3 Protocole d'évaluation

Pour évaluer nos propositions en RI sémantique, nous les avons implémentées à travers un système de recherche sémantique basé sur le modèle vectoriel. Dans ce système, les index sémantiques des documents et requêtes, assimilés à des vecteurs de concepts pondérés, sont comparés lors de la recherche à travers notre mesure (décrite en section 3.4) d'évaluation sémantique des requêtes.

En particulier, nous avons implémenté :

- Notre approche (avec différentes variantes) d'indexation sémantique par les concepts, incluant :
  - Notre approche d'identification des collocations.
  - Nos trois approches de désambiguïsation des sens des mots.
- Nos deux approches de pondération des concepts,
- Notre approche d'évaluation sémantique des requêtes.

[30] http://terrier.org/

[31] http://trec.nist.gov/trec_eval/

Dans ces implémentations, nous avons utilisé la mesure de Resnik [Resnik, 99] pour calculer les similarités sémantiques entre concepts dans les étapes concernées.

L'évaluation est faite selon le protocole TREC en utilisant le module d'évaluation *Trec-eval* de la plateforme de RI *Terrier*. Pour chaque requête, Terrier examine les 100 premiers documents restitués par le système de recherche sémantique et renvoie les précisions *P@x* à différents points *x* (*x*=5, 10, 15, 20, 30, 50, 100), ainsi que la moyenne arithmétique des précisions moyennes des requêtes utilisées dans la recherche (*MAP)*.

La précision au point *x* (*P@x)*, est le ratio des documents pertinents parmi les *x* premiers documents restitués. Formellement:

$$P@x = \frac{Nbre(doc_{Perts})_x}{x}$$

Où :

- $Nbre(doc_{Perts})_x$ est le nombre de documents pertinents parmi les *x* premiers documents sélectionnés par le système.

La *MAP* (*Mean Average Precision*) est la moyenne des précisions moyennes *AP* calculées pour l'ensemble des requêtes. Formellement:

$$MAP = \frac{1}{|Q|}\sum_{q=1}^{Q} AP_q$$

Où :

- *Q* est l'ensemble des requêtes utilisées dans la recherche,
- |*Q*| est le nombre de requêtes,
- AP_q est la moyenne des précisions calculées aux différents rangs *k* où un document pertinent est restitué par le système pour une requête *q*.

Pour mesurer les performances de notre système de RI sémantique, nous avons évalué et comparé ses résultats par rapport à ceux obtenus par des systèmes de référence (baselines) basés sur l'indexation classique par des mots clés pondérés. En particulier, nous avons considéré deux baselines qui sont:

- *Classic_tf*idf :* système basé sur un index classique par mots clés pondérés par le schéma de pondération *tf*idf* [Salton et al., 73],

- *Classic_BM25 :* système basé sur un index classique par mots clés pondérés par le schéma de pondération *Okapi-BM25* [Robertson et al., 94].

Nous avons mené trois séries d'expérimentations:

(1). la première série vise à évaluer les performances de notre approche d'indexation sémantique,

(2). la seconde série vise à évaluer les performances de nos approches de pondération des concepts (*Ct-Ict* et *Tidf*),

(3). la troisième vise à évaluer les performances de notre approche d'évaluation sémantique des requêtes.

Afin d'évaluer l'impact de la collection utilisée sur les résultats de notre approche de RI sémantique, nous avons réalisé nos expérimentations sur deux collections de test : une collection du domaine général, la collection TIME et une collection du domaine biomédical, la collection Muchmore.

Remarque : A cause de la complexité des calculs induits par les méthodes de désambiguïsation locale, globale et mixte, la collection Muchmore, de taille relativement importante, n'a été utilisée que pour expérimenter notre approche d'indexation sémantique basée sur la désambiguïsation globale et nos approches de pondération des concepts (*Ct-Ict* et *Tidf*) issus de cette désambiguïsation.

4.3.1 La collection TIME

La collection TIME est une petite collection composée de 423 documents issus d'articles de presse du magazine Time de 1963, et d'un nombre important de requêtes (83). Parmi ces requêtes, nous avons sélectionné 50 d'entre elles qui fournissent les résultats non nuls dans une recherche classique à base de mots clés pondérés par *tf*idf* sous la plateforme Terrier. Ces 50 requêtes sont utilisées dans nos expérimentations avec leurs jugements de pertinences associés.

Chaque document de la collection est identifié comme suit :

**TEXT id-doc date-doc PAGE n°page*

Exemple : Le document numéro 20 du 01/04/63, page 21, est présenté comme suit :

**TEXT 020* 01/04/63 *PAGE 021*

THE ROAD TO JAIL IS PAVED WITH NONOBJECTIVE ART SINCE THE KREMLIN'S SHARPEST BARBS THESE DAYS ARE AIMED AT MODERN ART AND " WESTERN ESPIONAGE, " IT WAS JUST A MATTER OF TIME BEFORE THE KGB'S COPS WOULD TURN UP A VICTIM WHOSE WRONGDOINGS COMBINED BOTH EVILS . HE TURNED OUT TO BE A LENINGRAD PHYSICS TEACHER WHOSE TASTE FOR ABSTRACT PAINTING ALLEGEDLY LED HIM TO JOIN THE US SPY SERVICE. POLICE SAID THEY FIRST SPOTTED THE TEACHER, ONE RUDOLF FRIEDMAN, AS HE MUTTERED UNCOMPLIMENTARY REMARKS ABOUT SOCIALIST REALISM WHILE STROLLING THROUGH LENINGRAD'S RUSSIAN MUSEUM. A WELL-DRESSED US TOURIST APPROACHED HIM, ENTHUSIASTICALLY SHARED HIS SENTIMENTS, AND PROMISED TO SEND FRIEDMAN REPRODUCTIONS OF AVANT-GARDE PAINTINGS FROM AMERICA. THE PICTURE FRIEDMAN LIKED BEST, SAID THE COPS INDIGNANTLY, WAS A " CHAOS OF BLACK, RED AND BLUE SPLOTCHES CAPTIONED I NEED YOU TONIGHT." SOON, THEY SAID, THE TEACHER WAS GETTING MESSAGES FROM THE US WRITTEN IN INVISIBLE INK. JUST AS FRIEDMAN PREPARED TO DELIVER INFORMATION "VERY REMOTE FROM THEORETICAL ARGUMENTS ABOUT ABSTRACT ART," POLICEMOVED IN AND HUSTLED HIM OFF TO JAIL.

Exemples de requêtes TIME

```
1    KENNEDY ADMINISTRATION PRESSURE ON NGO DINH DIEM TO STOP SUPPRESSING
THE BUDDHISTS.
2    EFFORTS OF AMBASSADOR HENRY CABOT LODGE TO GET VIET NAM'S PRESIDENT
DIEM TO CHANGE HIS POLICIES OF POLITICAL REPRESSION.
3    NUMBER OF TROOPS THE UNITED STATES HAS STATIONED IN SOUTH VIET NAM AS
COMPARED WITH THE NUMBER OF TROOPS IT HAS STATIONED IN WEST GERMANY.
```

Exemples de jugements de pertinence

1 268.txt 0 0
1 288.txt 0 0
2 326.txt 0 0
2 334.txt 0 0
3 326.txt 0 0

La TIME et WordNet

Comme le montre le tableau 4.1, le taux de couverture du vocabulaire de la TIME par WordNet est de 92.01% du vocabulaire des documents de la collection, et 94.16% du vocabulaire utilisé dans les 50 requêtes. Les mots orphelins (n'ayant pas d'entrée dans WordNet) représentent seulement 7.99% des termes des documents de la collection et 5.84% des termes des 50 requêtes.

	A partir Des documents de la collection TIME	**A partir Des 50 requêtes utilisées**
Nombre de termes identifiés	122107	428
Nombre de termes identifiés, couverts par WordNet	111133	403
(%) taux de couverture de WordNet	92.01%	94.16%

Tableau 4.1 : Nombre de termes identifiés dans TIME couverts par Wordnet.

4.3.2 La collection Muchmore

La collection Muchmore est une collection de taille moyenne, composée de 7823 documents issus de résumés de revues médicales du site web Springer Link[32]. Elle propose en outre 25 requêtes et des jugements de pertinences associés à ces requêtes. Dans ces jugements de pertinence, nous avons constaté que 176 documents, figurant dans la liste des documents pertinents, n'appartiennent pas à la collection Muchmore. Par conséquent, nous avons élargi cette collection en la complétant avec ces documents pertinents à partir du corpus bilingue

[32] http://link.springer.com/

Muchmore Springer[33]. La nouvelle taille de notre collection Muchmore est de 7999 documents.

La collection Muchmore est distribuée en deux langues (anglais-allemand) avec une version annotée et l'autre sans annotation. Dans nos tests d'évaluation menés, dans le cadre de ce travail, nous avons utilisé la version anglaise non annotée. Pour des raisons de simplicité et de réduction des temps de traitement, nous nous sommes limités, dans nos expérimentations, sur un sous-ensemble composé de 1954 documents parmi les 7999 de la collection. Ce sous-ensemble est construit à partir des 100 premiers documents renvoyés par la baseline *Classic_tf*idf* pour les 25 requêtes de Muchmore. Toutes les requêtes ont été utilisées dans nos tests.

Chaque document de la collection est identifié comme suit :

Nom-revue.id-doc eng.abstr

Exemple de document Arthroskopie.00130030.eng.abstr

```
The treatment of acute posterior instabilities can involve a reconstructive procedure as well as an antidislocation augmentation. Tibial bony avulsion ruptures must be reconstructed. For the treatment of chronic posterior instabilities we used the patella tendon graft, which can be implanted by an anterior or posterior approach. In cases of posterolateral instabilities, additional procedures are necessary.
```

Exemples de requêtes

```
1: Arthroscopic treatment of cruciate ligament injuries
2: Complications of arthroscopic interventions
```

Exemples de jugements de pertinence

```
1 0 Arthroskopie /80110304 1
2 0 Arthroskopie/90120198 1
```

La Muchmore et WordNet

Le vocabulaire utilisé dans les documents et les requêtes de la collection Muchmore est largement couvert par WordNet. Comme indiqué dans le Tableau 4.2, le taux de couverture représente 88.86% du vocabulaire des documents et 87.35% du vocabulaire utilisé dans les requêtes, ce qui nous a permis ainsi d'exploiter WordNet comme source d'évidence pour notre approche d'indexation des documents (et requêtes) de la collection et la pondération des concepts de nos index sémantiques.

[33]http://muchmore.dfki.de/resources2.htm

	A partir Des documents du sous-ensemble la collection Muchmore	**A partir Des 25 requêtes utilisées**
Nombre de termes identifiés	156024	87
Nombre de termes identifiés, couverts par WordNet	138654	76
(%) taux de couverture de WordNet	88.86	87.35

Tableau 4.2 : Nombre de termes identifiés dans Muchmore couverts par WordNet.

4.4. Evaluation avec la collection TIME

4.4.1 Evaluation de l'approche d'indexation sémantique dans TIME

Notre objectif à travers ces expérimentations est d'étudier l'impact de notre représentation sémantique des documents et requêtes sur l'efficacité de la recherche. Diverses expérimentations ont été menées dans l'objectif de :

- Evaluer notre approche d'indexation par les collocations et les mots simples (mots orphelins et mots simples ayant une entrée dans WordNet),
- Evaluer nos approches de désambiguïsation (locale, globale, mixte) des sens des mots,
- Evaluer notre approche de désambiguïsation par les domaines.

Nous avons en outre mené des expérimentations supplémentaires afin de:

- Evaluer l'impact de l'indexation sémantique par les concepts-noms,
- Evaluer l'apport de l'indexation sémantique combinée {Collocations + Sens},
- Evaluer l'apport des domaines en désambiguïsation des sens des mots.

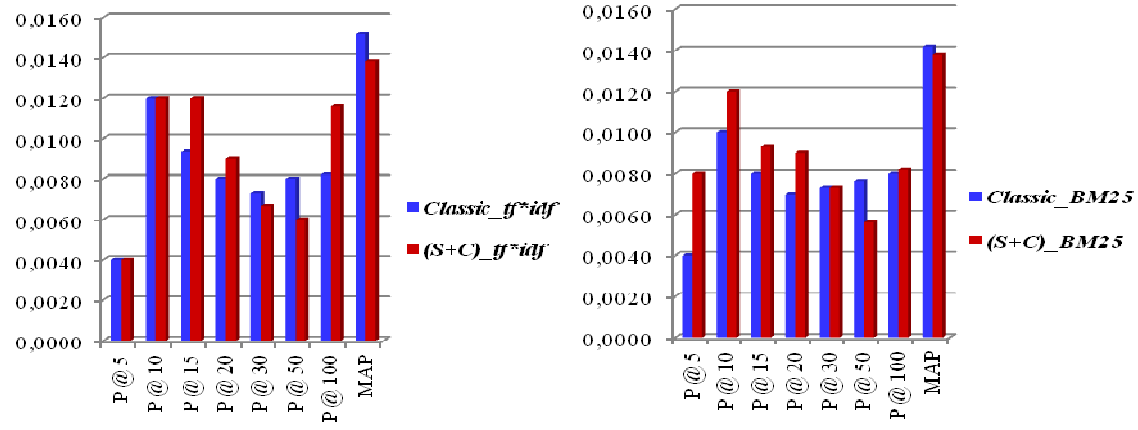
Dans ces expérimentations, nous avons implémenté un système de recherche basé sur une pondération classique (*tf*idf* ou *Okapi-BM25*) en utilisant la mesure de pertinence cosinus [Salton et al., 83] dans le modèle de recherche. Le but est d'évaluer uniquement l'apport de l'indexation sémantique sur les performances du système sans tenir compte des approches de pondération proposées au préalable et indépendamment de la mesure de pertinence sémantique proposée.

4.4.1.1 Evaluation de l'approche d'indexation par l'index combiné {mots simples + collocations}

Pour évaluer l'apport de la combinaison des collocations et les mots simples (représentés par les mots orphelins et les mots simples ayant une entrée dans WordNet) dans l'indexation des documents (et requêtes), nous avons comparé les résultats issus des baselines *Classic_tf*idf* et *Classic_BM25* à ceux des index suivants :

1. *(S+C)_tf*idf :* index combiné {mots simples + collocations} pondéré par *tf-idf*,
2. *(S+C)_BM25 :* index combiné {mots simples + collocations} pondéré par *Okapi-BM25*.

Les résultats des comparaisons obtenus avec la collection TIME sont détaillés ci-après à travers la figure 4.1 suivante.



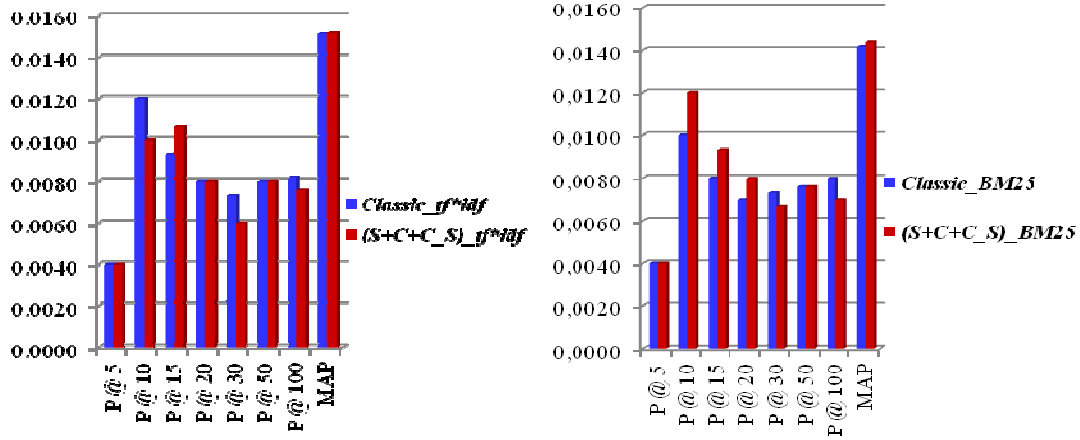
(a) *Classic_tf*idf* vs *(S+C)_tf*idf* (b) *Classic_BM25* vs *(S+C)_BM25*

Figure 4.1: Indexation classique basée mots clés *vs* Indexation basée {mots simples + collocations}.

Les graphiques de la figure 4.1 montrent que les résultats obtenus par les index *(S+C)_tf*idf* et *(S+C)_BM25* présentent une amélioration seulement aux premiers points de précision : *P@5, P@10, P@15* et *P@20.* Les gains de performances de l'index *(S+C)_tf*idf*, par rapport à la baseline *Classic_tf*idf*, sont de 29.03% pour *P@15* et 12.5% pour *P@20.* D'autre part, les gains de performances de l'index *(S+C)_BM25*, par rapport à la baseline *Classic_BM25*, sont de 100%, 20%, 16.25% et 28.57% pour les précisions *P@5*, *P@10*, *P@15* et *P@20* respectivement. Cette amélioration aux niveaux des rangs : 5, 10, 15, et 20 peut être interprétée par le fait que les collocations de certaines requêtes rendent les performances de notre système de recherche meilleures dans les premiers documents restitués par le système. Cependant, on note une dégradation de la performance de la *MAP* avec un taux de décroissement de -10.14% par rapport à l'index *Classic_tf*idf* et de -2.81% par rapport à la l'index *Classic_BM25.* Ceci peut être expliqué d'une part par le fait que le taux des collocations n'est pas important dans les documents et les requêtes de la collection TIME (seulement 4,54% dans les documents et 10,28% dans les requêtes utilisées), et d'autre part par le fait que l'appariement document-requête est lexical basé sur la présence ou l'absence

d'un terme de la requête dans le document. A titre d'exemple, lors de la recherche pour la collocation *president_kennedy* de la requête 55: *"SUGGESTION MADE BY PRESIDENT KENNEDY FOR A NATO NUCLEAR MISSILE FLEET MANNED BY INTERNATIONAL CREWS"*, le système renvoie les seuls documents qui contiennent cette collocation et ne retourne pas les documents pourtant pertinents contenant des termes sémantiquement liés (par exemple le terme *JFK*, ou *Kennedy*).

Pour tenter de pallier à ce problème, nous avons enrichi les deux index *(S+C)_tf*idf* et *(S+C)_BM25* par l'ajout des mots simples (non vides) qui composent leurs collocations respectives, obtenant ainsi les index enrichis respectifs suivants : *(S+C+S_C)_tf*idf* et *(S+C+S_C)_BM25*. Nous avons ensuite comparé les résultats de la recherche obtenus par *(S+C+S_C)_tf*idf* et *(S+C+S_C)_BM25* par rapport à ceux des baselines. Les comparaisons réalisées sont données en figure 4.2.



(a) *Classic_tf*idf* vs *(S+C+C_S)_tf*idf*

(b) *Classic_BM25* vs *(S+C+S_C)_BM25*

Figure 4.2: Indexation classique basée mots clés *vs* Indexation sémantique basée (mots simples+collocations+mots simples des collocations).

Les résultats obtenus montrent que les index *(S+C+S_C)_tf*idf* et *(S+C+S_C)_BM25* ne dégradent pas les performances de la recherche par rapport aux baselines. On note même pour l'index *(S+C+S_C)_BM25*, une légère amélioration de la *MAP* (de l'ordre de 1,5%) par rapport à l'index *Classic_BM25*.

De ces résultats, il ressort que l'index combiné {mots simples +collocations} a augmenté les performances de la recherche, par rapport à un index classique basé mots-clés, uniquement dans les premiers documents restitués, alors que l'index combiné {mots simples + collocations + mots-clés des collocations} n'a pas apporté d'amélioration significative des performances par rapport à une indexation classique.

4.4.1.2 Evaluation de l'approche d'indexation sémantique combinée {mots simples orphelins + collocations + sens}

Pour évaluer l'apport de notre approche d'indexation sémantique basée sur les concepts, incluant collocations et sens identifiés à l'issue de la désambiguïsation (nos trois approches de désambiguïsation locale, globale et mixte sont évaluées), par rapport à une indexation classique, nous avons comparé aux résultats issus des baselines *Classic_tf*idf* et *Classic_BM25*, les résultats issus des index sémantiques suivants :

1. *Sem_L_tf*idf* : index sémantique {mots orphelins + collocations + sens} pondéré par *tf*idf*, (les sens sont ici identifiés par une désambiguïsation contextuelle locale),
2. *Sem_L_BM25* : index sémantique {mots orphelins + collocations + sens} pondéré par *Okapi-BM25*, (les sens sont ici identifiés par une désambiguïsation contextuelle locale)
3. (3) *Sem_G_tf*idf* : index sémantique {mots orphelins + collocations + sens} pondéré par *tf*idf*, (les sens sont ici identifiés par une désambiguïsation contextuelle globale),
4. *Sem_G_BM25* : index sémantique {mots orphelins + collocations + sens} pondéré par *Okapi-BM25*, (les sens sont ici identifiés par une désambiguïsation contextuelle globale)
5. *Sem_M_tf*idf* : index sémantique {mots orphelins + collocations + sens} pondéré par *tf*idf*, (les sens sont ici identifiés par une désambiguïsation contextuelle mixte),
6. *Sem_M_BM25* : index sémantique {mots orphelins + collocations + sens} pondéré par *Okapi-BM25*, (les sens sont ici identifiés par une désambiguïsation contextuelle mixte)

Les comparaisons réalisées respectivement entre les résultats de la baseline *Classic_tf*idf* et ceux des index sémantiques *Sem_L_tf*idf*, *Sem_G_tf*idf* et *Sem_M_tf*idf* sont représentées à travers le graphique de la figure 4.3. Celles réalisées entre les résultats issus de l'index *Classic_BM25* avec ceux obtenus par nos index sémantiques *Sem_L_BM25*, *Sem_G_BM25* et *Sem_M_BM25* respectivement sont présentés en figure 4.4.

De la figure 4.3, il ressort que **l'index sémantique *Sem_G_tf*idf* produit de meilleurs performances de recherche à tous les points de précision** par rapport à l'index *Classic_tf*idf* d'une part et par rapport aux autres index sémantiques *Sem_L_tf_idf* et *Sem_M_tf*idf* d'autre part. Les gains de performances significatifs (supérieurs à 10%) de notre index *Sem_G_tf_idf* relativement à l'index *Classic_tf_idf* sont de 100%, 18.33%, 43.01%, 37.50%, 46.57%, 21.95% et 10.53% pour les précisions *P@5, P@10, P@15, P@20, P@30, P@100* et la *MAP* respectivement. Par ailleurs, les taux d'accroissement significatifs (supérieurs à 10%) apportés par rapport à *Sem_L_tf*idf,* sont de 18.33%, 43.01%, 37.50%, 22.99%, 11.11%, 19.04% et 10.53% pour les précisions *P@10, P@15, P@20, P@30, P@50, P@100* et la *MAP* respectivement. Tandis que les gains de performances observés par rapport à *Sem_M_tf*idf,* sont de 18.33%, 66.25%, 37.50%, 46.57%, 21.95% et 7.74% pour les précisions *P@10, P@15, P@20, P@30, P@100* et la *MAP* respectivement.

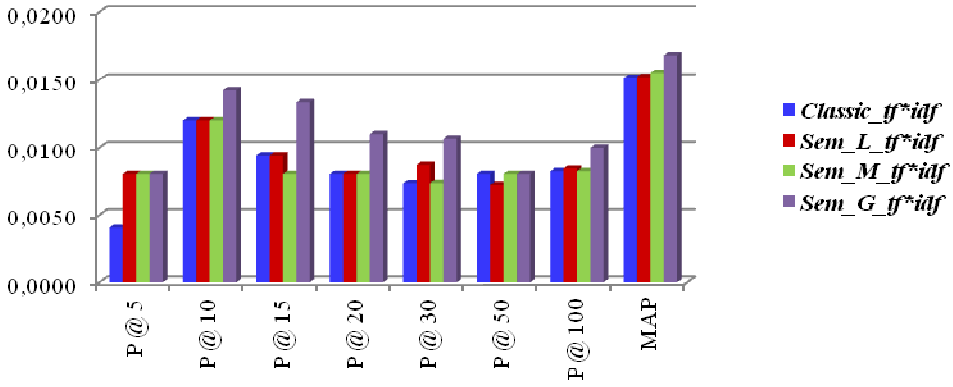


Figure 4.3 : Indexation classique basée mots clés simples *vs* Indexation sémantique basée concepts pondérés par *tf*idf*.

De la figure 4.4, il ressort que **l'index sémantique *Sem_L_BM25* dégrade globalement les performances de la recherche** par rapport à l'index *Classic_BM25* basé mots-clés. Les taux de décroissement sont de -16.25%, -27.39%, -9.58%, -5.26% et -2.11% pour les précisions *P@15, P@20, P@30, P@50* et la *MAP* respectivement. De plus, l'index *Sem_L_BM25* n'est pas performant comparé aux autres index sémantiques *Sem_M_BM25* et *Sem_G_BM25*. **Tandis que l'index sémantique *Sem_G_BM25* présente de meilleurs résultats à tous les points de précision** par rapport à l'index *Classic_BM25*, avec des gains de performances de 2%, 50%, 42.85%, 46.56%, 5.26%, 17.5%, et 5.63% pour les précisions *P@10, P@15, P@20, P@30, P@50, P@100* et la *MAP* respectivement. Par ailleurs, l'index sémantique *Sem_G_BM25* est plus performant que les autres index sémantiques *Sem_L_BM25* et *Sem_M_BM25*. Les taux d'accroissement par rapport à *Sem_L_BM25* sont de 2%, 29.85%, 50%, 33.75%, 11.11%, 9.30% et 7.91% pour les précisions *P@10, P@15, P@20, P@30, P@50, P@100* et la *MAP* respectivement. Tandis que ses gains de performances par rapport à *Sem_M_BM25* sont de 2%, 50%, 40%, 46.57%, 5.26%, 17.5% et 2.73% pour les précisions *P@10, P@15, P@20, P@30, P@50, P@100* et la *MAP* respectivement.

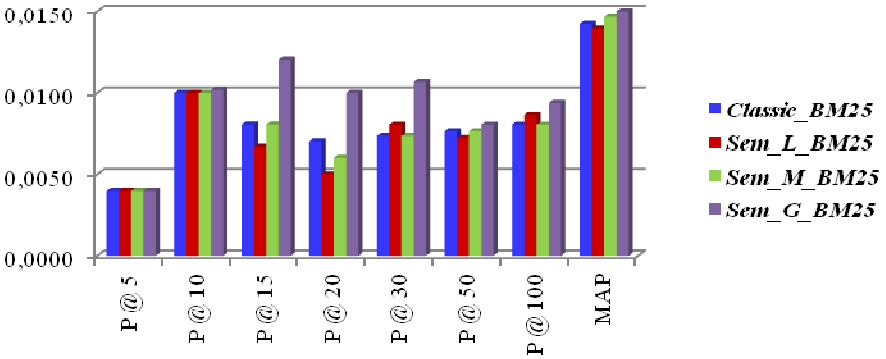


Figure 4.4 : Indexation classique basée mots clés simples *vs* Indexation sémantique basée concepts pondérés par *Okapi-BM25*.

De ces résultats, il ressort que la désambiguïsation locale dégrade globalement les performances, tandis que la désambiguïsation globale semble être le meilleur choix pour une indexation à base de concepts-sens.

Pour conforter nos résultats, nous avons analysé de près les concepts identifiés dans la collection TIME, par chacune de nos méthodes de désambiguïsation (locale, globale et mixte). Nous avons relevé que certains adjectifs et adverbes étaient partiellement désambiguïsés. Ceci peut être expliqué par le fait que la mesure de similarité utilisée [Resnik, 99] dans le score de désambiguïsation, est fondée sur les relations sémantiques de la taxonomie *is-a* des verbes et noms de WordNet. De ce fait, les adverbes et les adjectifs ne sont pas désambiguïsés. Par ailleurs, les relations sémantiques entre les verbes ne sont pas bien élaborées dans WordNet [Fellbaum, 98] [Mallak, 11]. Ainsi, les verbes sont incorrectement désambiguïsés. Ceci a pour effet de générer du bruit lors de la recherche. Pour pallier à cette situation, nous avons entrepris d'indexer les documents et requêtes par les concepts des noms uniquement. Les index sémantiques suivants sont évalués relativement aux baselines et relativement aux index sémantiques introduits en section 4.4.1.2:

- *Sem_Nom_L_tf*idf* : index sémantique basé sur les mots orphelins et les concepts-noms (des collocations et sens issus de la désambiguïsation locale), pondéré par *tf*idf*,
- *Sem_Nom_L_BM25* : index sémantique basé sur les mots orphelins et les concepts-noms (des collocations et sens issus de la désambiguïsation locale), pondéré par *Okapi-BM25*,
- *Sem_Noms_G_tf*idf* : index sémantique basé sur les mots orphelins et les concepts-noms (des collocations et sens issus de la désambiguïsation globale), pondéré par *tf*idf*,
- *Sem_Noms_G_BM25* : index sémantique basé sur les mots orphelins et les concepts-noms (des collocations et sens issus de la désambiguïsation globale), pondéré par *Okapi-BM25*,
- *Sem_Noms_M_tf*idf* : index sémantique basé sur les mots orphelins et les concepts-noms (des collocations et sens issus de la désambiguïsation mixte), pondéré par *tf*idf*,
- *Sem_Noms_M_BM25* : index sémantique basé sur les mots orphelins et les concepts-noms (des collocations et sens issus de la désambiguïsation mixte), pondéré par *Okapi-BM25*.

Les résultats de ces évaluations sont représentés à travers les graphiques des figures 4.5, 4.6, 4.7 et 4.8 suivantes.

De la figure 4.5, il ressort que les index sémantiques *Sem_Noms_L_tf*idf* et *Sem_Noms_L_BM25* présentent globalement de meilleurs résultats à tous les points de précision comparés aux résultats issus des baselines *Classic_tf*idf* et *Classic_BM25*. Les taux d'accroissement significatifs (supérieurs à 5%) produits par l'index *Sem_Noms_L_tf*idf* par rapport à l'index *Classic_tf*idf* sont de 100% pour *P@5*, 16.13% pour *P@15*, 8.75% pour *P@20*, 54.79% pour *P@30*, 20% pour *P@50* et 7.31% pour *P@100*. Le gain de performance de la *MAP* est non significatif mais reste néanmoins positif de 1.97%. Par ailleurs, les gains de performances significatifs apportés par *Sem_Noms_L_BM25* par rapport à *Classic_BM25*

sont de 33.75%, 71.42%, 64.38%, 36.84%, 20% et 14.08% pour les précisions: *P@15, P@20, P@30, P@50, P@100* et *MAP* respectivement.

De plus, les index sémantiques *Sem_Noms_L_tf*idf* et *Sem_Noms_L_*BM25 sont plus performants que les index sémantiques *Sem_L_tf*idf* et *Sem_L_BM25*. Une légère amélioration de la *MAP* est observée pour les deux index sémantiques *Sem_L_Noms_tf*idf* et *Sem_L_Noms_BM25* par rapport aux index *Sem_L_tf*idf* et *Sem_L_BM25* respectivement (de l'ordre de 1.97% pour *Sem_Noms_tf*idf* et de 16.54% pour *Sem_Noms_BM25*).

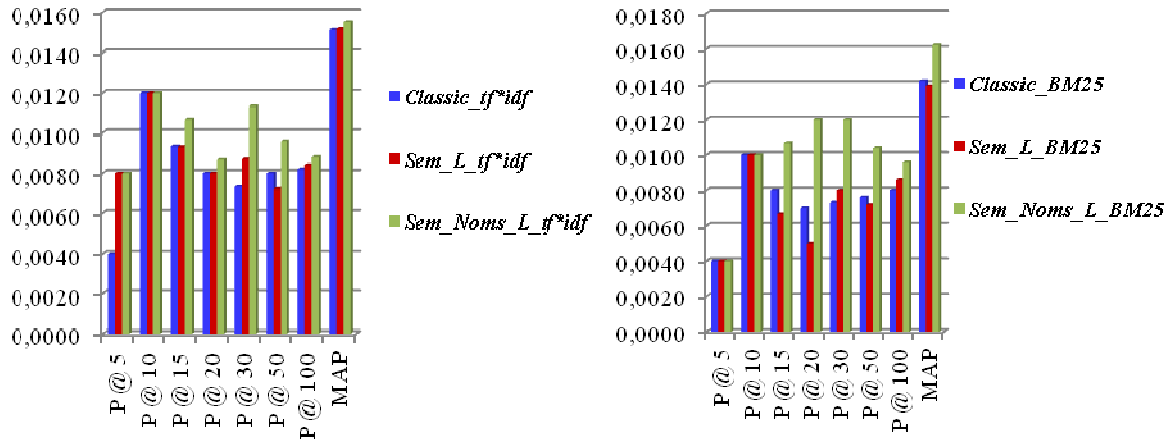


Figure 4.5 : Impact des concepts noms (issus de la désambiguïsation locale) sur les résultats de la recherche.

La figure suivante (figure 4.6) illustre les résultats des index *Sem_Noms_G_tf*idf* et *Sem_Noms_G_BM25* comparativement à ceux des baselines d'une part et à ceux des index sémantiques *Sem_G_tf*idf* et *Sem_G_BM25* d'autre part.

De cette figure, il apparaît clairement que les résultats issus des index sémantiques *Sem_Noms_G_tf_idf* et *Sem_Noms_G_BM25* sont nettement meilleurs comparés aux résultats obtenus des baselines *Classic_tf*idf* et *Classic_BM25* respectivement. D'une part, ldes gains de performances significatifs de l'index *Sem_Noms_G_tf*idf* par rapport à *Classic_tf*idf* sont observés, de l'ordre de 200% pour *P@5*, 18.33% pour *P@10*, 43.01% pour *P@15*, 37.5% pour *P@20*, 46.57% pour *P@30*, 10% pour *P@50* et 17.07% pour *P@100*. L'amélioration de la performance observée pour la *MAP* est très significative (de l'ordre de 91.44%). D'autre part, les taux d'accroissement significatifs observés pour l'index sémantique *Sem_Noms_G_BM25* par rapport à la baseline *Classic_BM25* sont de 200% pour *P@5*, 20.07% pour *P@10*, 50% pour *P@15*, 57.14% pour *P@20*, 46.57% pour *P@30*, 21.05% pour *P@50* et 124.64% pour la *MAP*.

De plus, les index sémantiques *Sem_Noms_G_tf*idf* et *Sem_Noms_G_*BM25 sont plus performants que les index *Sem_G_tf*idf* et *Sem_G_BM25* respectivement. L'amélioration des précisions apportée au niveau de la précision au rang 5 (*p@5)* est de 50.30% pour *Sem_Noms_G_tf*idf* et de 200% pour *Sem_Noms_G_BM25*. En outre, on note un taux

d'accroissement significatif de la *MAP* de l'ordre de 73.21% pour l'index *Sem_Noms_G_tf*idf* et de 112.66% pour l'index *Sem_Noms_G_BM25*.

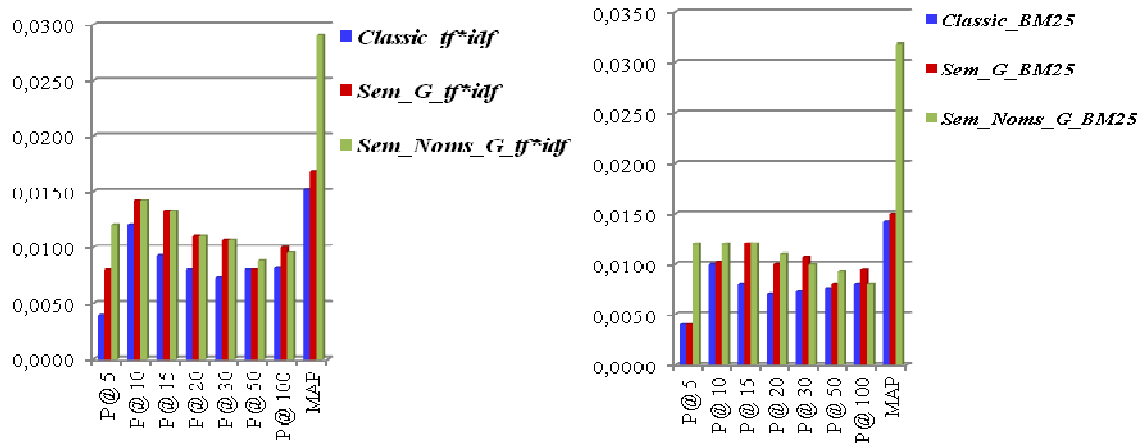


Figure 4.6 : Impact des concepts-noms (issus de la désambiguïsation globale) sur les résultats de la recherche.

Les graphiques de la figure 4.7 présentent les résultats des index sémantiques *Sem_M_tf*idf* et *Sem_M_BM25* comparativement à ceux des baselines d'une part et à ceux des index sémantiques *Sem_Noms_M_tf*idf* et *Sem_Noms_BM25* d'autre part.

De cette évaluation, il est apparait que nos index sémantiques *Sem_Noms_M_tf*idf* et *Sem_Noms_M_BM25* sont globalement meilleurs que les index classiques (baselines). Les gains de performances apportés par notre index *Sem_Noms_M_tf*idf* par rapport à *Classic_tf*idf* sont de 100%, 29.03%, 75%, 54.79%, 20%, 9.75% et 3.94% pour les précisions *P@5, P@15, P@20, P@30, P@50, P@100* et *MAP* respectivement. D'autre part, les taux d'accroissement de notre index sémantique *Sem_Noms_M_BM25* par rapport à *Classic_BM25* sont de 100%, 33.75%, 100%, 64.38%, 36.84%, 20% et 16.19% pour les précisions *P@5, P@15, P@20, P@30, P@50, P@100* et *MAP* respectivement.

Par ailleurs, les index sémantiques *Sem_Noms_M_tf*idf* et *Sem_Noms_M_BM25* présentent globalement de meilleures performances par rapport aux index *Sem_M_tf*idf* et *Sem_M_BM25*. Les gains de performance apportés par *Sem_Noms_M_tf*idf*, par rapport à *Sem_M_tf*idf* sont de 50% pour *P@15*, 75% pour *P@20*, 54.79% pour *P@30*, 20% pour *P*@50, 9.75% pour *P@100* et 1.93% pour la *MAP*. L'index *Sem_Noms_M_BM25* présente par rapport à *Sem_M_BM25* des taux d'accroissement significatifs de l'ordre de 100% pour *P@5*, 33.75% pour *P@15*, 133.33% pour *P@20*, 64.38% pour *P@30*, 36.84% pour *P*@50, 20% pour *P@100* et 13.01% pour la *MAP*.

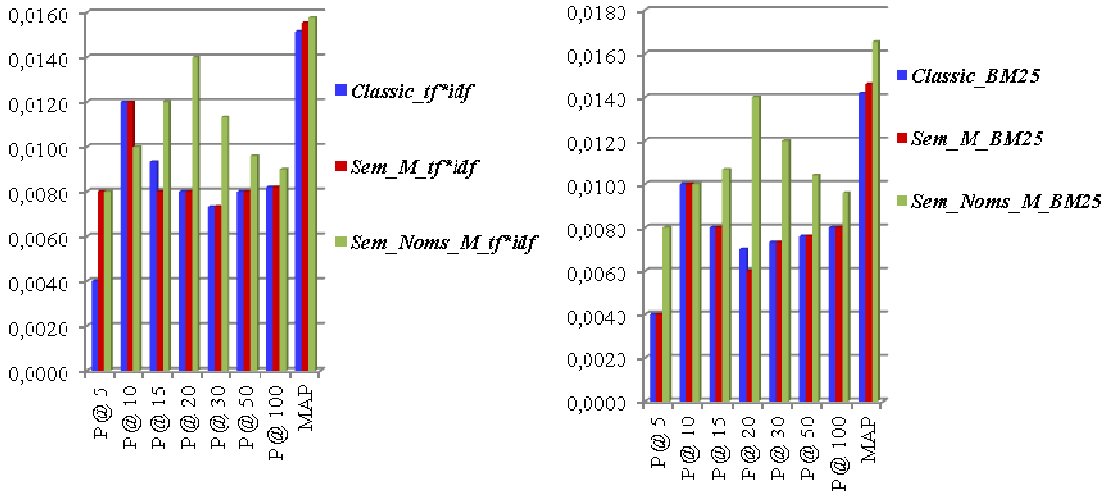


Figure 4.7 : Impact des concepts noms (issus de la désambiguïsation mixte) sur les résultats de la recherche.

De ces évaluations, il ressort clairement que l'indexation sémantique par les concepts-noms est un meilleur choix que l'indexation classique par les mots clés ou que l'indexation sémantique incluant des concepts de catégories syntaxiques quelconques.

Pour désigner l'index le plus performant parmi les index sémantiques basés sur les concepts-noms, nous avons réalisé les comparaisons données à travers les graphiques de la figure 4.8.

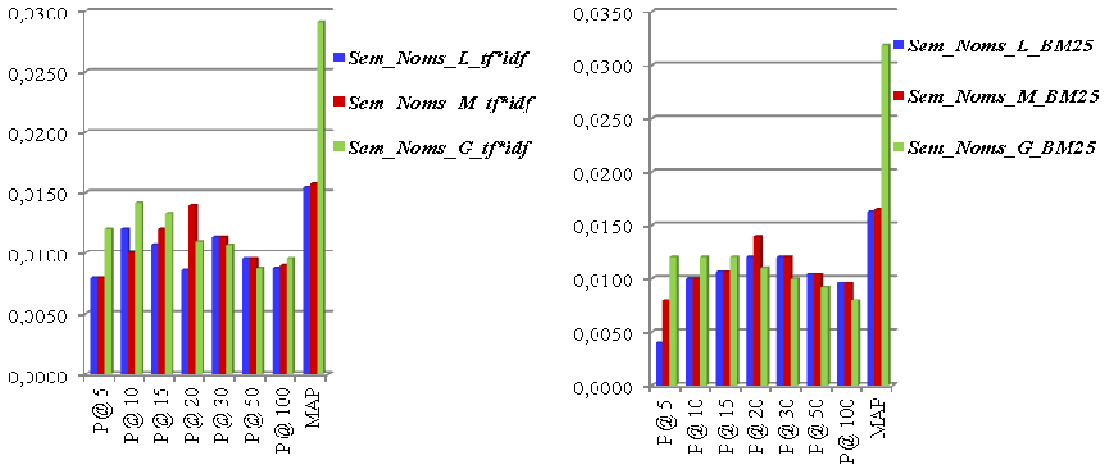


Figure 4.8 : Résultats des Comparaisons entre nos index sémantiques basés concepts-noms (identifiés par nos différentes techniques de désambiguïsation).

De ces graphiques, il ressort que :

- L'index sémantique *Sem_Noms_M_** présente de meilleurs résultats que l'index sémantique *Sem_Noms_L_**. Les gains de performances apportés par *Sem_Noms_M_tf*idf* par rapport à *Sem_Noms_L_tf*idf* sont de 12.14% pour *P@15*, 6.08% pour *P@20* et 1.93% pour la *MAP*. De plus, on observe une légère amélioration des résultats de *Sem_Noms_M_BM25* par rapport à *Sem_Noms_L_BM25* avec des taux d'accroissement de 100% pour *P@5*, 16.66% pour *P@20* et 1.85% pour la *MAP*.

- L'index sémantique *Sem_Noms_G_** est plus performants que les deux autres index sémantiques basés concepts-noms *Sem_Noms_L_**, et *Sem_Noms_M_**. Les taux d'accroissement significatifs de l'index *Sem_Noms_G_tf*idf* par rapport à *Sem_Noms_L_tf*idf*, varient de 50%, 35%, 24.29%, 26.43% et 87.74% pour *P@5, P@10, P@15, P@20* et la *MAP* respectivement. Ses taux d'accroissement par rapport à *Sem_Noms_M_tf*idf* , sont de 50% pour *P@5*, 42% pour *P@10*, 10.83% pour *P@15* et 84.17% pour la *MAP*. Par ailleurs, les taux d'amélioration significatifs de l'index sémantique *Sem_Noms_G_BM25* par rapport à l'index sémantique *Sem_Noms_L_BM25* sont de 200% pour *P@5*, 20% pour *P@10*, 12.14% et 96.91% pour la *MAP*. Ses taux d'accroissement par rapport à *Sem_Noms_M_BM25* sont de 50% pour *P@5*, 20% pour *P@10*, 12.14% et 93.33% pour la *MAP*.

- Finalement, l'index sémantique *Sem_Noms_G_BM25* est plus performant que l'index sémantique *Sem_Noms_G_tf*idf*, avec un taux d'amélioration de la *MAP* de l'ordre de 9.59%.

4.4.1.3 Evaluation de l'approche d'indexation par l'index combiné {concepts + mots-clés}

L'objectif à travers ces expérimentations est de tester l'impact de la combinaison de notre index sémantique basé mots-clés (les mots orphelins sont inclus) et concepts (collocations + sens issus de la désambiguïsation -locale, globale ou mixte-) avec l'index classique basé mots-clés sur les résultats de la recherche. Pour cela, nous avons évalué relativement aux baselines d'une part et aux index sémantiques correspondants précédemment cités (en section 4.4.1.2) d'autre part, les résultats issus des index sémantiques suivants :

1. *(Sem_L+Bas)_tf*idf* : l'index combiné mots-clés et concepts (collocations + sens issus de la désambiguïsation locale), pondéré par *tf*idf*,
2. *(Sem_L+Bas) _BM25* : l'index combiné mots-clés et concepts (collocations + sens issus de la désambiguïsation locale), pondéré par *Okapi-BM25*,
3. *(Sem_G+Bas) _tf*idf* : l'index combiné mots-clés et concepts (collocations + sens issus de la désambiguïsation globale), pondéré par *tf*idf*,
4. *(Sem_G+Bas) _BM25* : l'index combiné mots-clés et concepts (collocations + sens issus de la désambiguïsation globale), pondéré par *Okapi-BM25*,

5. *(Sem_M+Bas) tf*idf* : l'index combiné mots-clés et concepts (collocations + sens issus de la désambiguïsation mixte), pondéré par *tf*idf*,
6. *(Sem_M+Bas)_ BM25* : l'index combiné mots-clés et concepts (collocations + sens issus de la désambiguïsation mixite), pondéré par *Okapi-BM25*.

Les résultats des comparaisons effectuées sont représentés à travers les graphiques des figures 4.9 à 4.12 suivantes.

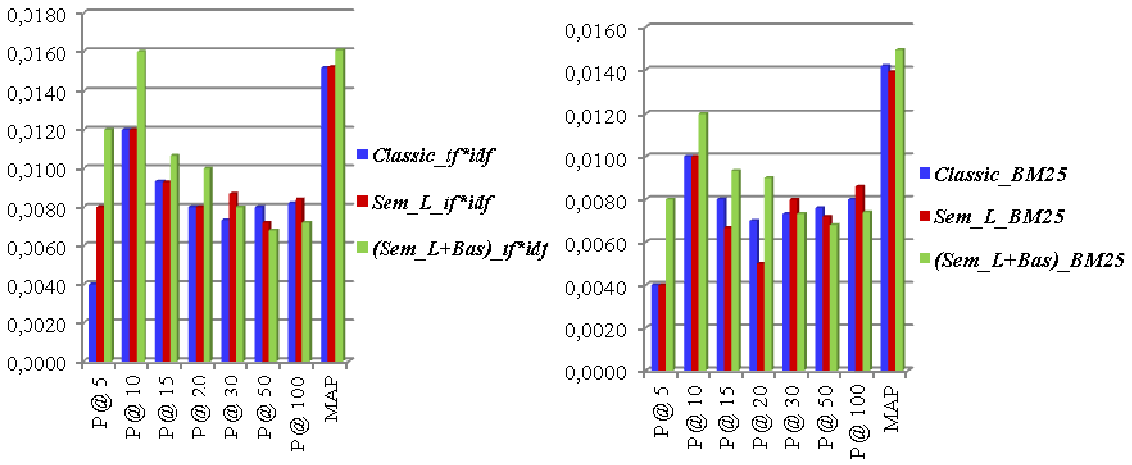


Figure 4.9 : Apport de l'indexation par les concepts (identifiés par notre approche de désambiguïsation locale) combinés à des mots clés simples.

La figure 4.9, présente les comparaisons des résultats issus des index combinés {mots-clés + collocations + sens issus de la désambiguïsation locale} *(Sem_L+Bas) _** aux résultats issus des index classiques (*Classic_tf*idf* et *Classic_BM25*) d'une part et à ceux obtenus par les index sémantiques *Sem_L_tf*idf* et *Sem_L_BM25* d'autre part. Des graphiques de la figure 4.9, il ressort que l'indexation combinée {mots clés + collocations + sens issus d'une désambiguïsation locale} *(Sem_L+Bas) _** présente de meilleurs résultats à tous les points de précision, par rapport aux baselines (*Classic_tf*idf* et *Classic_BM25*), tant avec la pondération *tf*idf* (*(Sem_L+Bas) _tf*idf*) qu'avec la pondération *Okapi-BM25* (*(Sem_L+Bas)_BM25*). Les gains de performance significatifs de l'index *(Sem_L+Bas)_tf*idf* comparé à l'index *Classic_tf*idf* sont respectivement de 299.90% pour *P@5*, 33.33% pour *P@10*, 15.05% pour *P@15*, 25% pour *P@20*, 9.58% pour *P@30* et 5.92% pour la *MAP*. D'autre part, les taux d'accroissement pour l'index combiné *(Sem_L+Bas)_BM25*, par rapport à l'index *Classic_BM25*, sont respectivement de 100% pour *P@5*, 20% pour *P@10*, 16.5% pour *P@15*, 28.57% pour *P@20* et 4.92% pour la *MAP*.

Par ailleurs, les index combinés *(Sem_L+Bas) _tf*idf* et *(Sem_L+Bas) _BM25* sont globalement plus efficaces que les index sémantiques *Sem_L_tf*idf* et *Sem_L_BM25* respectivement. Le taux d'accroissement de la *MAP*, observé pour l'index combiné *(Sem_L+Bas) _**, est de 5.92% par rapport à *Sem_L_tf*idf* et de 7.19% par rapport à *Sem_L_BM25*. Cette amélioration est due probablement à la réintroduction dans les index *(Sem_L+Bas) _**, par le biais des mots-clés, les sens (concepts) de certains termes qui ont été

mal désambiguïsés par notre approche de désambiguïsation locale (index *Sem_L_tf*idf* et *Sem_L_BM25*) ce qui a pour effet d'augmenter la précision.

La figure 4.10, présente les comparaisons des résultats issus des index combinés {mots-clés + collocations + sens issus de la désambiguïsation globale} *(Sem_G+Bas)* _* aux résultats issus des index classiques (*Classic_tf*idf* et *Classic_BM25*) d'une part et à ceux obtenus par les index sémantiques *Sem_G_tf*idf* et *Sem_G_BM25* d'autre part. Des graphiques de la figure 4.10, il ressort clairement que les résultats des index *(Sem_G +Bas)_tf*idf* et *(Sem_G+Bas)_BM25* sont nettement meilleurs que ceux des baselines *Classic_tf*idf* et *Classic_BM25* respectivement. Les taux d'accroissement significatifs de l'index *(Sem_G +Bas)_tf*idf* par rapport à *Classic_tf*idf* sont de 200%, 50%, 43.01%, 37.5%, 27.39%, 7.31% et 16.44% pour *P@5*, *P@10*, *P@15*, *P@20*, *P@30*, *P@100* et la *MAP* respectivement. Tandis que les taux d'accroissement de *(Sem_G+Bas)_BM25* par rapport à *Classic_BM25* sont de 100%, 20%, 62.5%, 52.85%, 23.68%, 2.63%, 15% et 23.94.44% pour *P@5*, *P@10*, *P@15*, *P@20*, *P@30*, *P@50*, *P@100* et la *MAP* respectivement.

En outre, les index *(Sem_G+Bas)_tf*idf* et *(Sem_G+Bas)_BM25* sont plus performants que les index sémantiques *Sem_G_tf*idf* et *Sem_G_BM25* respectivement, en particulier dans les premiers documents renvoyés par le système de recherche. Les gains de performances de *(Sem_G+Bas)_tf*idf* aux points *P@5*, *P@10* et la *MAP* sont respectivement de 50%, 26.76% et 5.35% par rapport à *Sem_G_tf*idf*. Les taux d'accroissement de *(Sem_G+Bas)_BM25*, par rapport à *Sem_G_BM25*, sont de 100%, 17.64%, 8.33%, 7% et 17.33% pour les précisions *P@5*, *P@10*, *P@15*, *P@20* et la *MAP* respectivement. Ces performances peuvent être expliquées par le fait que les deux index *(Sem_G+Bas)_tf*idf* et *(Sem_G+Bas)_BM25* ont permis de réintroduire, par le biais des mots clés, les sens des mots qui ont été mal désambiguïsés par l'approche de désambiguïsation globale (*Sem_G_tf*idf* et *Sem_G_BM25*) ce qui a pour effet d'améliorer la précision du système.

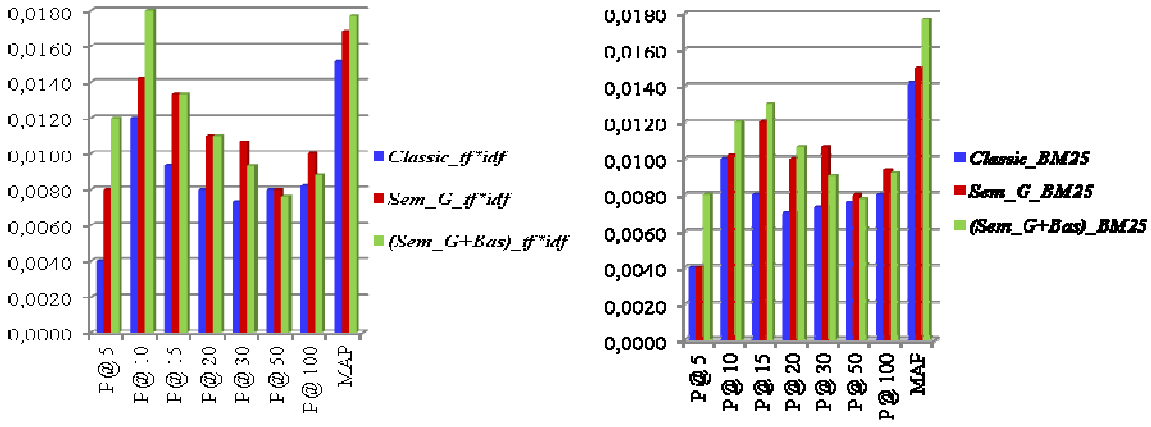


Figure 4.10 : Apport de l'indexation par les concepts (identifiés par notre approche de désambiguïsation globale) combinés à des mots clés simples.

La figure 4.11 présente les comparaisons des résultats issus des index combinés {mots-clés + collocations + sens issus de la désambiguïsation mixte} *(Sem_M+Bas)* _* aux résultats issus des index classiques (*Classic_tf*idf* et *Classic_BM25*) d'une part et à ceux obtenus par les index sémantiques *Sem_M_tf*idf* et *Sem_M_BM25* d'autre part. Des graphiques de la figure 4.11, il ressort que les index *(Sem_M+Bas)_tf*idf* et *(Sem_M+Bas)_BM25* présentent globalement de meilleurs résultats de recherche notamment dans les premiers documents restitués par notre système de recherche (*P@5, P@10, P@15* et *P@20*). Les gains de performances significatifs de l'index *(Sem_M+Bas)_tf*idf* comparé à la baseline *Classic_tf*idf* sont de 200%, 16.66%, 15.05%, 12.5% et 6.57% pour les précisions *P@5, P@10, P@15, P@20* et la *MAP* respectivement. En outre, les gains de performances de l'index *(Sem_M+Bas)_BM25* par rapport à la baseline *Classic_BM25* sont : 100%, 20%, 28.57% et 6.33% pour les précisions *P@5, P@10, P@20* et la *MAP* respectivement.

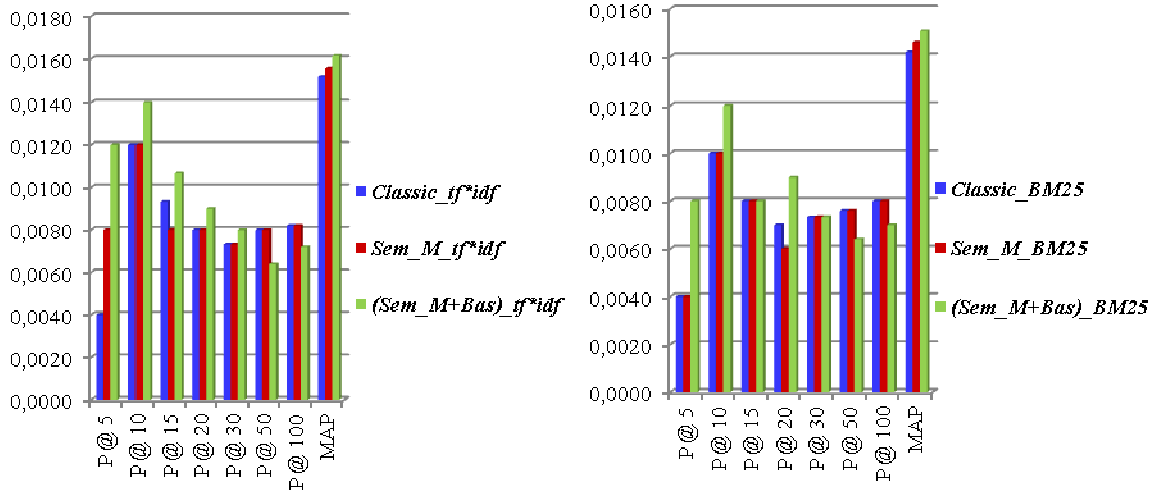


Figure 4.11 : Apport de l'indexation par les concepts (identifiés par notre approche de désambiguïsation mixte) combinés à des mots clés simples.

Finalement, la figure 4.12 permet de comparer les approches de désambiguïsation locale, globale et mixte à travers les résultés des index sémantiques combinés *(Sem_L +Bas)_*, (Sem_G +Bas)_** et *(Sem_M +Bas)_** qui les intègrent. Les graphiques de la figure 4.12 montrent que :

- l'indexation combinée *(Sem_G +Bas)_** présente des résultats nettement meilleurs, à tous les points de précision, que ceux de l'indexation combinée *(Sem_L +Bas)_** d'une part et ceux de l'indexation combinée *(Sem_M +Bas)_** d'autre part. Les taux d'accroissement de *(Sem_G +Bas) _tf*idf*, par rapport à *(Sem_L +Bas)_tf*idf* sont (supérieurs à 5%) : 12.5%, 24.29%, 10% , 16.25%, 11.76%, 22.21% et 9.93% pour les précisions *P@10, P@15, P@20, P@30, P@50, P@100* et la *MAP* respectivement. Tandis que les taux d'accroissement de *(Sem_G +Bas)_BM25*, par rapport à *(Sem_L +Bas)_BM25* sont de 39.78%, 14.7%, 25% et 18.12% pour les *P@30, P@50, P@100* et la *MAP* respectivement. D'autre part, les gains de performances de *(Sem_G +Bas)_tf*idf*, par rapport à *(Sem_M +Bas)_tf*idf* sont : 28.57%, 24.29%, 22.22% , 16.25%, 18.75%, 22.21%

et 9.25% pour les précisions *P@10*, *P@15*, *P@20*, *P@30*, *P@50*, *P@100* et la *MAP* respectivement. Tandis que les gains de performances de *(Sem_G+Bas)_BM25*, par rapport à *(Sem_M+Bas)_BM25* sont : 39.78%, 18.88%, 24.65%, 21.87%, 31.42% et 16.55% pour les précisions *P@15*, *P@20*, *P@30*, *P@50*, *P@100* et la *MAP* respectivement.

- l'indexation combinée *(Sem_M+Bas)_** présente une légère amélioration de la *MAP* par rapport à l'indexation combinée *(Sem_L+Bas)_**, avec un taux d'accroissement (non significatif) de 0.62% par rapport à *(Sem_L +Bas)_tf*idf* et de 1.34% par rapport à *(Sem+Bas)_L_BM25*. D'autre part, les index *(Sem_M+Bas)_tf*idf* et *(Sem_M+Bas)_BM25* ne sont pas plus performants que les index *(Sem_G+Bas)_tf*idf* et *(Sem_G+Bas)_BM25* respectivement.

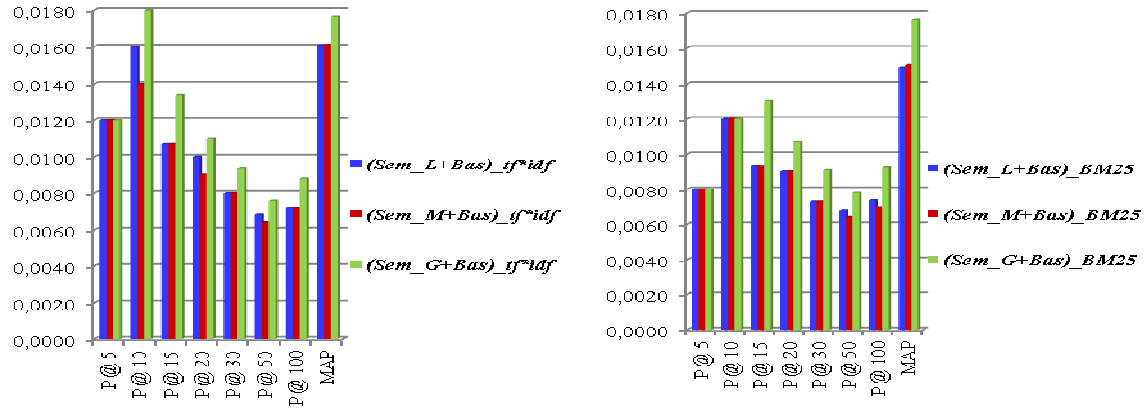


Figure 4.12: Résultats des comparaisons entre les index sémantiques basés mots clés simples combinés à des concepts identifiés par nos différentes approches de désambiguïsation (locale, globale ou mixte).

De ces expérimentations, on peut conclure que nos index sémantiques combinés aux mots-clés sont d'un apport certain dans l'amélioration des performances de la recherche.

4.4.1.4 Impact des collocations sur l'approche d'indexation sémantique

Afin de mesurer l'apport des collocations en indexation sémantique et leur impact sur l'efficacité de la recherche, nous avons expérimenté et comparé aux index classiques d'une part, puis aux index sémantiques *Sem_G_** *et Sem_M_** d'autre part, les résultats issus des index sémantiques suivants qui n'incluent pas les collocations :

1. *Sem_G_Sans_C_tf*idf* : index sémantique combiné {mots-simples orphelins + sens issus de la désambiguïsation globale), pondéré par *tf*idf*,
2. *Sem_G_Sans_C_BM25* : index sémantique combiné {mots-simples orphelins + sens issus de la désambiguïsation globale), pondéré par *Okapi-BM25*,

3. *Sem_M_Sans_C_tf*idf* : index sémantique combiné {mots-simples orphelins + sens issus de la désambiguïsation mixte), pondéré par *tf*idf*,
4. *Sem_M_Sans_C_BM25* : index sémantique combiné {mots-simples orphelins + sens issus de la désambiguïsation mixte), pondéré par *Okapi-BM25*.

Ces évaluations sont illustrées à travers les figures 4.13 et 4.14.

Les graphiques de la figure 4.13 comparent les résultats issus de *Sem_G_Sans_C_tf*idf* et *Sem_G_Sans_C_BM25* à ceux issus des index *Sem_G_tf*idf* et *Sem_G_BM25* respectivement. De ces graphiques, il apparaît que les index sémantiques *Sem_G_Sans_C_tf*idf* et *Sem_G_Sans_C_BM25* présentent des dégradations des performances. Les taux de décroissement observés dans *Sem_G_Sans_C_tf*idf* par rapport à *Sem_G_tf*idf* sont de -43.66% pour *P@10*, -9.77% pour *P@15*, -18.18% pour *P@20*, -8% pour *P@100* et -4.76% pour la *MAP*. D'autre part, les taux de décroissement de l'index *Sem_G_Sans_C_BM25*, par rapport à *Sem_G_BM25*, sont de -1.96% pour *P@10*, -10.83% pour *P@15*, -10% pour *P@20*, -18.69% pour *P@30*, -8.51% pour *P@100* avec une légère diminution de la *MAP* (de l'ordre de -1.33%).

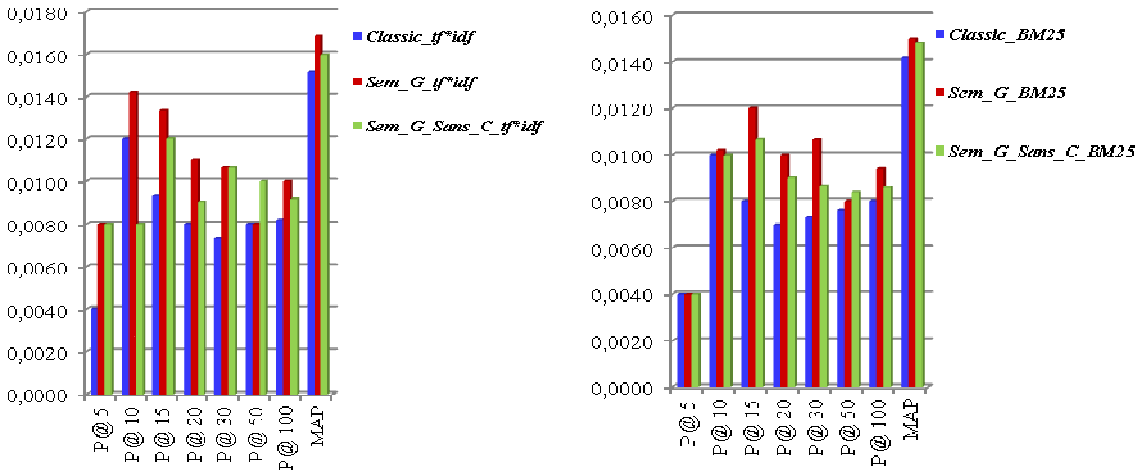


Figure 4.13: Impact des collocations sur notre approche d'indexation sémantique basée sur les concepts issus de la désambiguïsation globale.

Néanmoins, les index sémantiques *Sem_G_Sans_C_tf*idf* et *Sem_G_Sans_C_BM25* restent globalement plus performants que les baselines *Classic_tf*idf* et *Classic_BM25* respectivement. Les gains de performance observés pour l'index *Sem_G_Sans_C_tf*idf* par rapport à *Classic_tf*idf* sont de 100%, 29.03%, 12.5%, 46.57%, 25%, 12.91% et 5.26% pour les précisions *P@5*, P@15, *P@20*, *P@30*, *P@50*, *P@100* et la *MAP* respectivement. En outre, les gains de performances de l'index *Sem_G_Sans_C_BM25* par rapport à *Classic_BM25* sont de 33.75%, 28.87%, 19.17%, 10.52%, 7.5% et 4.22% pour les précisions P@15, *P@20*, *P@30*, *P@50*, *P@100* et la *MAP* respectivement. **Cette amélioration des performances de la recherche peut être interprétée par le fait que notre approche de**

désambiguïsation globale a permis d'identifier les sens corrects des mots-clés simples dans le contexte global.

Les graphiques de la figure 4.14 comparent les résultats issus de *Sem_M_Sans_C_tf*idf* et *Sem_G_Sans_C_BM25* à ceux issus des index sémantiques *Sem_M_tf*idf* et *Sem_M_BM25* respectivement. On observe une légère diminution de la *MAP* pour les index *Sem_M_Sans_C_tf*idf* et *Sem_M_Sans_C_BM25* par rapport aux index *Sem_M_tf*idf* et *Sem_M_BM25* respectivement, avec des taux de décroissement (non significatifs) de -1.29% pour *Sem_M_Sans_C_tf*idf* et de –2.05% pour *Sem_M_Sans_C_BM25*.

Toutefois, les index *Sem_M_Sans_C_tf*idf* et *Sem_M_Sans_C_BM25* présentent une amélioration non significatifs de la *MAP* comparés aux baselines, de l'ordre de 0.65% pour l'index *Sem_M_Sans_C_tf*idf* par rapport à *Classic_tf*idf* et de l'ordre de 0.70% pour l'index *Sem_M_Sans_C_BM25* par rapport à *Classic_BM25*. **Cette légère amélioration peut être interprétée par le fait que notre désambiguïsation contextuelle mixte a désambiguïsé correctement certains mots des documents.**

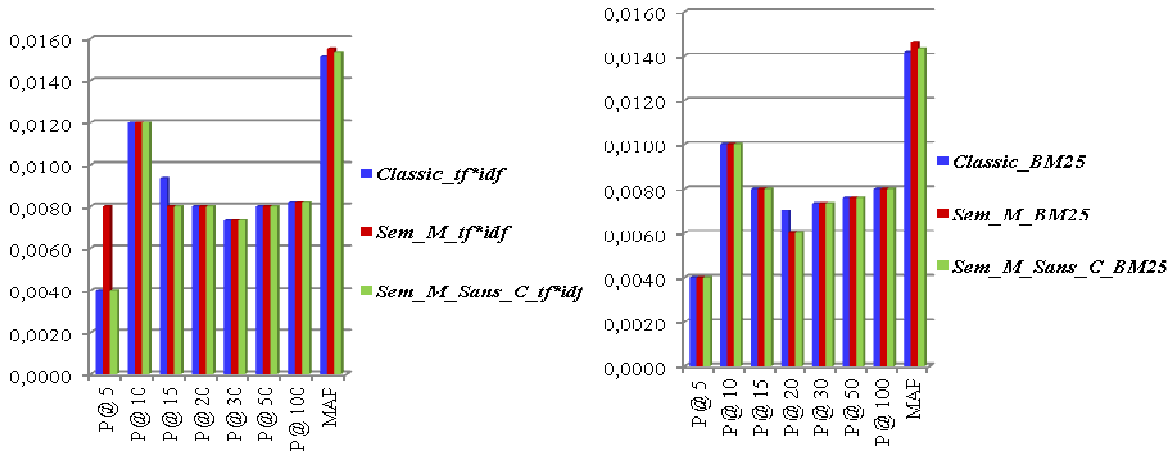


Figure 4.14: Impact des collocations sur notre approche d'indexation sémantique basée sur les concepts issus de la désambiguïsation mixte.

De ces expérimentations, nous concluons que l'utilisation des collocations en indexation sémantique des documents (et requêtes) a légèrement amélioré les résultats de la recherche. Cette amélioration légère est due au fait que le taux de collocations dans les requêtes et les documents de la TIME est faible (10,28% dans les 50 requêtes et 4,54% dans les documents). Or, on peut supposer que l'existence d'un nombre important de collocations dans les requêtes et les documents pourrait nettement améliorer les performances de la recherche.

4.4.1.5 Impact des domaines des concepts sur l'approche d'indexation sémantique

Pour évaluer l'impact de notre approche de désambiguïsation par les domaines, nous avons expérimenté et évalué relativement aux baselines d'une part eu aux index sémantiques (section 4.4.1.2) d'autre part, les résultats issus des index sémantiques suivants dont les concepts-sens sont identifiés sans la désambiguïsation par les domaines :

1. *Sem_G_Sans_Dom_tf*idf* : index sémantique combiné {mots-clés orphelins + collocations + sens identifiés hors domaines dans une désambiguïsation globale}, pondéré par *tf*idf*,
2. *Sem_G_Sans_Dom_BM25* : index sémantique combiné {mots-clés orphelins + collocations+ sens identifiés hors domaines dans une désambiguïsation globale}, pondéré par *Okapi-BM25*,
3. *Sem_M_Sans_Dom_tf*idf* : index sémantique combiné {mots-clés orphelins + collocations+ sens identifiés hors domaines dans une désambiguïsation mixte}, pondéré par *tf*idf*,
4. *Sem_M_Sans_Dom_BM25* : index sémantique combiné {mots-clés orphelins + collocations+ sens identifiés hors domaines dans une désambiguïsation mixte}, pondéré par *Okapi-BM25*.

Les résultats de ces évaluations sont présentés dans les graphiques des figures 4.15 et 4.16.

La figure 4.15 permet de comparer les résultats issus des index *Sem_G_Sans_Dom_tf*idf* et *Sem_G_Sans_Dom_BM25* à ceux issus des index *Sem_G_tf*idf* et *Sem_G_BM25* d'une part, et à ceux des baselines d'autre part. De ses graphiques, il ressort que les index *Sem_G_Sans_Dom_tf*idf* et *Sem_G_Sans_Dom_BM25* diminuent nettement les performances du système de recherche comparés respectivement aux index sémantiques *Sem_G_tf*idf* et *Sem_G_BM25*, avec des taux de décroissement de -100 pour *P@5*, -29.57% pour *P@10*, -19.54% pour *P@15*, -13.08% pour *P@30*, -5% pour *P@50*, -18% pour *P@100* et -14.28% pour *MAP* pour l'index *Sem_G_Sans_Dom_tf*idf* . En outre, les taux de décroissement de *Sem_G_Sans_Dom_BM25* par rapport à *Sem_G_BM25* sont respectivement de -1.98% pour *P@10*, -44.16% pour *P@15*, -10% pour *P@20*, -18.69% pour *P@30*, -10.63% pour *P@100*, -10.66% pour *MAP*. Par ailleurs, *Sem_G_Sans_Dom_tf*idf* et *Sem_G_Sans_Dom_BM25* donnent de moins bons résultats que les baselines *Classic_tf*idf* et *Classic_BM25* respectivement, avec des taux de décroissement de la *MAP* de -5.26% pour *Sem_G_Sans_Dom_tf*idf* et de -5.63% pour *Sem_G_Sans_Dom_BM25*.

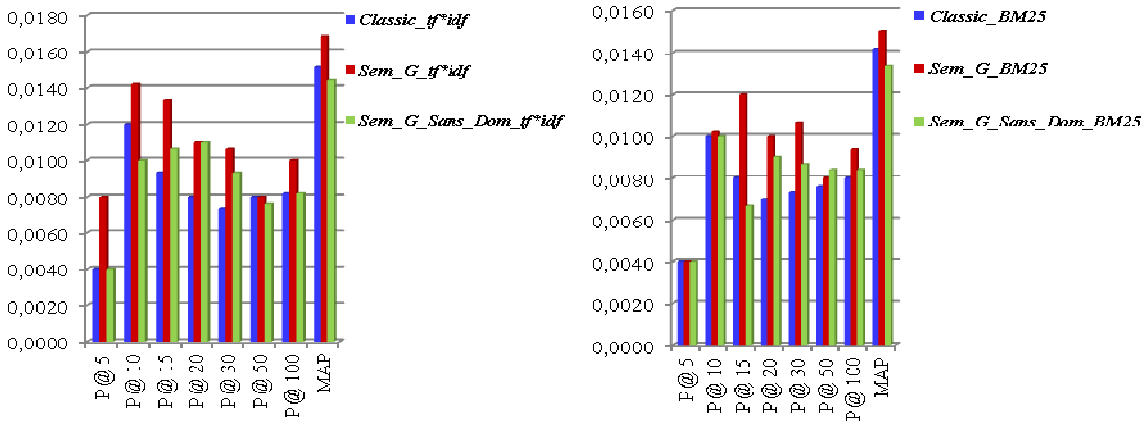


Figure 4.15: Apport de notre approche de désambiguïsation des domaines des mots dans la technique de désambiguïsation contextuelle globale.

De ces résultats, on peut déduire que la désambiguïsation par les domaines utilisée dans notre technique de désambiguïsation globale fournit de meilleurs résultats de recherche.

La figure 4.16 permet de comparer les résultats issus des index *Sem_M_Sans_Dom_tf*idf* et *Sem_M_Sans_Dom_BM25* à ceux issus des index *Sem_M_tf*idf* et *Sem_M_BM25* d'une part, et ceux des baselines d'autre part.

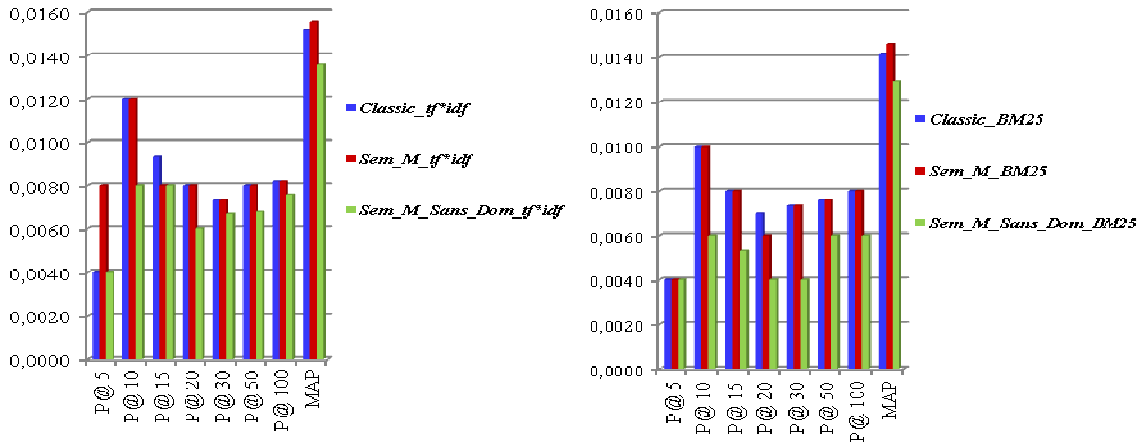


Figure 4.16: Apport de notre approche de désambiguïsation des domaines des mots dans la technique de désambiguïsation contextuelle mixte.

De ses graphiques, il ressort que la désambiguïsation sans les domaines provoque une nette dégradation des performances de la recherche à tous les points de précision par rapport à la désambiguïsation par les domaines. Les taux de décroissement observés pour l'index *Sem_M_Sans_Dom_tf*idf* par rapport à l'index *Sem_M_tf*idf* sont de -100%, -33.33%, -25%, -8.21%, -15%, -7.3% et -12.25% pour les précisions *P@5, P@10, P@20, P@30, P@50, P@100* et la *MAP* respectivement. De même, les taux de décroissement observés

pour *Sem_M_Sans_Dom_BM25* par rapport à *Sem_M_BM25* sont de -100%, -33.33%, -25%, -8.21%, -15%, -7.3% et -12.25% pour les précisions *P@5, P@10, P@20, P@30, P@50, P@100* et la *MAP* respectivement. En outre, les index *Sem_M_Sans_Dom_tf*idf* et *Sem_M_Sans_Dom_BM25* sont moins performants que les baselines *Classic_tf*idf* et *Classic_BM25*. Les taux de décroissement sont observés au niveau de la *MAP* de l'ordre de -10,52% pour l'index *Sem_M_Sans_Dom_tf*idf* et de -9.15% pour l'index *Sem_M_Sans_Dom_BM25*.

De ces résultats, on peut déduire que la désambiguïsation par les domaines utilisée dans notre technique de désambiguïsation mixte fournit de meilleurs résultats de recherche.

A ce niveau de l'évaluation, il apparaît que les domaines des mots renforcent la précision de la désambiguïsation des sens des mots permettant ainsi d'améliorer les performances de la recherche

4.4.1.6 Etude comparative entre notre approche de désambiguïsation des domaines et l'approche de désambiguïsation des domaines proposée par Kolte

Nous avons dans ces expérimentations évalué notre approche de désambiguïsation des domaines comparativement à celle proposée dans [Kolte et al., 09]. Pour ce faire, nous avons comparé les résultats issus de notre approche d'indexation sémantique basée sur la désambiguïsation globale (respectivement mixte), représentée par les index sémantiques *Sem_G_tf*idf* et *Sem_G_BM25(respectivement Sem_M_tf*idf* et *Sem_M_BM25*) à ceux obtenus par les deux index sémantiques suivants :

1. *Sem_G_Kolte_tf*idf :* index issu de notre approche d'indexation sémantique basée sur la désambiguïsation globale dans laquelle l'approche de désambiguïsation des domaines est remplacée par celle de [Kolte et al., 09], pondéré par *tf*idf,*
2. *Sem_G_Kolte_BM25 :* issu de notre approche d'indexation sémantique basée sur la désambiguïsation globale en remplaçant notre méthode de désambiguïsation des domaines par la méthode proposée dans [Kolte et al., 09], pondéré par *Okapi-BM25,*
3. *Sem_M_Kolte_tf*idf :* index issu de notre approche d'indexation sémantique basée sur la désambiguïsation mixte dans laquelle l'approche de désambiguïsation des domaines est remplacée par celle de [Kolte et al., 09], pondéré par *tf*idf,*
4. *Sem_M_Kolte_BM25 :* issu de notre approche d'indexation sémantique basée sur la désambiguïsation mixte en remplaçant notre méthode de désambiguïsation des domaines par la méthode proposée dans [Kolte et al., 09], pondéré par *Okapi-BM25,*

Les résultats des comparaisons sont donnés à travers les figures 4.17 et 4.18 respectivement.

La figure 4.17 présente les résultats issus des index *Sem_G_Kolte_tf*idf* et *Sem_G_Kolte_BM25* relativement à ceux des baselines d'une part et à ceux de nos index sémantiques *Sem_G_tf*idf* et *Sem_G_BM25* d'autre part. De cette figure, il apparaît que les index sémantiques *Sem_G_Kolte_tf*idf* et *Sem_G_Kolte_BM25* produisent des résultats de recherche meilleurs que ceux obtenus par les baseline *Classic*tf*idf* et *Classic_BM25* respectivement. Les taux d'accroissement de la *MAP* sont de l'ordre de 5.25% pour *Sem_G_Kolte_tf*idf* et 3.52% pour *Sem_G_Kolte_BM25*.

Par conséquent, **on peut conclure que la désambiguïsation des domaines de [Kolte et al., 09] utilisée dans notre approche de désambiguïsation globale, a permis d'améliorer les résultats de la recherche par rapport à une indexation classique sans désambiguïsation**.

Par ailleurs, nos index sémantiques *Sem_G_tf*idf* et *Sem_G_BM25* présentent de meilleurs résultats à tous les points de précision comparés respectivement aux index sémantiques *Sem_G_Kolte_tf*idf* et *Sem_G_Kolte_BM25*. Des gains de performances significatifs sont observés pour l'index *Sem_G_tf*idf* par rapport à *Sem_G_Kolte_tf*idf* , de l'ordre de 100% pour *P@5*, 42% pour *P@10*, 10.83% pour *P@15*, 10% pour *P@20*, 46.57% pour *P@30*, 10% pour *P@100* et 5% pour la *MAP*. D'autre part, des gains de performances sont aussi observés pour l'index *Sem_G_BM25* par rapport à l'index *Sem_G_Kolte_BM25* de l'ordre de 12.14% pour *P@10*, 11.11% pour *P@20*, 46.57% pour *P@30*, 6.81% pour *P@100* et 2.04 % pour la *MAP*.

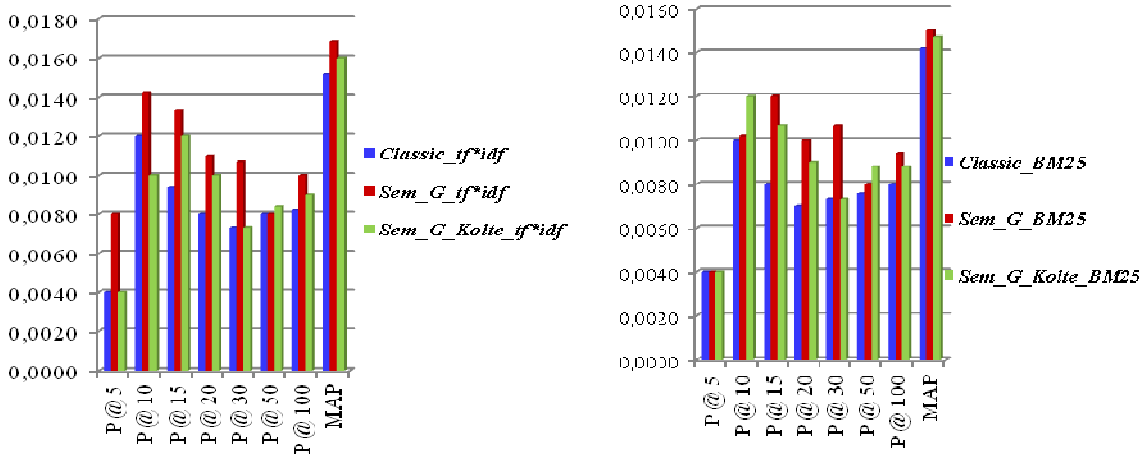


Figure 4.17: Résultats des comparaisons entre notre approche de désambiguïsation des domaines et l'approche de désambiguïsation des domaines de Kolte [Kolte, 09] dans l'indexation sémantique à base de concepts-sens issus de la désambiguïsation globale.

La figure 4.18 présente les résultats issus des index *Sem_M_Kolte_tf*idf* et *Sem_M_Kolte_BM25* relativement à ceux des baselines d'une part et à ceux de nos index sémantiques *Sem_M_tf*idf* et *Sem_M_BM25* d'autre part. De cette figure, il apparaît que les index sémantiques *Sem_M_Kolte_tf*idf* et *Sem_M_Kolte_BM25* apportent une amélioration (non significative) de la *MAP* par rapport aux baselines respectives *Classic_tf*idf* et

Classic_BM25 (de l'ordre de 0.65% pour *Sem_M_Kolte_tf*idf* et de l'ordre de 1.4% pour *Sem_M_Kolte_BM25* comparés). De ce fait, on peut déduire que la désambiguïsation des domaines de Kolte [Kolte et al., 08], utilisée dans notre indexation sémantique basée sur la désambiguïsation mixte, est légèrement meilleure qu'une indexation classique.

Par ailleurs, nos index sémantiques *Sem_M_tf*idf* et *Sem_M_BM25* apportent une légère amélioration de la *MAP* par rapport aux index sémantiques *Sem_M_Kolte_tf*idf* et *Sem_M_Kolte_BM25* respectivement, avec un gain de performance de l'ordre de 1.3% pour *Sem_G_tf*idf* et de l'ordre de 1.39% pour *Sem_G_BM25*.

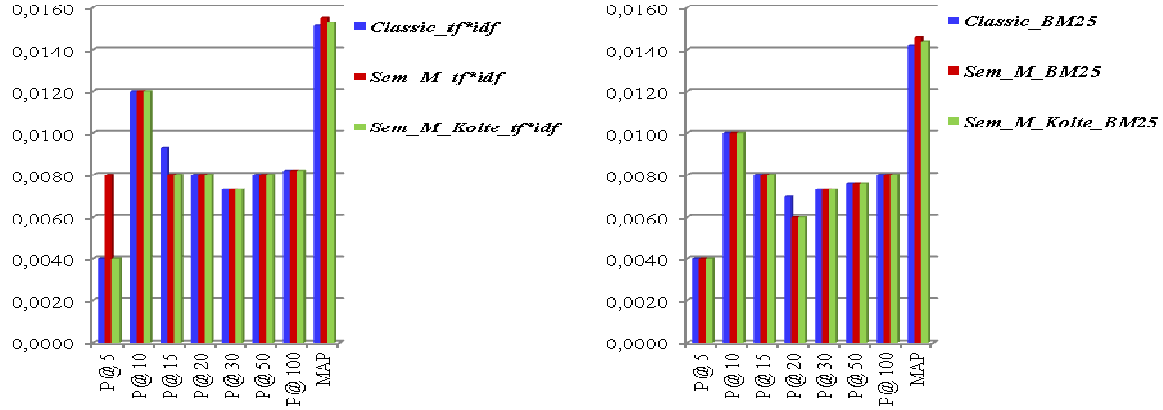


Figure 4.18: Résultats des comparaisons entre notre approche de désambiguïsation des domaines et l'approche de désambiguïsation des domaines de [Kolte, 09] dans l'indexation sémantique à base de concepts-sens issus de la désambiguïsation mixte.

De ces expérimentations, il ressort que notre approche de désambiguïsation des domaines est plus performante que l'approche de désambiguïsation des domaines proposée dans [Kolte et al., 09]. Ceci est probablement dû au fait que notre approche de désambiguïsation des domaines est plus précise que celle de [Kolte et al., 09], car tenant compte des relations sémantiques entre les domaines des mots contrairement à cette dernière.

4.4.2 Evaluation des approches de pondération des concepts dans TIME

La deuxième série d'expérimentations menées sur la collection TIME concerne l'évaluation de nos approches de pondération sémantique *Ct-Ict* et *Tidf* (définies en section 3.3.3.3). Nous avons testé ces deux schémas pour la pondération des concepts issus des différentes techniques de désambiguïsation proposées : locale, globale et mixte.

4.4.2.1 Evaluation de l'approche de pondération *Ct-Ict*

Notre objectif à travers cette évaluation est de comparer notre approche de pondération (sémantique) des concepts, par le schéma *Ct-Ict*, par rapport à leur pondération classique. Une étape préalable de choix de la valeur du paramètre de pondération α est nécessaire. Le paramètre α utilisé dans le schéma de pondération *Ct-Ict* (formule [3.7]), permet de balancer entre la fréquence d'un concept et son importance sémantique par rapport aux autres concepts dans le document (ou la requête). Pour déterminer la valeur adéquate de ce paramètre, nous avons affecté à α différentes valeurs (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 et 0.9) qui ont conduit à différents schémas de pondération paramétrés *Ct-Ict*(*α*). Chacun de ces schémas est utilisé pour pondérer les termes de notre index sémantique. Les index pondérés ainsi obtenus sont ensuite évalués à travers les résultats de recherche qu'ils retournent. L'index sémantique qui fournit les meilleurs résultats pour une valeur donnée à α, comparativement à des schémas de pondération classiques (*tf*idf* et *Okapi-BM25*) appliqués à l'index sémantique détermine la meilleure valeur de α. Dans nos évaluations, nous avons considéré les trois index suivants :

1. l'index *Sem_L_Ct-Ict* : index sémantique issu de notre méthode de désambiguïsation locale, pondéré par *Ct-Ict*,
2. l'index *Sem_G_Ct-Ict* : index sémantique issu de notre méthode de désambiguïsation globale, pondéré par *Ct-Ict*,
3. l'index *Sem_M_Ct-Ict* : index sémantique issu de notre méthode de désambiguïsation mixte, pondéré par *Ct-Ict*.

L'évaluation de ces index est présentée dans les sections suivantes.

a) Evaluation de *Sem_L_Ct-Ict*. Le tableau 4.3 présente les résultats de l'évaluation de notre index sémantique *Sem_L_Ct-Ict* pour différentes valeurs de α.

	Sem_L_Ct-Ict								
	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
P @ 1	**0.0600**	0.0400	0.0400	0.0400	0.0400	0.0200	0.0200	0.0200	0.0000
P @ 2	**0.0400**	0.0300	0.0300	0.0300	0.0300	0.0300	0.0200	0.0200	0.0100
P @ 3	**0.0267**	0.0200	0.0200	0.0200	0.0200	0.0200	0.0200	0.0133	0.0133
P @ 4	**0.0200**	0.0150	0.0150	0.0150	0.0150	0.0150	0.0150	0.0150	0.0150
P @ 5	**0.0160**	0.0160	0.0160	0.0160	0.0160	0.0160	0.0160	0.0120	0.0160
P @ 10	**0.0180**	0.0160	0.0160	0.0160	0.0160	0.0180	0.0140	0.0140	0.0100
P @ 15	0.0133	0.0120	0.0120	0.0133	0.0133	0.0133	0.0133	**0.0147**	0.0133
P @ 20	0.0100	0.0100	0.0100	0.0100	0.0100	**0.0110**	**0.0110**	**0.0110**	**0.0110**
P @ 30	**0.0127**	**0.0127**	0.0113	0.0113	0.0107	0.0107	0.0100	0.0107	0.0100
P @ 50	**0.0112**	0.0100	0.0100	0.0100	0.0092	0.0092	0.0084	0.0084	0.0092
P @ 100	**0.0100**	0.0096	0.0098	0.0098	0.0098	0.0098	0.0098	0.0098	0.0098
MP@x	**0.0216**	0.0178	0.0173	0.0174	0.0173	0.0157	0.0143	0.0135	0.0107
MAP	**0.0337**	0.0305	0.0304	0.0304	0.0303	0.0284	0.0275	0.0260	0.0213

Tableau 4.3 : Résultats d'évaluation de l'index sémantique Sem_L_Ct-Ict pour les différentes valeurs données à α.

Ces résultats montrent que l'index *Sem_L_Ct-Ict* pondéré par *Ct-Ict*(α=0.1) présente globalement les meilleures performances en particulier aux niveaux des précisions aux points *x* (*x* <=10 et *x* >=30), de la MAP et de la moyenne des précisions à *x* (*MP@x*). Aux rangs 15 et 20, les performances de cette pondération sont très proches des performances maximales (une différence non significative de l'ordre de 0.0014 pour *P@15* et de l'ordre de 0.001 pour *P@20*). Par conséquent, nous sélectionnons le schéma *Ct-Ict*(α=0.1), pour la pondération des concepts de la collection TIME issus de notre méthode de désambiguïsation locale.

L'index sémantique *Sem_L_Ct-Ict*(α=0.1) est ensuite évalué comparativement aux baselines *Classic_tf*idf* et *Classic_BM25* d'une part, et aux index sémantiques classiquement pondérés *Sem_L_tf*idf* et *Sem_L_BM25* d'autre part. Les résultats de ces comparaisons sont donnés à travers la figure 4.19.

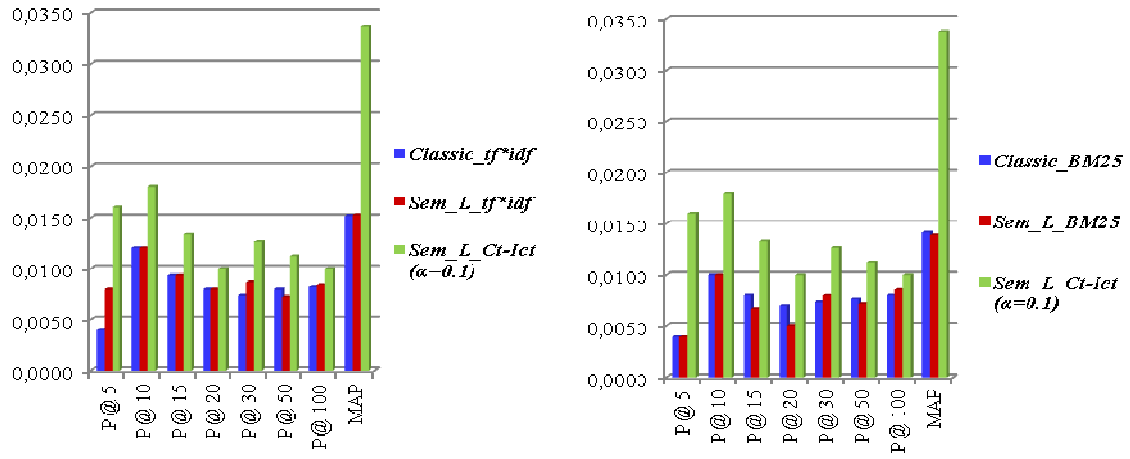


Figure 4.19: Apport de notre pondération sémantique des concepts (issus de la désambiguïsation locale) avec le schéma *Ct-Ict*.

De la figure 4.19, il ressort que l'index sémantique *Sem_L_Ct-Ict* améliore les résultats de la recherche à tous les points de précision, par rapport aux baselines *Classic_tf*idf* et *Classic_BM25*. Les gains de performances significatifs par rapport à *Classic_tf*idf* sont de 300%, 50%, 43%, 25%, 73.97%, 40%, 21.95% et 123.02% pour les précisions respectives *P@5, P@10, P@20, P@30, P@50, P@100* et la MAP. Les gains de performances significatifs par rapport à *Classic_BM25* sont de 300%, 80%, 66.25%, 42.85%, 73.97%, 47.36%, 25% et 137.32% pour les précisions *P@5, P@10, P@20, P@30, P@50, P@100* et la *MAP* respectivement.

Par ailleurs, on note que l'index sémantique *Sem_L_Ct-Ict* est plus performant que les index sémantiques classiquement pondérés Sem*_L_tf*idf* et *Sem_L_BM25*, avec des taux d'accroissement de la MAP de 121.71% et 142.44% respectivement. **De ces résultats, il ressort que notre pondération des concepts (issus de la désambiguïsation locale) par le schéma *Ct-Ict(α=0.1)* est plus performante qu'une pondération classique.**

b) Evaluation de *Sem_G_Ct-Ict* .

Le tableau 4.4 présente les résultats d'évaluation de notre index sémantique Sem_G_Ct-Ict pour les différentes valeurs affectées à α.

	Sem_G_Ct-Ict								
	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**
P @ 1	**0.0400**	**0.0400**	**0.0400**	**0.0400**	**0.0400**	**0.0400**	**0.0400**	**0.0400**	**0.0400**
P @ 2	**0.0400**	**0.0400**	0.0300	0.0200	0.0300	0.0300	0.0200	0.0200	0.0300
P @ 3	**0.0267**	**0.0267**	**0.0267**	0.0200	0.0267	0.0267	0.0200	0.0133	0.0200
P @ 4	**0.0200**	**0.0200**	**0.0200**	0.0150	**0.0200**	**0.0200**	**0.0200**	**0.0200**	**0.0200**
P @ 5	**0.0240**	**0.0240**	**0.0240**	**0.0240**	**0.0240**	**0.0240**	**0.0240**	0.0200	0.0200
P @ 10	0.0200	**0.0220**	**0.0220**	**0.0220**	**0.0220**	0.0200	0.0200	0.0200	0.0140
P @ 15	0.0160	0.0160	0.0160	0.0160	0.0160	**0.0173**	**0.0173**	**0.0173**	0.0160
P @ 20	0.0120	0.0120	0.0120	0.0130	0.0130	0.0130	**0.0140**	**0.0140**	0.0130
P @ 30	0.0127	**0.0133**	0.0127	0.0127	0.0127	**0.0133**	0.0127	0.0120	0.0120
P @ 50	**0.0104**	0.0100	0.0100	0.0100	0.0103	0.0100	0.0096	0.0096	**0.0104**
P @ 100	0.0100	0.0100	0.0100	0.0100	0.0102	0.0104	**0.0106**	**0.0106**	**0.0106**
MP@x	0.0211	**0.0213**	0.0203	0.0184	0.0204	0.0204	0.0189	0.0179	0.0187
MAP	0.0349	**0.0354**	0.0348	0.0348	0.0348	0.0348	0.0336	0.0315	0.0315

Tableau 4.4 : Résultats d'évaluation de l'index sémantique *Sem_G_Ct-Ict* pour les différentes valeurs données à α.

Ces résultats montrent que l'index *Sem_G_Ct-Ict* pondéré par *Ct-Ict* (α=0.2) présente globalement les meilleures performances en particulier aux niveaux des précisions P@10 et au *P@30*, au niveau de la MAP et au niveau de la moyenne des précisions à *x* (*MP@x*). Les performances aux points *x* (*x*=15, 20, 50 et 100) sont très proches des valeurs maximales.

Par conséquent, nous sélectionnons le schéma *Ct-Ict*(α=0.2), pour la pondération des concepts de la collection TIME issus de notre méthode de désambiguïsation globale.

L'index sémantique *Sem_G_Ct-Ict*(α=0.2) est ensuite évalué comparativement aux baselines *Classic_tf*idf* et *Classic_BM25* d'une part, et aux index sémantiques classiquement pondérés *Sem_G_tf*idf* et *Sem_G_BM25* d'autre part. Les résultats de ces comparaisons sont donnés à travers la figure 4.20. Des graphiques de cette figure, il apparaît que les résultats obtenus par notre index sémantique *Sem_G_Ct-Ict*(α=0.2) sont meilleurs à tous les points de précisions que ceux obtenus par les baselines. Les taux d'accroissement de la *MAP* sont de 132.89% par rapport à *Classic_tf*idf* et de 149.29% par rapport à *Classic_BM25*.

De plus, *Sem_G_Ct-Ict*(α=0.2) présente de meilleurs performances de recherche à tous les points de précisions comparativement aux index sémantiques *Sem_G_tf*idf* et *Sem_G_BM25*. Des taux d'accroissement significatifs sont observés d'une part par rapport à *Sem_G_tf*idf* de l'ordre de 200% pour *P@5*, 54.92% pour *P@10*, 20.30% pour *P@15*, 9.09% pour *P@20*, 24.29% pour *P@30*, 25% pour *P*@50 et 110.07% pour la *MAP*, et d'autre part par rapport à *Sem_G_BM25* de l'ordre de 500% pour *P@5*, 115.68% pour *P@10*, 33.33% pour *P@15*, 20% pour *P@20*, 24.29% pour *P@30*, 25% pour *P*@50, 6.38% pour *P*@100 et 136% pour la *MAP*.

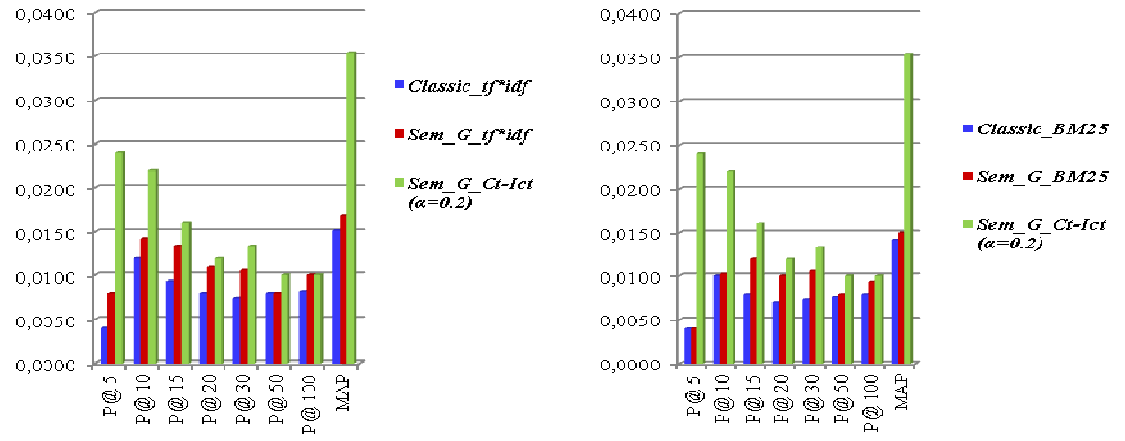


Figure 4.20: Apport de notre pondération sémantique des concepts (issus de la désambiguïsation globale) avec le schéma *Ct-Ict*.

De ces résultats, il ressort que notre pondération des concepts (issus de la désambiguïsation globale) par le schéma *Ct-Ict(α=0.2)* est plus performante que la pondération classique. Cette performance peut s'expliquer par le fait que notre poids sémantique *Ct-Ict* qui tient compte de la centralité d'un concept (mesurée par ses similarités sémantiques avec les autres concepts) est plus expressif et plus représentatif de l'importance d'un concept qu'un schéma de pondération classique *tf*idf* et *Okapi-BM25*.

c) Evaluation de *Sem_M_Ct-Ict*.

Les résultats de l'évaluation de notre index sémantique *Sem_M_Ct-Ict* pour les différentes valeurs de α, illustrés à travers le tableau 4.5, montrent que les meilleurs résultats de recherche sont obtenus pour la valeur α=0.1. Les performances de la pondération *Ct-Ict* (α=0.1*)* sont maximales aux points *x* (*x*=1, 2, 3, 4, 5 et 30), au niveau de la *MAP* et de la moyenne des précisions à *x* (*MP@x*). Par conséquent, nous sélectionnons le schéma *Ct-Ict*(α=0.1), pour la pondération des concepts de la collection TIME issus de notre méthode de désambiguïsation mixte.

L'index sémantique *Sem_M_Ct-Ict*(α=0.1) est ensuite évalué comparativement aux baselines *Classic_tf*idf* et *Classic_BM25* d'une part, et aux index sémantiques classiquement pondérés *Sem_M_tf*idf* et *Sem_M_BM25* d'autre part. Les résultats de ces comparaisons sont donnés à travers la figure 4.21. Des graphiques de cette figure, il apparaît que les résultats obtenus par notre index sémantique *Sem_M_Ct-Ict*(α=0.1) sont meilleurs à tous les points de précisions que ceux obtenus par les baselines. Les taux d'accroissement de la *MAP* sont de 85.52% par rapport à *Classic_tf*idf* et de 98.59% par rapport à *Classic_BM25*.

	Sem_M_Ct-Ict								
	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**
P @ 1	**0.0400**	**0.0400**	**0.0400**	**0.0400**	**0.0400**	**0.0400**	0.0200	0.0200	0.0200
P @ 2	**0.0200**	**0.0200**	**0.0200**	**0.0200**	**0.0200**	**0.0200**	**0.0200**	**0.0200**	0.0100
P @ 3	**0.0200**	**0.0200**	**0.0200**	**0.0200**	**0.0200**	**0.0200**	**0.0200**	**0.0200**	**0.0200**
P @ 4	**0.0150**	**0.0150**	**0.0150**	**0.0150**	**0.0150**	**0.0150**	**0.0150**	**0.0150**	**0.0150**
P @ 5	**0.0160**	0.0120	0.0120	0.0120	0.0120	0.0120	0.0120	0.0120	**0.0160**
P @ 10	0.0140	0.0140	0.0140	0.0140	0.0120	0.0120	0.0140	**0.0160**	0.0120
P @ 15	0.0107	0.0107	0.0107	0.0107	**0.0133**	**0.0133**	**0.0133**	**0.0133**	**0.0133**
P @ 20	0.0120	**0.0130**	**0.0130**	**0.0130**	**0.0130**	**0.0130**	0.0120	0.0110	0.0110
P @ 30	**0.0113**	0.0107	0.0107	0.0107	0.0100	0.0100	0.0100	0.0107	0.0093
P @ 50	0.0100	0.0100	0.0100	0.0100	**0.0104**	0.0096	0.0096	0.0092	0.0096
P @ 100	0.0092	0.0092	0.0092	0.0094	0.0092	**0.0096**	**0.0096**	**0.0096**	0.0090
MP@x	**0.0162**	0.0159	0.0159	0.0159	0.0159	0.0159	0.0141	0.0143	0.0132
MAP	**0.0282**	0.0275	0.0275	0.0274	0.0274	0.0272	0.0254	0.0248	0.0224

Tableau 4.5 : Résultats d'évaluation de l'index sémantique *Sem_M_Ct-Ict* pour les différentes valeurs données à α.

De plus, *Sem_M_Ct-Ict*(α=0.1) présente de meilleurs performances de recherche comparativement aux index sémantiques *Sem_M_tf*idf* et *Sem_M_BM25*. Des gains de performance significatifs sont observés d'une part par rapport à *Sem_M_tf*idf* de l'ordre de de 100%, 16.66%, 33.75%, 50%, 54.79%, 25%, 12.19% et 81.93% pour les précisions *P@5, P@10, P@20, P@30, P@50, P@100* et la *MAP* respectivement, et d'autre part par rapport à *Sem_M_BM25* de l'ordre de 300%, 40%, 33.75%, 100%, 54.79%, 31.57%, 15% et 93.15% pour les précisions *P@5, P@10, P@20, P@30, P@50, P@100* et la *MAP* respectivement.

De ces résultats, il ressort que notre pondération des concepts (issus de la désambiguïsation mixte) par le schéma *Ct-Ict(α=0.1)* est plus performante que la pondération classique.

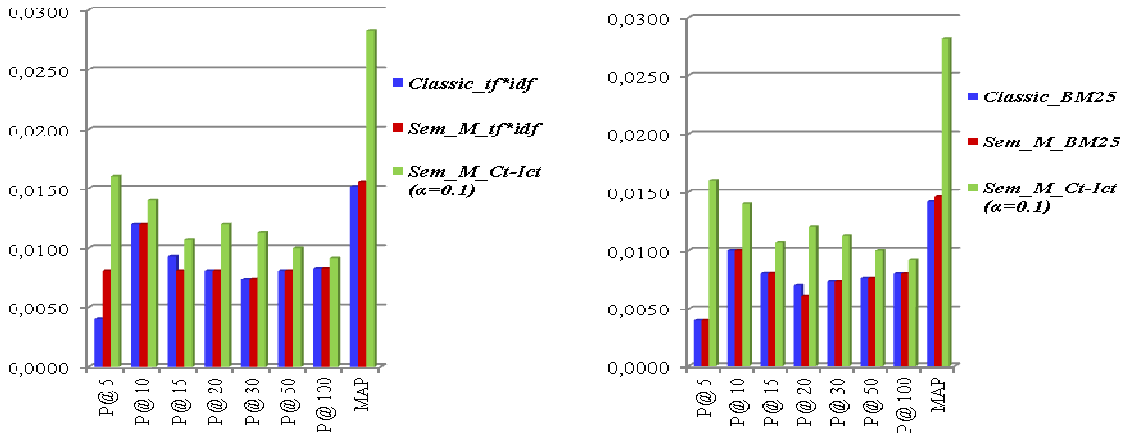


Figure 4.21: Apport de notre pondération sémantique des concepts (issus de la désambiguïsation mixte) avec le schéma *Ct-Ict*.

Par ailleurs, les comparaisons réalisées entre les index sémantiques, issus de nos différents index sémantiques basés sur la pondération *Ct-Ict* (illustrées en figure 4.22) montrent que :

- l'index *Sem_G_Ct-Ict*(α=0.2) est plus performant que les index *Sem_L_Ct-Ict*(α=0.1) et *Sem_M_Ct-Ict*(α=0.1). Des taux de performances variables sont observés par rapport aux premiers documents restitués par le système, de l'ordre de 50%, 22.22%, 20.30% et 5.04%. pour les précisions respectives *P@5, P@10 , P@15* et *MAP* comparativement à *Sem_L_Ct-Ict*. Tandis que les gains de performance observés par rapport à *Sem_M_Ct-Ict* sont de l'ordre de 50%, 57.14%, 49.53% et 25.53% pour les précisions *P@5, P@10, P@15 et MAP* respectivement.

- l'index *Sem_L_Ct-Ict*(α=0.1) est plus performant que l'index *Sem_M_Ct-Ict*(α=0.1). Des taux de performances significatifs de l'ordre de 28.57%, 24.29%, 20% et 19.50% sont observés pour les précisions *P@10, P@15, P@20* et la *MAP* respectivement.

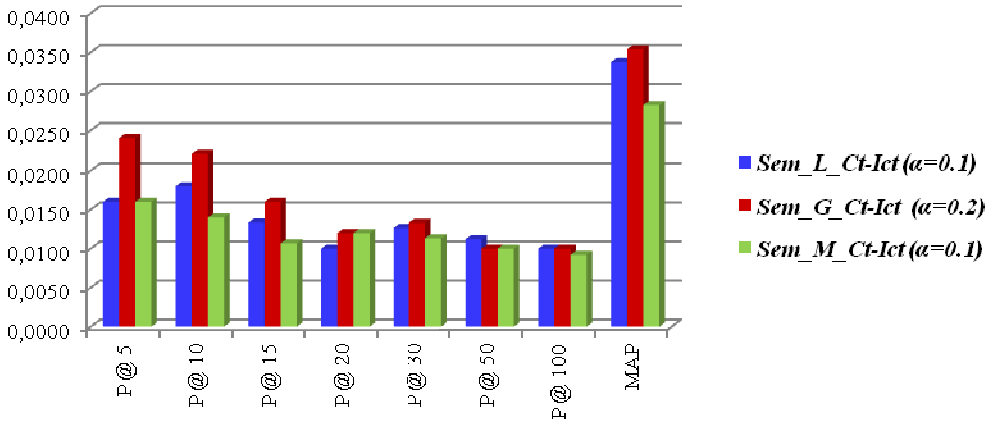


Figure 4.22: Résultats des comparaisons entre les index sémantiques à base de concepts, issus des différentes méthodes de désambiguïsation (locale, globale, mixte), pondérés par le schéma *Ct-Ict*.

A ce niveau de l'évaluation, nous concluons que quelque soit l'approche de désambiguïsation utilisée, notre pondération des concepts par le schéma *Ct-Ict* est toujours plus performante que la pondération classique. Cette performance peut s'expliquer par le fait que notre poids sémantique *Ct-Ict* qui tient compte de la centralité d'un concept (mesurée par ses similarités sémantiques avec les autres concepts) est plus expressif et plus représentatif de l'importance d'un concept qu'un schéma de pondération classique *tf*idf* et *Okapi-BM25*. De plus, il apparaît qu'une indexation sémantique basée sur la désambiguïsation globale donne de meilleurs résultats qu'une indexation basée sur la désambiguïsation locale ou mixte.

4.4.2.2 Evaluation de l'approche de pondération *Tidf*

Dans cette section, nous présentons l'évaluation de notre schéma de pondération *Tidf* (décrit en section 3.3.3.3.2). Pour cela, nous avons comparé les résultats issus de notre index sémantique pondéré par *Tidf* à ceux obtenus à partir des baselines d'une part et à partir de notre index sémantique avec classiquement pondéré (deux schéma classiques ont été utilisés : *tf*idf* et *Okapi-BM25*) d'autre part. Nos principaux objectifs à travers ces comparaisons consistent à :

- mesurer l'apport de l'indexation sémantique par les concepts pondérés par *Tidf* par rapport à l'indexation classique basée mots clés pondérés par *tf*idf* ou par *Okapi-BM25*,

- tester l'efficacité de notre pondération sémantique des concepts par *Tidf* par rapport à leur pondération classique par *tf*idf* et *Okapi-BM25*.

Dans nos expérimentations, nous avons considéré les index sémantiques suivants :

1. l'index *Sem_L_Tidf* : index sémantique issu de notre méthode de désambiguïsation locale, pondéré par *Tidf*,
2. l'index *Sem_G_Tidf* : index sémantique issu de notre méthode de désambiguïsation globale, pondéré par *Tidf*,
3. l'index *Sem_M_Tidf* : index sémantique issu de notre méthode de désambiguïsation mixte, pondéré par *Tidf*,

Les résultats d'évaluation de ces index seront présentés dans les sections suivantes.

a) Evaluation de *Sem_L_Tidf* .

Les résultats de l'évaluation de notre index sémantique *Sem_L_Tidf* ainsi que ceux des baselines *Classic_tf*idf* et *Classic_BM25* et des index sémantiques classiquement pondérés *Sem_L_tf*idf* et *Sem_L_BM25* sont présentés à travers les graphiques de la figure 4.23. De ces graphiques, il ressort que les résultats obtenus par l'index *Sem_L_Tidf* sont nettement meilleurs, à tous les points de précision, que ceux obtenus par les baselines *Classic_tf*idf* et *Classic_BM25*. Des taux d'accroissement significatifs sont observés par rapport à la baseline *Classic_tf*idf* de l'ordre de 150%, 44.16%, 101.07%, 108.75%, 134.24%, 122.50%, 32.92% et 56.57% pour les précisions *P@5, P@10,P@15, P@20, P@30, P@50, P@100* et la *MAP* respectivement. Les gains de performances réalisés par rapport à *Classic_BM25* varient de 150%, 73%, 101.07%, 128.75%, 138.24%, 134.21%, 36.25% et 67.60% pour les précisions *P@5, P@10,P@15, P@20, P@30, P@50, P@100* et la *MAP* respectivement. En outre, on note une amélioration de la *MAP* dans l'index *Sem_L_Tidf* par rapport à *Sem_L_tf*idf* et *Sem_L_BM25* respectivement de l'ordre de 56.57% et de 71.22%.

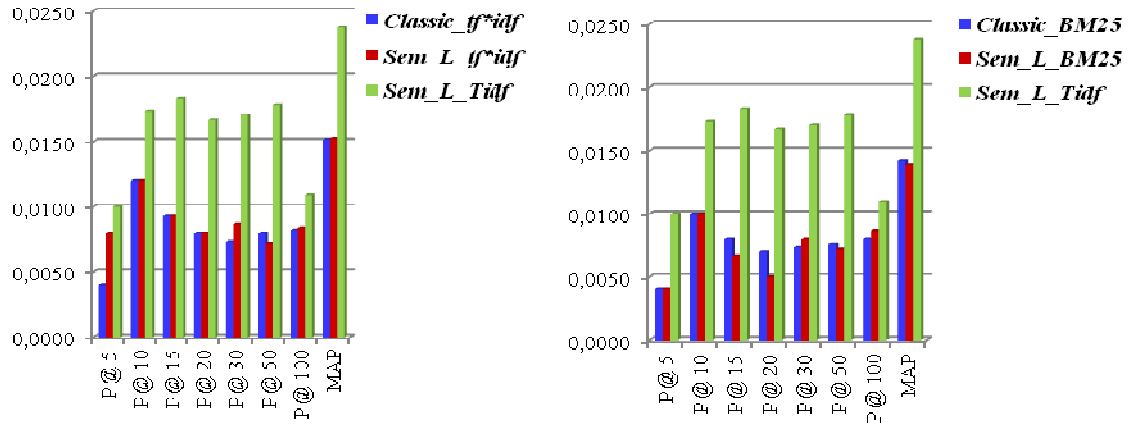


Figure 4.23: Apport de notre pondération sémantique des concepts (issus de la désambiguïsation locale) avec le schéma *Tidf*.

De ces résultats, nous concluons que notre indexation sémantique basée sur les concepts issus de la désambiguïsation locale et pondérés par notre schéma *Tidf* est plus efficace qu'une indexation classique basée mots-clés. De plus, la pondération des concepts par le schéma *Tidf* apporte des résultats meilleurs que ceux apportés par leur pondération par les schémas classiques (*tf*idf* et *Okapi-BM25*).

b) Evaluation de *Sem_G_Tidf*

Les résultats de la recherche issus respectivement de *Sem_G_Tidf* et ceux issus des baselines et des index sémantiques classiquement pondérés *Sem_G_tf*idf* et *Sem_G_BM25* sont représentées à travers les graphiques de la figure 4.24.

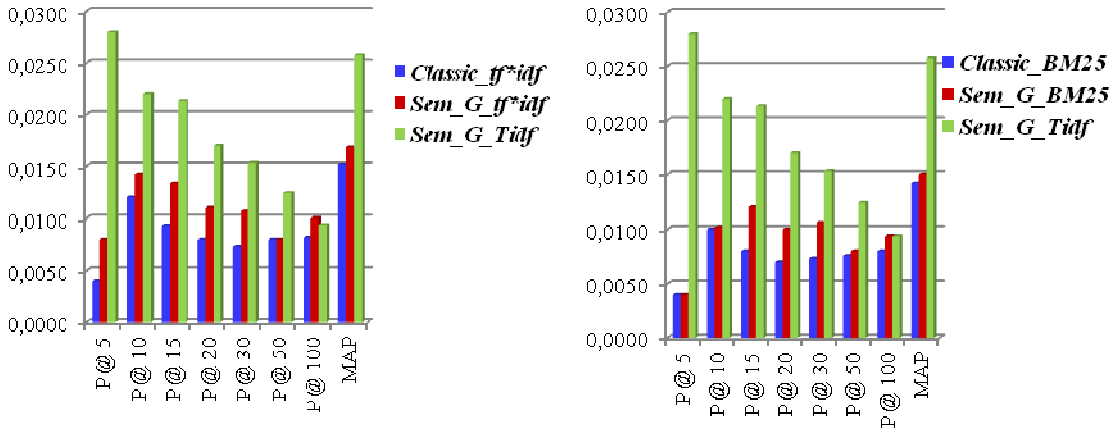


Figure 4.24: Apport de notre pondération sémantique des concepts (issus de la désambiguïsation globale) avec le schéma *Tidf*.

De ces résultats, il apparaît que *Sem_G_Tidf* est plus performant que les index classiques *Classic_tf*idf* et *Classic_BM25*. Des taux d'amélioration significatifs sont observés par rapport à *Classic_tf*idf* et *Classic_BM25* de l'ordre de 600% pour *P@5*, 83.33% pour *P@10*, 123.65% pour *P@15*, 112.50% pour *P@20*, 109.58% pour *P@30*, 55.01% pour *P*@50 et 69.73% pour la *MAP*. Les taux d'accroissement obtenus par rapport à *Classic_BM25* sont de 600% pour *P@5*, 120% pour *P@10*, 166.25% pour *P@15*, 142.85% pour *P@20*, 109.58% pour *P@30*, 63.15% pour *P*@50 et 81.69% pour la *MAP* respectivement.

De plus, *Sem_G_Tidf* produit des résultats nettement meilleurs que *Sem_G_tf*idf* et *Sem_G_BM25*. Les gains de performances réalisés par rapport *Sem_G_tf*idf* sont de 250%, 65%, 60.15%, 54.54%, 42.99%, 55% et 53.57% pour les précisions *P@5, P@10, P@15, P@20, P@30, P@50* et la *MAP* respectivement. Les gains de performance par rapport à *Sem_G_BM25* sont de l'ordre de 600%, 115.68%, 77.5%, 70%, 42.99%, 55% et 72% pour les précisions *P@5, P@10, P@15, P@20, P@30, P@50* et la *MAP* respectivement.

De cette évaluation, il ressort que notre indexation sémantique basée sur les concepts issus de la désambiguïsation globale et pondérés par notre schéma *Tidf* est plus efficace qu'une indexation classique basée mots-clés. De plus, la pondération des concepts par le schéma *Tidf* apporte de meilleurs résultats à la recherche que leur pondération par les schémas classiques *tf*idf* et *Okapi-BM25*.

c) Evaluation de *Sem_M_Tidf*

Les graphiques de la figure suivante (figure 4.25) présente les résultats de la recherche issus respectivement de *Sem_M_Tidf*, des baselines *Classic_tf*idf* et *Classic_BM25*, et des index sémantiques *Sem_M_tf*idf* et *Sem_M_BM25* classiquement pondérés.

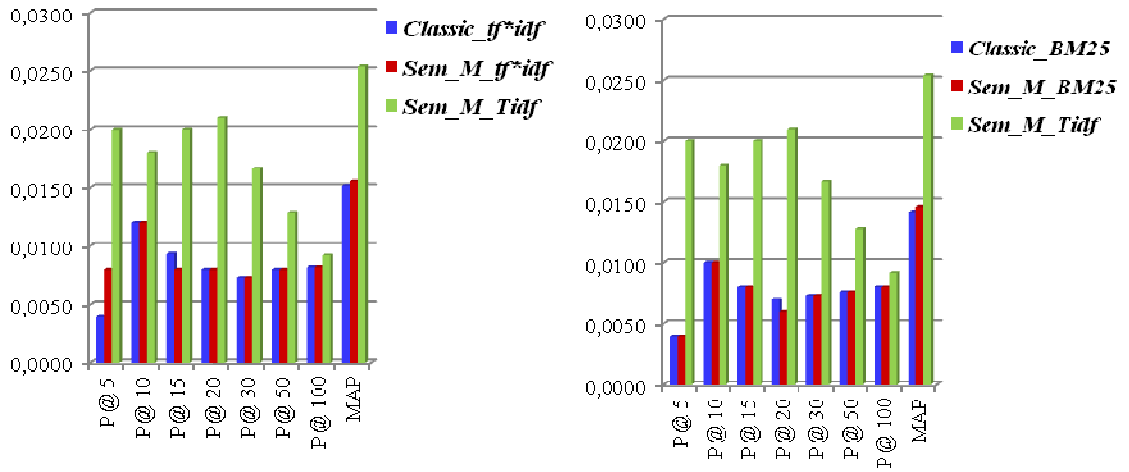


Figure 4.25: Apport de notre pondération sémantique des concepts (issus de la désambiguïsation mixte) avec le schéma *Tidf*.

Ces résultats montrent clairement que l'index sémantique *Sem_M_Tidf* est plus performant que *Classic_tf*idf* et *Classic_BM25*. Des taux d'accroissement significatifs sont observés par rapport à *Classic_tf*idf* de l'ordre de 400% pour *P@5,* 59.90% pour *P@10,* 115.05% pour *P@15,* 162.50% pour *P@20,* 128.76% pour *P@30,* 58.53% pour *P@*50 et 67.10% pour la *MAP*. Les taux d'amélioration significatifs obtenus par rapport à *Classic_BM25* sont de l'ordre de 400% pour *P@5,* 80% pour *P@10,* 150% pour *P@15*, 250% pour *P@20,* 128.76% pour *P@30,* 68.42% pour *P@*50 et 78.87% pour la *MAP*.

De plus, l'index *Sem_M_Tidf* donne de meilleurs résultats de recherche, à tous les points de précision notamment dans les premiers documents restitués, comparativement à *Sem_M_tf*idf* et à *Sem_G_BM25*. Les gains de performances réalisés par rapport *Sem_G_tf*idf* sont de 150%, 50%, 150%, 162.50%, 128.76%, 60% et 63.87% pour les précisions *P@5, P@10, P@15, P@20, P@30, P@50* et *MAP* respectivement. En outre, les taux d'accroissement réalisés par rapport à *Sem_M_BM25* sont de l'ordre de 400%, 80%, 150%, 250%, 128.76% et 68.42% et 73.97% pour les précisions *P@5, P@10, P@15, P@20, P@30, P@50* et *MAP* respectivement.

De ces résultats, nous déduisons que notre indexation sémantique basée sur les concepts, issus de la désambiguïsation mixte et pondérés par notre schéma *Tidf*, est plus performante qu'une indexation classique. De plus, la pondération des concepts par le schéma de pondération *Tidf* apporte de meilleurs résultats de recherche par rapport à leur pondération classique par *tf*idf* ou *Okapi-BM25*.

A cette échelle de l'évaluation, nous concluons que, quelque soit l'approche de désambiguïsation pour l'identification des concepts, notre pondération sémantique *Tidf* produit de meilleures performances que les schémas classiques *tf*idf* et *Okapi-BM25*.

Par ailleurs, en comparant entre les résultats issus des différents index sémantiques *Sem_L_Tidf, Sem_G_Tidf* et *Sem_M_Tidf* (résultats représentés à travers la figure 4.26), il apparaît que :

- l'index *Sem_G_Tidf* présente de meilleurs résultats aux rangs *x* (*x*<=15) par rapport aux index *Sem_L_Tidf* et *Sem_M_Tidf*. Des taux d'accroissement significatifs sont observés par rapport à *Sem_L_Tidf* de l'ordre de 50%, 27.16%, 16.39% et 18.18% pour les précisions *P@5, P@10, P@15* et MAP respectivement. Des taux d'accroissement sont aussi observés par rapport à *Sem_M_Tidf* sont de 40%, 22.22% et 6.5% pour les précisions *P@5, P@10* et *P@15* respectivement. Une amélioration non significative de la *MAP* de l'ordre de 1.57% est aussi notée.

- l'index *Sem_M_Tidf* est légèrement plus performant que *Sem_L_Tidf* avec un taux d'accroissement de la *MAP* de l'ordre de 6.72%.

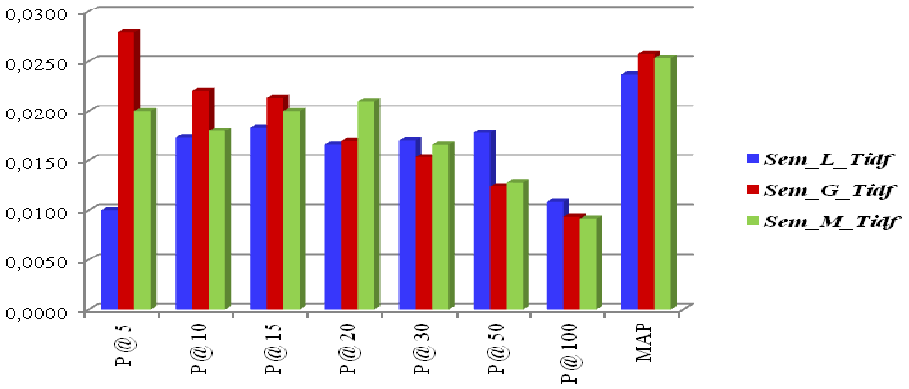


Figure 4.26: Résultats des comparaisons entre les index sémantiques à base de concepts, issus des différentes méthodes de désambiguïsation (locale, globale, mixte), pondérés par le schéma *Tidf*

4.4.2.3 Etude comparative entre les approches de pondération sémantique : *Ct-Ict* et *Tidf*

Les évaluations de nos approches de pondération sémantique *Ct-Ict* et *Tidf* ont montré que ces deux schémas de pondération sont d'un apport certain dans l'amélioration des performances de la recherche. Néanmoins, nous nous sommes posés la question suivante : « laquelle de ces deux approches *Ct-Ict* ou *Tidf* est la plus performante? ».

Pour répondre à cette question, nous avons comparé les résultats issus des index *Sem_L_Tidf*, *Sem_G_Tidf* et *Sem_M_Tidf* avec ceux issus des index respectifs *Sem_L_Ct-Ict*(α=0.1), *Sem_G_Ct-Ict*(α=0.2) et *Sem_M_Ct-Ict*(α=0.1). Les résultats de ces comparaisons sont représentés à travers les graphiques de la figure 4.27.

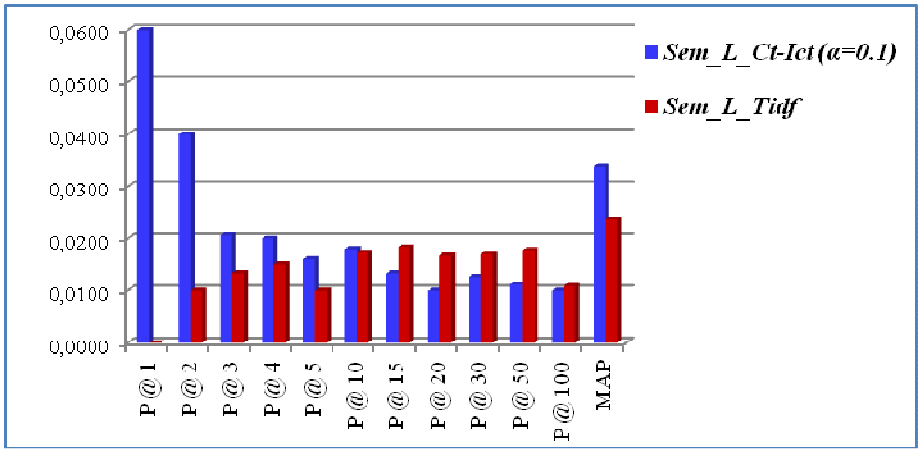
De ces résultats, il apparaît que :

- *Sem_L_Ct-Ict*(α=0.1) présente de meilleures performances aux précisions *P@x* (*x*<=10) par rapport à *Sem_L_Tidf*. Les taux d'amélioration observés sont de 300%, 55.63%, 33.33%, 60%, 4.04% et 41.59% pour les précisions *P@2*,*P@3*, *P@4*, *P@5*, *P@10* et MAP respectivement. Au rang 1, la précision atteint la valeur 0.0600 tandis qu'elle est nulle pour *Sem_L_Tidf*. Notons cependant que ses résultats sont en deçà de ceux de *Sem_L_Tidf* à partir du point *P@15*. Des taux de décroissement de l'ordre de -37.47%pour *P@15*, -66,66% pour *P@20*, -34.74% pour *P@30*, -58.92% pour *P*@50 et -9% pour *P@100* sont observés.

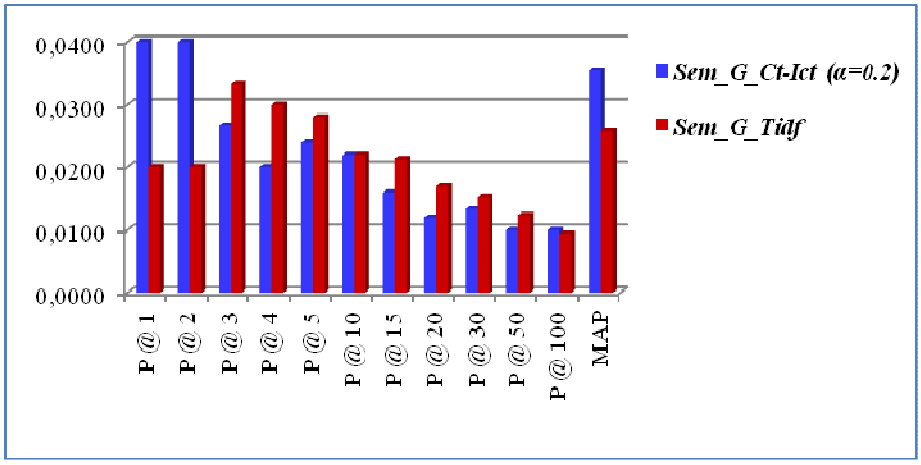
- *Sem_G_Ct-Ict*(α=0.2) présente de meilleurs performances comparativement à *Sem_G_Tidf* seulement aux rangs 1 et 2 avec un taux d'accroissement de 100%. Le gain de performance de la *MAP* est de 37.20%. Cependant, les résultats issus de *Sem_G_Tidf*, comparés à ceux de *Sem_G_Ct-Ict*, sont meilleurs à tous les points de précision *P@x* tels que *x*>=3. Les taux d'amélioration de *Sem_Tidf* par rapport à *Sem_G_Ct-Ict* sont de 24.71% pour *P@3*, 50% pour *P@4*, 16.66% pour *P@5*, 33.12% pour *P*@15, 41.66% pour *P@20*, 15.03% pour *P@30* et 24% pour *P@50*.

- *Sem_M_Ct-Ict*(α=0.1) présente de meilleurs résultats de recherche par rapport à *Sem_M_Tidf* au point *P@3* avec un taux d'accroissement de 50.37%. On ne note cependant pas d'améliorations aux points *P@x* (pour *x* = 1, 2, 4, 100), mais il n'y a pas de dégradation des performances. Cependant, *Sem_M_Tidf* est plus performant aux points P@x (pour x= 5, 15, 20, 30 et 50) et produit des gains de performances par rapport à *Sem_M_Ct-Ict* de l'ordre de 25% pour *P@5*, 28.57% pour *P@10*, 86.91% pour *P@15*, 75% pour *P*@20, 47.78% pour *P@30* et 28% pour *P@50*.

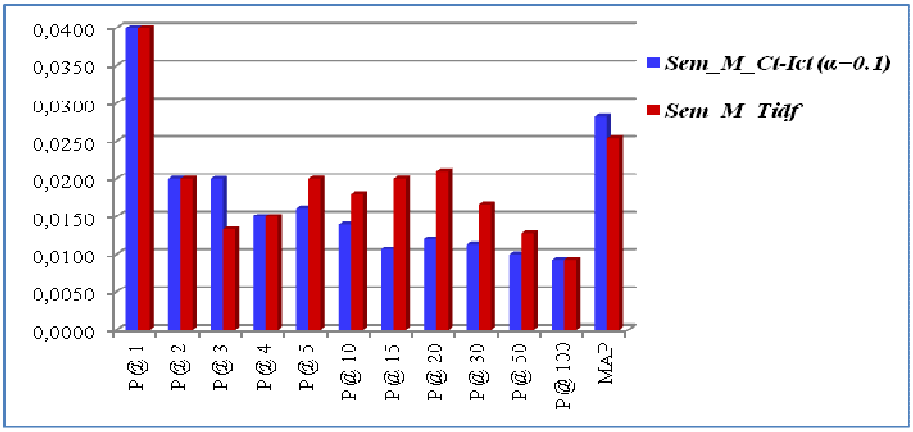
De cette étude, nous concluons que la pondération *Ct-Ict* apporte globalement de meilleures performances par rapport à la pondération *Tidf* dans les premiers documents restitués par le système dans la collection TIME. Ceci signifie que la pondération *Ct-Ict* a permis de sélectionner un nombre important de documents pertinents dans les premiers rangs ce qui n'est pas le cas de la pondération *Tidf*.



(a) *Sem_L_Ct-Ict (α=0.1)* vs *Sem_L_Tidf*



(b) *Sem_G_Ct-Ict (α=0.2)* vs *Sem_G_Tidf*



(c) *Sem_M_Ct-Ict (α=0.1)* vs *Sem_M_Tidf*

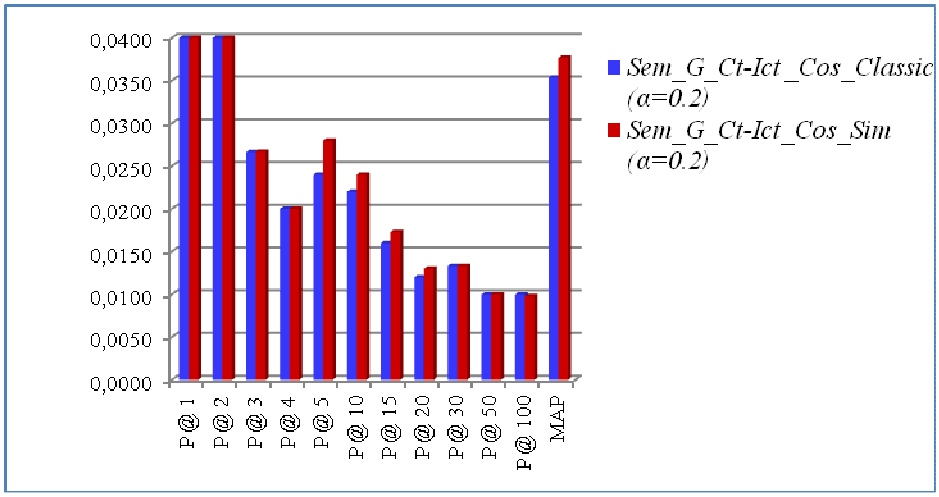
Figure 4.27: Résultats des comparaisons entre l'approche de pondération sémantique *Ct-Ict* et l'approche de pondération sémantique *Tidf*.

4.4.3 Evaluation de la mesure sémantique du score d'appariement documents-requête dans TIME

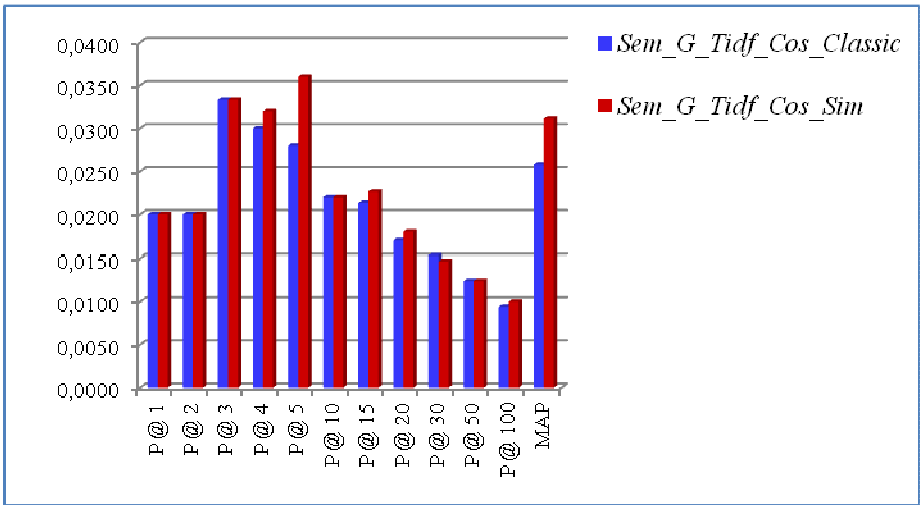
Les dernières expérimentations réalisées sur la collection TIME concernent l'évaluation du score sémantique d'appariement (ou de pertinence) document-requête proposé (formule 3.12). Dans ces expérimentations, les documents et les requêtes de la collection TIME sont indexés par notre index sémantique basé sur la désambiguïsation globale (vu que c'est cette dernière qui offre les meilleures performances). Les deux schémas de pondération que nous avons proposés, *Ct-Ict* et *Tidf*, ont été successivement utilisés pour pondérer les termes de cet index sémantique. Nous avons alors considéré les modèles de recherche suivants, caractérisés chacun par l'index et le modèle d'appariement document-requête utilisés:

- *Sem_G_Ct-Ict_Cos_Sim(α=0.2)*: dans ce modèle, l'index utilisé est l'index sémantique *Sem_G_Ct-Ict(α=0.2)* pondéré par *Ct-Ict* (pour α=0.2). Le modèle d'appariement repose sur le score sémantique de pertinence proposé au préalable (formule 3.12).
- *Sem_G_Tidf_Cos_Sim*: dans ce modèle, l'index utilisé est l'index sémantique *Sem_G_Tidf* pondéré par le schéma *Tidf*. Le modèle d'appariement repose sur le score sémantique de pertinence proposé au préalable (formule 3.12).
- *Sem_G_Ct-Ict_Cos_Classic(α=0.2)*: dans ce modèle, l'index utilisé est l'index sémantique *Sem_G_Ct-Ict(α=0.2)* pondéré par *Ct-Ict* (pour α=0.2). Le modèle d'appariement repose sur la mesure de cosinus du modèle vectoriel [Salton et al., 83].
- *Sem_G_Tidf_Cos_Classic*: dans ce modèle, l'index utilisé est l'index sémantique *Sem_G_Tidf* pondéré par le schéma *Tidf*. L'évaluation des requêtes de la collection, dans ce modèle, repose sur la mesure de cosinus du modèle vectoriel [Salton et al., 83]..

Les résultats de recherche issus des modèles *Sem_G_Ct–Ict_Cos_Sim (α=0.2)* et *Sem_G_Tidf_Cos_Sim* sont comparés respectivement à ceux de deux modèles *Sem_G_Ct-Ict_Cos_Classic(α=0.2)* et *Sem_G_Tidf_Cos_Classic*. L'objectif étant de comparer notre évaluation sémantique des requêtes par rapport à une évaluation classique basée sur la mesure du cosinus. Les résultats des ces comparaisons sont donnés à travers les graphiques de la figures 4.28.



(a) *Sem_G_Ct-Ict (α=0.2)* vs *Sem_G_Ct-Ict_Cos_Sim*



(b)*Sem_G_Tidf* vs *Sem_G_Tidf_Cos_Sim*

Figure 4.28: Apport de la mesure sémantique de l'évaluation des requêtes.

De ces graphiques, il ressort que :

- Le modèle *Sem_G_Ct-Ict_Cos_Sim* présente globalement de meilleurs résultats de recherche par rapport à *Sem_G_Ct-Ict_Cos_Classic* avec des gains de performances de 16.66% pour *P@5*, 9.09% pour *P@10*, 8.12% pour *P@15*, 8.33% pour *P*@20 et 6.49% pour la *MAP*. On ne note pas d'amélioration au point *P@30*, mais il n y pas de dégradation de performances. Une diminution non significative de la précision est observée au rang 100 (de l'ordre de -2%).
- Le modèle *Sem_G_Tidf_Cos_Sim* produit des résultats de recherche meilleurs comparativement à *Sem_G_Tidf_Cos_Classic* avec des gains de performances de 28.57% pour *P@5*, 6.57% pour *P@15*, 5.88% pour *P@20*, 6.38% pour *P*@100 et 20.54% pour la *MAP*. On ne note pas d'amélioration au point *P@10*, mais il n'y a pas

de dégradation de performances. Une diminution non significative de la précision est observée au point *P@30* (de l'ordre de -3.92%).

Par conséquent, nous concluons que notre score de pertinence sémantique apporte de meilleurs résultats de recherche comparativement à la mesure du cosinus du modèle vectoriel. Ceci peut être dû à la prise en compte, dans notre score de pertinence, des proximités sémantiques entre les concepts respectifs document-requête. Notons cependant que cette caractéristique est coûteuse et l'approche présente une complexité algorithmique très grande. Ainsi, il est difficile d'appliquer notre score de pertinence pour la recherche des documents dans des collections de tailles volumineuses.

Par ailleurs, nous avons comparé entre nos deux modèles basés sur l'appariement sémantique proposé avec les deux schémas de pondération sémantique *Ct-Ict* et *Tidf*, soient *Sem_G_Ct-Ict_Cos_Sim* (α=0.2) et *Sem_G_Tidf_Cos_Sim*. Les résultats de cette comparaison sont représentés à travers le graphique de la figure 4.29.

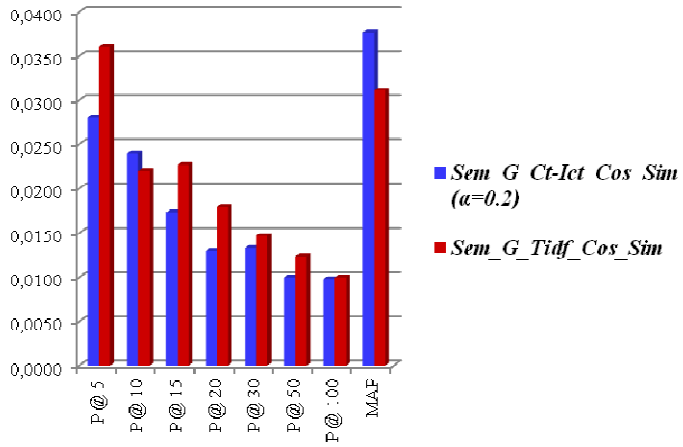


Figure 4.29: Résultats des comparaisons de notre modèle RI basé sur la pondération *Ct-Ict* d'une part et sur la pondération *Tidf* d'autre part.

La figure 4.29 montre que *Sem_G_Ct-Ict_Cos_Sim* (α=0.2) produit de meilleurs résultats de recherche aux points *P@x* (*x*= 1, 2, 10). Le taux d'amélioration observé aux deux premiers rangs (*P@1* et *P@2*) est significatif (de l'ordre 100%). Le gain de performances au point *P@10* est seulement de 9.09%. Cependant, le modèle *Sem_G_Tidf_Cos_Sim* est plus performant aux autres rangs avec des taux d'accroissement respectivement de 24.71% pour *P@3*, 60% pour *P@4*, 28.57% pour *P@5*, 31.21% pour *P*@15, 38.46% pour *P@20*, 10.52% pour *P@30*, 60% pour *P@50*, 28.57% pour *P@100* et 20.54% pour la *MAP*.

De cette évaluation, nous concluons que notre modèle de recherche *Sem_G_Ct-Ict_Cos_Sim* (α=0.2) est plus performant que notre modèle *Sem_G_Tidf_Cos_Sim*.

La figure suivante (Figure 4.30) résume les caractéristiques (résultats) de l'ensemble de toutes les approches évaluées. L'objectif étant de fournir une vue d'ensemble de toutes ces approches et de conclure à la plus performante.

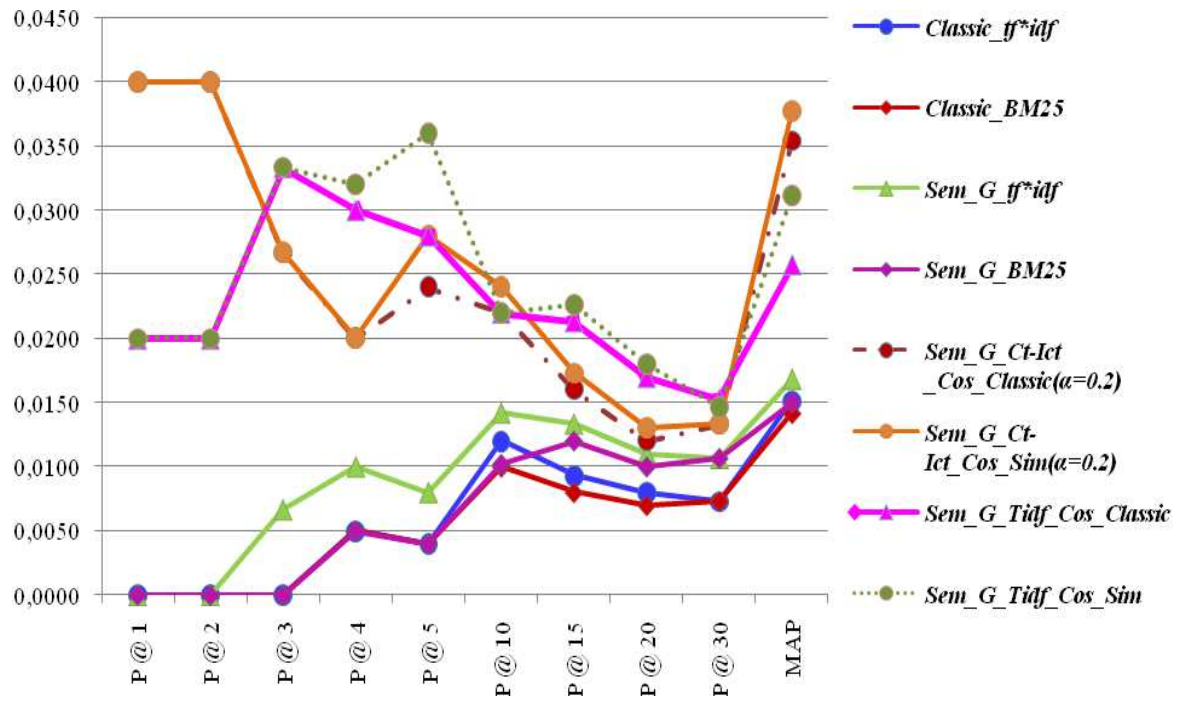


Figure 4.30 : Vue globale sur l'ensemble des résultats des approches évaluées.

D'après les résultats des approches évaluées sur la collection TIME, illustrés dans la figure 4.30, nous concluons que **l'indexation sémantique (*Sem_G_tf*idf* et *Sem_G_BM25*) par nos concepts, issus de l'approche de désambiguïsation globale proposée, présente une nette amélioration des résultats de la recherche par rapport à une indexation classique basée mots-clés (*Classic_tf*idf* et *Classic_BM25*). Néanmoins, les performances du SRI s'augmentent significativement en pondérant ces index sémantiques par l'un des schémas de pondération *Ct-Ict* (*Sem_G_Ct-Ict*) ou *Tidf* (*Sem_G_Tidf*), comparées à celles obtenues par leur pondération classique (*Sem_G_tf*idf* et *Sem_G_BM25*). En outre, en remplaçant dans le processus de recherche le score de pertinence classique (*Sem_G_Ct-Ict_Cos-Classic* et *Sem_G_Tidf_Cos-Classic*), basé sur la mesure de cosinus, par notre score de pertinence sémantique proposé en section 3.4 (*Sem_G_Ct-Ict_Cos-Sim* et *Sem_G_Tidf_Cos-Sim*), nous obtenons de meilleures performances du système. Par conséquent, on peut déduire que notre modèle de RI sémantique est plus performant qu'un modèle de RI classique.**

4.5 Evaluation avec la collection médicale Muchmore

4.5.1 Evaluation de l'approche d'indexation sémantique dans Muchmore

Les premières expérimentations sur la collection Muchmore consistent à tester l'impact de notre approche d'indexation sémantique sur l'efficacité de la recherche par rapport à une indexation classique basé mots-clés. Pour cela, nous avons comparé les résultats de recherche obtenus par les baselines (*Classic_tf*idf* et *Classic_BM25*) à ceux restitués par notre index sémantique, à base de mots orphelins et concepts (collocations et sens identifiés par la méthode de désambiguïsation globale), pondéré par *tf*idf* (*Sem_G_tf*idf*) puis par *Okapi-BM25* (*Sem_G_BM25*).

Les graphiques de la figure 4.31 montrent que nos index *Sem_G_tf*idf* et *Sem_G_BM25* produisent une nette dégradation des résultats de la recherche comparés respectivement à ceux des baselines *Classic_tf*idf* et *Classic_BM25*. Les taux de diminution des précisions du système de recherche varient entre -25% et –40%. Les résultats obtenus sont bien en deçà de ce qui était attendu. Les problèmes à l'origine de ces insuffisances sont causés par :

- l'imprécision de la désambiguïsation dans les requêtes Muchmore relativement à leurs petites tailles (en moyenne 3 termes utiles non vides par requête). Les concepts identifiés dans ces requêtes ne correspondent pas aux sens corrects des mots, provoquant ainsi une dégradation des résultats de la recherche. A contrario, la taille du contexte de la requête TIME est en moyenne de 9 termes (non vides) par requête. Ce qui a pour effet de désambiguïser correctement la plus part des termes des requêtes, en augmentant ainsi les performances de la recherche par rapport à une indexation classique basée mots-clés.

- l'utilisation de la ressource WordNet comme source d'évidence dans l'identification des termes de la Muchmore ne permet pas de détecter certaines collocations représentatives du domaine biomédical, qui sont plus riches sémantiquement que l'ensemble des mots qui les composent pris isolément. A titre d'exemple, *Catheter ablation*, *Posterior cruciate ligament*, *Ligamentum Collaterale Tibiale*, *patellar tendon*, *os trigonum…etc*. Ces collocations n'existent pas dans le langage de l'ontologie WordNet mais elles sont reconnues par d'autres ressources sémantiques médicales telles que le thesaurus MeSH ou le métathesaurus UMLS. De plus, la majorité des mots orphelins de Muchmore sont des termes médicaux ne contribuant pas à la désambiguïsation des autres mots ambigus de la collection.

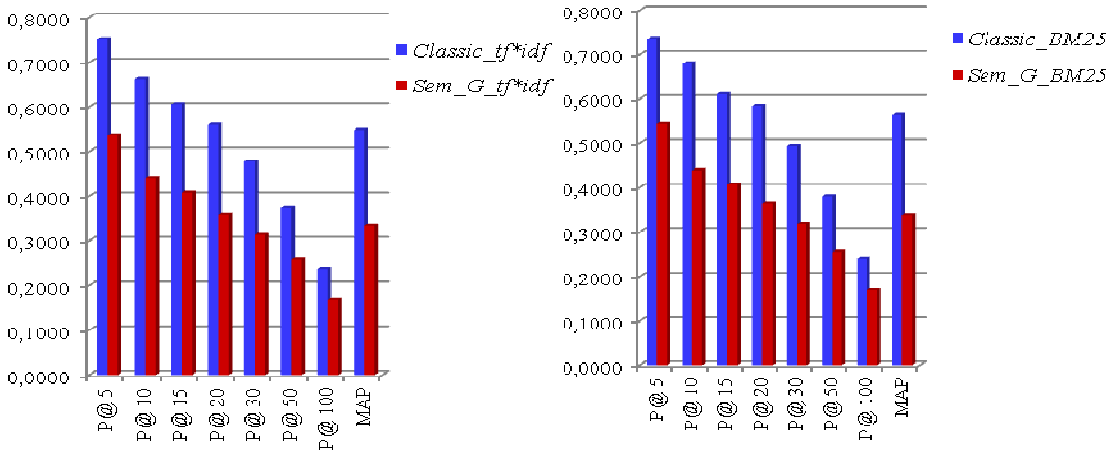


Figure 4.31: Impact de l'indexation sémantique de la collection Muchmore, basée sur les concepts de la désambiguïsation globale.

De ces résultats, il ressort que la qualité de notre approche d'indexation sémantique dépend du choix de la ressource externe utilisée dans l'identification des concepts des mots dans les documents (et les requêtes) d'une collection donnée. Nous supposons qu'une ressource sémantique qui couvre la totalité du domaine de la collection peut donner une meilleure précision de la désambiguïsation des mots permettant ainsi d'augmenter les performances de la recherche. De plus, la taille du contexte d'un mot joue un rôle important dans la définition de son sens.

4.5.2 Evaluation des approches de pondération des concepts dans Muchmore

Nous avons constaté, dans la sous-section précédente, que notre index sémantique issu de la désambiguïsation globale, a produit une diminution des performances de la recherche avec les schémas de pondération classique *tf*idf* (*Sem_G_tf*idf*) et *Okapi-BM25* (*Sem_G_BM25*) par rapport à une indexation classique basée mots-clés. Dans les expérimentations suivantes, nous tentons de tester si notre pondération sémantique (*Ct-Ict* et *Tidf*) de l'index *Sem_G* apporterait-elle de meilleurs résultats de recherche sur la collection Muchmore par rapport à une indexation classique? Pour cela, nous avons comparé aux résultats de la recherche obtenus par les baselines (*Classic_tf*idf* et *Classic_BM25*), les résultats issus de notre index sémantique pondéré respectivement par *Ct-Ict* (*Sem_G_Ct-Ict*) puis par *Tidf* (*Sem_G_Tidf*). L'objectif à travers cette évaluation est de mesurer la performance de notre pondération sémantique (*Ct-Ict* et *Tidf*) sur l'efficacité de la recherche dans la collection Muchmore.

4.5.2.1 Evaluation de l'approche de pondération *Ct-Ict*

Les résultats expérimentaux représentés dans le tableau 4.6 montrent que la valeur α=0.9 présente les meilleures performances de la pondération *Ct-Ict* à tous les points de précisions par rapport aux autres valeurs de α. Par conséquent, nous choisissons le schéma *Sem_G_Ct-Ict(α=0.9)* pour la pondération des concepts, de la collection Muchmore, issus de notre méthode de désambiguïsation globale.

	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**
P @ 1	0.5200	0.4800	0.4800	0.5200	0.5600	0.5600	0.5600	0.6000	**0.6000**
P @ 2	0.4200	0.4000	0.4400	0.4200	0.4400	0.4800	0.4600	0.5000	**0.5400**
P @ 3	0.4267	0.4000	0.4133	0.4267	0.4533	0.4533	0.4400	0.4667	**0.5067**
P @ 4	0.4300	0.4100	0.4100	0.4300	0.4400	0.4200	0.4100	0.4500	**0.4800**
P @ 5	0.4080	0.3760	0.3840	0.4000	0.4120	0.4320	0.4320	**0.4480**	**0.4800**
P @ 10	0.3560	0.3880	0.3440	0.3480	0.3280	0.3360	0.3440	0.3720	**0.4000**
P @ 15	0.3067	0.3013	0.2960	0.2932	0.2965	0.2933	0.3127	0.3260	**0.3627**
P @ 20	0.2940	0.2820	0.2800	0.2800	0.2780	0.2800	0.2780	0.2900	**0.3297**
P @ 30	0.2547	0.2480	0.2466	0.2427	0.2427	0.2413	0.2467	0.2587	**0.2907**
P @ 50	0.2032	0.2016	0.1960	0.1952	0.1952	0.2000	0.2064	0.2072	**0.2368**
P @ 100	0.1500	0.1484	0.1480	0.1472	0.1452	0.1444	0.1460	0.1476	**0.1600**
MP@x	0.3427	0.3305	0.3307	0.3366	0.3446	0.3491	0.3487	0.3696	**0.3988**
MAP	0.2411	0.2369	0.2400	0.2426	0.2425	0.2438	0.2445	0.2548	**0.2755**

Tableau 4.6 : Résultats d'évaluation de l'index sémantique *Sem_G_Ct-Ict* pour les différentes valeurs données à α.

Les comparaisons réalisées entre les résultats obtenus par l'index *Sem_G_Ct-Ict(α=0.9)* avec ceux des baselines sont données à travers la figure 4.32.

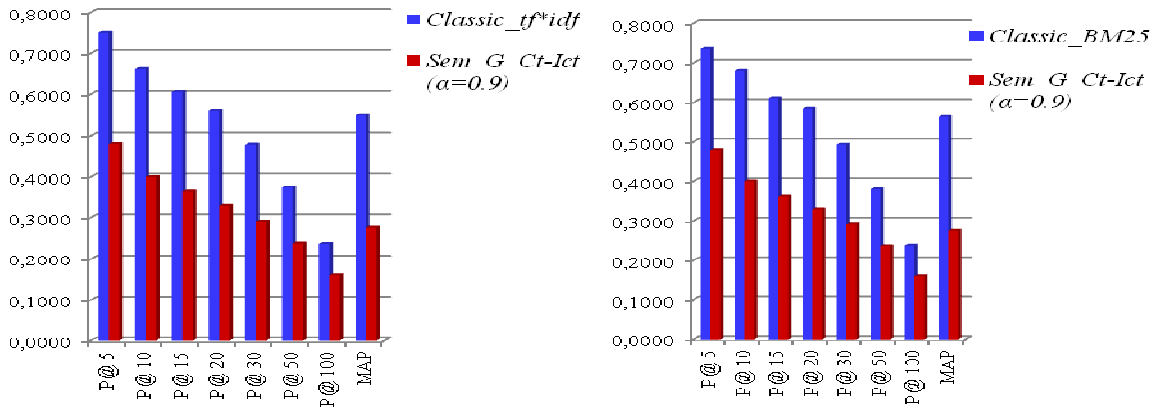


Figure 4.32: Apport de notre pondération sémantique des concepts, issus de la désambiguïsation globale, avec le schéma *Ct-Ict*.

De ces résultats, il apparait que l'index *Sem_G_Ct-Ict* présente une diminution significative (entre -30% et -40%) à tous les points de précisions comparé aux baselines *Classic_tf*idf* et *Classic_BM25*. Ces insuffisances de précisions sont dues probablement au problème de l'imprécision de la désambiguïsation globale des termes de la Muchmore, causée d'une part par les tailles des requêtes de la collection qui sont relativement petites et d'autre part par l'utilisation de la ressource du domaine général WordNet qui ne permet pas d'identifier correctement les concepts médicaux de la collection. De ce fait, **les concepts issus de la désambiguïsation des mots de la collection médicale Muchmore ne correspondent pas aux sens exacts de ces termes, provocant ainsi une dégradation dans les résultats de la recherche en utilisant notre pondération sémantique *Ct-Ict.***

4.5.2.2 Evaluation de l'approche de pondération *Tidf*

Les résultats des comparaisons réalisées entre *Sem_G_Tidf* d'une part et les index *Classic_tf*idf* et *Classic_BM25* d'autre part sont donnés à travers les graphiques de la figure suivante.

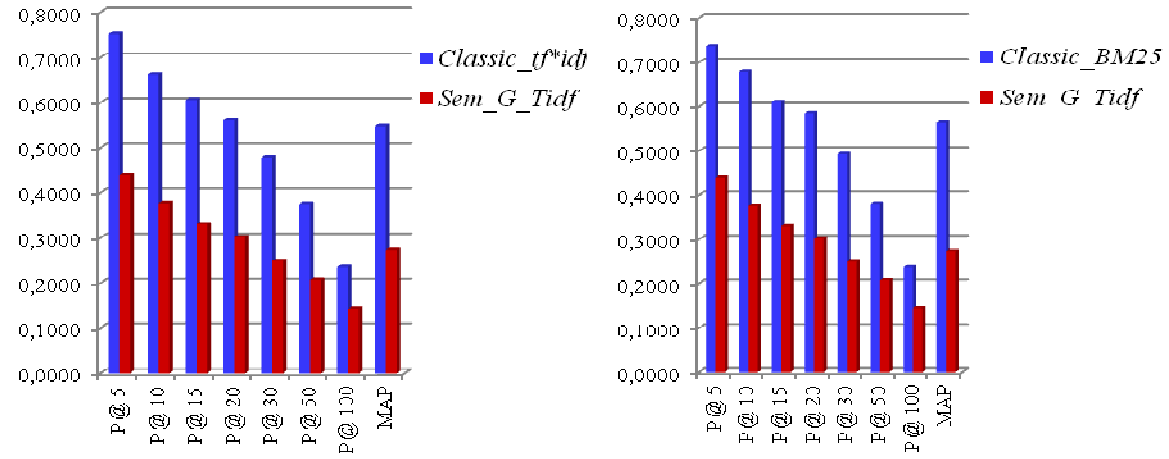


Figure 4.33: Apport de notre pondération sémantique des concepts, issus de la désambiguïsation globale, avec le schéma *Tidf*.

Des résultats de la figure 4.33, nous remarquons même avec notre schéma de pondération *Tidf*, l'index sémantique *Sem_G_Tidf* produit une dégradation significative des performances du système par rapport aux baselines. Le problème à l'origine est dû aux concepts des mots de la collection médicale Muchmore qui ont été mal désambiguïsés dans l'étape d'indexation. Ainsi, **le poids sémantique *Tidf* attribué à un concept d'un mot donné qui ne correspond pas au concept central exact du mot dans la collection, conduit à une dégradation des performances des résultats de la recherche.**

De ces évaluations, on peut conclure que les concepts (ou sens des mots) issus d'une désambiguïsation imprécise ne reflètent pas les concepts centraux exacts des mots dans les documents (ou les requêtes). Ainsi, leur pondération avec l'un de nos schémas de pondération sémantique *Ct-Ict* ou *Tidf*, fondée sur la centralité d'un concept dans le document (ou la requête), détériore les performances des résultats de la recherche par rapport à ceux d'une recherche classique basée sur les mots-clés.

4.6 Conclusion

Nous avons présenté au début de ce chapitre, l'environnement technique appliqué dans l'implémentation de notre modèle de RI sémantique. Par la suite, nous avons exposé les résultats expérimentaux obtenus des différents tests réalisés pour l'évaluation des performances de ce modèle dans le cadre de la recherche d'information en, s'appuyant sur les deux collections de tests : la collection TIME et la collection Muchmore.

Au vu des résultats de recherche dans la collection TIME, il apparait clairement que l'indexation sémantique proposée a permis d'améliorer les performances du système de recherche par rapport à une indexation traditionnelle, aussi bien avec les concepts utilisés seuls que lorsqu'ils sont combinés aux mots-clés des index classiques. Les concepts associés aux collocations de mots dans un texte donné ont servi à réduire l'ambiguïté de son contenu, permettant ainsi d'augmenter la performance du système. Les concepts des mots simples ont été identifiés par une technique de désambiguïsation contextuelle locale, ou globale ou mixte. Ces approches de désambiguïsation ont été toutes évaluées dans le but de définir parmi elles laquelle est la plus performante à retrouver le sens attendu du terme dans son contexte d'utilisation. Les résultats de ces évaluations ont montré que la désambiguïsation globale apporte plus de précisions qu'une désambiguïsation dans le contexte locale, confortant ainsi l'hypothèse sous-jacente à la définition du contexte global, qui stipule le «*One sense per discourse*».

Par ailleurs, nous avons constaté, à travers les expérimentations menées sur la collection TIME, que l'exploitation des domaines des mots dans le processus de désambiguïsation est une voie intéressante pour retrouver les sens (concepts) des mots qui les correspondent dans leur contexte d'utilisation. Les résultats obtenus, en intégrant dans le processus de désambiguïsation notre approche de désambiguïsation des domaines, sont satisfaisants et très encourageants. On peut déduire alors que la connaissance du domaine d'usage d'un mot renforce sa désambiguïsation sémantique dans son contexte d'apparition.

Néanmoins, avec une ressource linguistique ne couvrant pas la totalité du domaine d'application de l'information recherchée, comme dans le cas du WordNet qui ne reconnait pas certains termes médicaux de la collection Muchmore, notre désambiguïsation ne pourrait pas identifier correctement les concepts adéquats des mots ambigus dans les documents ou les requêtes, provoquant ainsi une dégradation des résultats du processus de recherche. De plus,

si la taille d'un contexte donné ne contient pas un nombre important de mots, la méthode de désambiguïsation serait incapable de retrouver leur sens exacts dans ce contexte.

Nous avons aussi testé dans nos expérimentations, l'apport de l'approche de pondération proposée avec les schémas : *Ct-Ict* ou *Tidf*, basés sur la centralité d'un concept dans le document. Les résultats obtenus prouvent que notre pondération sémantique est plus efficace qu'une pondération classique basée sur *tf*idf* ou *Okapi-BM25*. En outre, la pondération *Ct-Ict* est plus performante que la pondération *Tidf*, en particulier dans les précisions aux premiers rangs. Cependant, la pondération sémantique des concepts issus d'une mauvaise désambiguïsation, avec l'un de ces schémas proposés, dégradent nettement les performances du système.

Finalement, notre proposition de l'évaluation des requêtes par une nouvelle mesure de pertinence sémantique produit de meilleures performances du système par rapport à une évaluation classique fondée sur la mesure du cosinus [Salton et al., 83].

On peut ainsi conclure que le modèle de RI sémantique, proposé dans le cadre de ce travail, apporte plus d'intérêt qu'un modèle classique basé sur les mots-clés.

Conclusion et perspectives

Synthèse

Les travaux présentés dans ce mémoire s'inscrivent dans le contexte de la recherche d'information (RI) sémantique, en se basant sur deux volets différents mais néanmoins complémentaires dans un SRI. Le premier volet concerne une amélioration de la représentation de l'information dans les systèmes de recherche, en utilisant la sémantique des mots dans l'indexation des documents et requêtes. Le second volet concerne une évaluation sémantique des requêtes pour retrouver des documents sémantiquement pertinents aux requêtes utilisateurs. Dans ce cadre, nous avons défini un nouveau modèle de RI sémantique dans la recherche de documents textuels. Nos contributions ont porté sur un double aspect : d'une part, une proposition d'une approche d'indexation sémantique des documents et requêtes, et d'autre part, une proposition d'une approche d'évaluation sémantique des requêtes.

Notre première contribution a pour but d'apporter une meilleure représentation du contenu sémantique des documents et requêtes. Classiquement, les systèmes de recherches d'information indexent les documents, ou respectivement les requêtes, par les mots-clés qu'ils contiennent. La pertinence document-requête dans de tels systèmes est calculée par l'appariement de leur ressemblance en se basant sur le nombre de mots communs partagés. Or, les mots de langue sont par nature ambigus, réduisant ainsi la précision des résultats de la recherche et les rend incomplets. Par conséquent, des documents pourtant non pertinents seront retrouvés et des documents pourtant sémantiquement pertinents seront ignorés. Pour résoudre cette problématique, la recherche d'information sémantique se base sur les sens des mots, ou concepts, dans la représentation des documents et requêtes plutôt que par de simples mots-clés. C'est l'objectif de notre approche d'indexation sémantique proposée dans ce contexte. Elle s'articule par une succession de trois nouvelles approches :

(1) La première est une approche d'extraction des termes d'indexation par projection du texte, d'une part sur une liste préétablie de toutes les collocations de la ressource linguistique WordNet pour faire ressortir les expressions correspondant à des collocations de mots, et d'autre part sur l'ontologie WordNet pour retrouver les mots simples non vides appartenant au texte. Deux types de mots simples ont été définis : des mots simples ayant une entrée dans WordNet et des mots dits orphelins ne possédant aucune entrée dans WordNet.

(2) La seconde est une approche de désambiguïsation qui utilise conjointement WordNet et son extension aux domaines WordNetDomains comme sources d'évidence pour déterminer le bon sens de chaque mot ambigu dans son contexte. Nous avons suggéré dans cette approche de désambiguïser seuls les mots simples ayant une entrée dans WordNet puisque les collocations sont monosémiques. Pour vérifier laquelle des hypothèses suivantes est la plus vraisemblable : *one sense per discourse*, stipulant qu'un mot est généralement utilisé avec un seul sens dans un document, et *multiple senses per discourse*, stipulant qu'un mot peut être utilisé avec différents sens dans un même document, nous avons considéré deux types de

contextes d'utilisation d'un mot : le contexte local et le contexte global. A ces deux contextes, trois techniques de désambiguïsation ont été proposées :

- une désambiguïsation contextuelle locale,
- une désambiguïsation contextuelle globale,
- une désambiguïsation contextuelle mixte (désambiguïsation dans le contexte local puis dans le contexte global).

Ces techniques de désambiguïsation reposent sur les relations sémantiques entre les concepts des mots appartenant à leurs domaines d'usage respectifs dans le contexte considéré. La nouveauté dans ces approches, par rapport aux approches de désambiguïsation existantes, est la proposition de deux scores de désambiguïsation :

- le premier est un score de désambiguïsation des domaines d'un mot pour définir celui qui le correspond dans son contexte, en se basant sur une mesure de proximités sémantiques entre les domaines des sens des mots dans WordNetDomains,
- le second est un score de désambiguïsation sémantique des mots dans leurs domaines d'usage associés, en se basant sur les liens sémantiques entre les sens des mots appartenant au même domaine ou des domaines similaires.

Notre objectif à travers ces scores est de renforcer la désambiguïsation d'un mot afin d'obtenir le sens attendu dans son contexte.

(3) La troisième est une approche de pondération quantifiant le degré de représentativité d'un terme de l'index sémantique dans le document. Nous avons proposé une combinaison de deux pondérations :

- une pondération classique des termes orphelins par un schéma classique (par exemple : *tf*idf* ou *Okapi-BM25*),
- et une pondération sémantique des concepts (collocations et sens des mots simples) basée sur leur centralité. Nous avons défini la centralité d'un concept par son importance apparente, exprimée par sa fréquence relative, et son importance latente, exprimée par ses relations sémantiques avec les autres concepts dans le document. En particulier, deux schémas de pondération basés sur la centralité ont été proposés : *Ct-Ict* et *Tidf*.

Pour résoudre le problème de la synonymie dans le processus d'indexation, les concepts (sens des mots simples et collocations) sont représentés, dans cette approche, par les numéros de leurs synsets associés dans WordNet. Les index sémantiques ainsi obtenus sont composés de mots orphelins et concepts (numéros synsets dans WordNet) pondérés.

Notre deuxième contribution a pour but d'améliorer l'évaluation des requêtes pour retrouver des documents pertinents. Dans les modèles de recherche les plus intuitifs en RI, la pertinence document-requête est mesurée en fonction des termes (concepts éventuellement combinés aux mots clés) identiques de la requête et du document. Cependant, nous avons constaté qu'un document pourtant pertinent peut ne pas contenir aucun terme de la requête, même s'il possède des liens

sémantiques implicites entre ses concepts et les concepts de la requête. Pour découvrir ces liens et les intégrer dans la mesure de pertinence d'un document, nous avons proposé une approche d'évaluation sémantique des requêtes basée sur l'appariement de vecteurs. En particulier, les index sémantiques sont interprétés par des vecteurs de concepts (numéros synsets dans WordNet) combinés aux mots orphelins pondérés, tandis que la pertinence document-requête est calculée sur la base d'une nouvelle mesure sémantique traduisant sa pertinence apparente, exprimée par les poids de leurs termes communs, et sa pertinence sémantique latente, exprimée par les similarités sémantiques entre leurs concepts respectifs.

L'ensemble de nos contributions défini dans sa globalité un nouveau modèle de RI sémantique. Pour évaluer ses performances, nous avons mené plusieurs expérimentations sur deux collections de test : la collection TIME et la collection Muchmore.

L'analyse globale des résultats d'évaluation obtenus dans la collection TIME, ont montré l'intérêt de notre modèle de RI proposé au niveau de l'approche d'indexation sémantique et au niveau de l'évaluation sémantique des requêtes.

Au niveau de l'approche d'indexation proposée, les différents tests réalisés ont révélé que les concepts seuls ou combinés aux mots-clés des index classiques, apportent plus de précision dans la description du contenu informatif des documents et requêtes par rapport à leur indexation classique basée uniquement sur les mots-clés. Par conséquent, on peut déduire que les techniques de désambiguïsation contextuelle ont permis d'identifier les sens attendus de certains mots dans leurs contextes, particulièrement la méthode de désambiguïsation globale qui a produit de meilleures performances du SRI par rapport aux autres approches de désambiguïsation proposées. Ceci nous conforte alors à penser que le contexte global présente le meilleur choix pour définir les sens des mots employés dans le document (ou respectivement dans la requête). Par ailleurs, l'exploitation dans notre processus de désambiguïsation les concepts associés aux collocations d'une part et les domaines d'usages des termes dans leurs contextes d'une autre part, a permis de réduire le problème d'ambiguïté dans l'indexation et de renforcer la désambiguïsation de ces termes, apportant ainsi une amélioration significative des précisions des résultats de la recherche.

Toutefois, nous avons aussi constaté à travers nos expérimentations, que l'indexation sémantique par les concepts-noms est un meilleur choix que l'indexation classique par les mots clés ou que l'indexation sémantique incluant des concepts de catégories syntaxiques quelconques. Cela est dû à la mesure de similarité (en l'occurrence la mesure de Resnik [Resnik, 99]) utilisée dans le score de désambiguïsation sémantique des mots, qui est fondée sur les relations sémantiques de la taxonomie *is-a* des verbes et noms de WordNet. De ce fait, les adverbes et les adjectifs ont été mal désambiguïsés provocant ainsi une diminution de la précision de notre approche d'indexation sémantique avec les concepts de différents formes syntaxiques. Néanmoins, cette approche reste toujours meilleure qu'une indexation classique basée sur les mots-clés.

Par ailleurs, les résultats d'évaluation de notre approche de pondération des concepts, issus de n'importe quelle technique de désambiguïsation, ont montré son efficacité, tant avec le schéma *Ct-Ict* qu'avec le schéma *Tidf*, par rapport à une pondération classique par *tf*idf* ou *Okapi-BM25*. En particulier, la pondération *Ct-Ict* est plus performante que la pondération *Tidf*. Le facteur α utilisé dans la formule de pondération *Ct-Ict*, pour balancer entre l'importance apparente du concept et son

importance latente, est choisi expérimentalement dans la collection TIME par : la valeur 0.1 pour les concepts issus de la désambiguïsation locale et la désambiguïsation mixte, et la valeur 0.2 pour les concepts issus de la désambiguïsation globale. On peut ainsi remarquer que le choix des valeurs de ce paramètre dans TIME, favorise dans la formule de pondération *Ct-Ict* l'importance sémantique du concept dans le document, exprimée par ses similarités aux autres concepts, par rapport à sa fréquence relative. Cependant, il reste à déterminer dans quelle mesure ce choix de la valeur de ce paramètre est dépendant ou non de la collection utilisée et du contexte de désambiguïsation.

Au niveau de l'évaluation sémantique des requêtes, la mesure de pertinence proposée, qui intègre dans l'évaluation de le requête les liens sémantiques implicites entre les vecteurs de concepts respectifs d'un document et d'une requête, a présenté des résultats globalement meilleurs que la mesure classique du cosinus [Salton et al., 83].

Au vu de l'analyse des résultats d'évaluation obtenus dans la collection TIME, on peut conclure que notre modèle de RI sémantique est plus performant qu'un modèle de RI classique. Cependant, nous avons remarqué à travers les résultats issus de la collection Muchmore, que la taille du contexte d'un mot est importante pour sa désambiguïsation. Il serait difficile de retrouver le sens exact d'un mot ambigu dans un contexte qui ne contient pas un nombre important de termes. De plus, avec une ressource linguistique ne couvrant pas la totalité du domaine d'application de l'information recherchée, comme le cas de WordNet qui ne reconnait pas certains concepts médicaux, notre approche de désambiguïsation serait imprécise provocant ainsi une dégradation des performances de notre modèle de RI, même en appliquant notre pondération sémantique dans la recherche des documents. On peut alors conclure que la qualité de notre modèle de RI dépend du choix de la ressource linguistique utilisée dans l'identification des concepts des mots.

Perspectives

Les perspectives de notre modèle se déclinent en deux objectifs principaux:

(1) le premier consiste à tester le modèle de RI sémantique proposé sur plusieurs collections volumineuses, telles que les collections TREC, afin de valider nos propositions à grande échelle et de fixer la valeur du paramètre α de la formule de pondération sémantique *Ct-Ict*, qui permet de balancer entre l'importance sémantique du concept dans le document et sa fréquence relative ;

(2) le second consiste à apporter des améliorations futures au modèle par :

- une amélioration dans l'approche de détection des termes d'indexation, en utilisant une combinaison de plusieurs ressources linguistiques dans l'extraction des termes. Notre but est de couvrir le maximum de mots composés (collocations) ou simples appartenant aux différents domaines d'application dans les collections de test ;
- une amélioration dans l'approche de désambiguïsation des mots, par une proposition d'une nouvelle mesure de similarité entre les concepts, en se basant sur les différentes relations sémantiques dans les ressources linguistiques, pour identifier avec plus de précision les concepts des mots de catégories syntaxiques quelconques.

Références Bibliographiques

[Alvarez et al., 04] C. Alvarez, P. Langlais, J-Y. Nie. *Word pairs in language modeling for information retrieval*. Proceeding of the conference on computer assisted information retrieval, 2004

[Amati et al., 02] G. Amati, C .J. Van Riejsbergen. *Probabilistic models of information retrieval based on measuring the divergence from randomness*. ACM Transactions on Information System, 20(4), p. 357–389, 2002.

[Apidianaki, 08] M. Apidianaki. *Acquisition automatique de sens pour la désambiguïsation et la sélection lexicale en traduction*. Thèse de doctorat en linguistique théorique et descriptive et automatique, université Diderot, Paris 7, 2008.

[Aronson, 01] A.R. Aronson. *Effective mapping of biomedical text to the UMLS Metathesaurus : the MetaMap program*. Proceedings AMIA Symposium, p.17–21, 2001.

[Baeza-Yates et al., 99] R. A. Baeza-Yates, B. A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press. *Addison-Wesley*, 1999.

[Banerjee et Pedersen, 03] S. Banerjee, T. Pedersen. *Extended gloss overlaps as a measure of semantic relatedness*. Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI, Acapulco, Mexico), p. 805–810, 2003.

[Baziz et al., 04] M. Baziz, M. Boughanem, N. Aussenac-Gilles. *The Use of Ontology for Semantic Representation of Documents*. The 2nd Semantic Web and Information Retrieval Workshop(SWIR), SIGIR 2004, Sheffield UK, Ying Ding, Keith Van Riejsbergen, Iad Ounis, Joemon Jose (Eds.), p. 38-45, Juillet 2004.

[Baziz et al., 05a] M. Baziz, M. Boughanem and N. Aussenac-Gilles. *A Conceptual Indexing Approach based on Document Content Representation*. CoLIS5: Fifth International Conference on Conceptions of Libraries and Information Science, Glasgow, UK, 4 juin 8 juin 2005.

[Baziz, 05b] Baziz M. Indexation *Conceptuelle Guidée Par Ontologie Pour La Recherche d'Information.* Thèse de Doctorat en Informatique de l'Université Paul Sabatier de Toulouse (Sciences). Décembre 2005.

[Belkin et al., 92] J. N. Belkin, P. Ingwersen, A. M. Pejtersen. *Proceedings of the 15th Annual International ACM SIGIR, In Conference on Research and Development in Information Retrieval*, Copenhagen, Denmark, June 21-24, ACM 1992.

[Borlund et al., 98] P. Borlund, P. Ingwersen. *Measures of relative relevance and ranked half-life, performance indicators for interactive IR*. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR'98, p. 324–331, New York, NY, USA, ACM, 1998.

[Boubekeur, 08] F. Boubekeur. *Contribution à la définition de modèles de recherche d'information flexibles basés sur les CP-Nets*. Thèse de doctorat en informatique, Université Toulouse III - Paul Sabatier, 2008.

[Boubekeur et al., 08] F. Boubekeur, M. Boughanem, L. Tamine. *Une approche d'indexation conceptuelle de documents basée sur les graphes CP_Nets*. Dans: cinquième édition du colloque sur l'optimisation et les systèmes d'information COSI'08, Tizi-Ouzou, Algérie, 8-10 juin 2008.

[Boubekeur et al., 10a] F. Boubekeur, M. Boughanem, L. Tamine , M. Daoud. *Using WordNet for Concept-based document indexing in information retrieval*. Fourth International Conference on Semantic Processing (SEMAPRO), Florence, Italy, October 2010.

[Boubekeur el al., 10b] F. Boubekeur, M. Boughanem, L.Tamine, M. Daoud. *De l'utilisation de WordNet pour l'indexation conceptuelle des documents*. le 13 ème Colloque International sur le Document Electronique (CIDE 13), 16-17 Décembre 2010, INHA, Paris.

[Boughanem et al., 98] M. Boughanem, T. Dkaki, J. Mothe, C. Soulé-Dupuy. *Mercure at TREC7*. TREC 1998. p.355-360, 1998.

[Boughanem et al. 99] M. Boughanem, C. Chrisment, C. Soule-Dupuy. *Query modification based on relevance backpropagation in adhoc environment*. Information Processing and Management, 35, p. 121-139, 1999.

[Boughanem et al., 04] M. Boughanem, W. Kraaij, J.Y. Nie. *Modèles de langue pour la recherche d'informations*. Dans les systèmes de recherche d'informations - Modèles conceptuels, ed. M. Ihadjadene, Hermes, p. 163-184, 2004.

[Boughanem et al., 08] M. Boughanem, J. Savoy, editors. *Recherche d'information états des lieux et perspectives*. Hermès Science Publications, 2008.

[Boughanem et al., 10] M. Boughanem, I. Mallak, H. Prade. *A new factor for computing the relevance of a document to a query*. In IEEE World Congress on Computational Intelligence (WCCI 2010), Barcelone, 18/07/2010-23/07/2010.

[Buckley et al., 94] C. Buckley, G. Salton, J. Allan, A. Singhal. *Automatic query expansion using SMART, TREC 3.* In Text REtrieval Conference, 1994.

[Buckley et al., 95] C. Buckley, A. Singhal, M. Mitra. *New Retrieval Approaches Using SMART , TREC 4*. TREC 1995.

[Budanitsky et al., 06] A. Budanitsky, G. Hirst. *Evaluating WordNet-based Measures of Lexical Semantic Relatedness*. Association for Computational Linguistics, 2006.

[Buitelaar et al., 07] P. Buitelaar, B. Magnini, C. Strapparava, P. Vossen. *Domain specific word sense disambiguation, chapter 10*. In Word sense disambiguation : algorithms and applications, p.275–298.

[Burnard, 00] L. Burnard. *Reference Guide for the British National Corpus (World Edition*). Oxford University Computing Services, 2000.

[Chevallet., 09] J.P Chevallet. *Ressources endogènes et exogènes pour une indexation conceptuelle intermédia*. Mémoire d'Habilitation à Diriger des Recherches, 2009.

[Claypool et al., 01] M. Claypool, P. Le, M. Wased, D. Brown. *Implicit interest indicators*, p. 33-40, 2001.

[Cleverdon, 67] C. Cleverdon. *The cranfield tests on index language devices*. In Aslib Proceedings, Vol. 19, p. 173-193, 1967.

[Cleverdon, 70] C. Cleverdon. *Progress in documentation: Evaluation of information retrieval systems*. In Journal of Documentation 26, p. 55-67, 1970.

[Clinchant et al., 10] S. Clinchant, E. Gaussier. *Information-based models for ad-hoc IR*. In Proc. of Conference on Research and Development in Information Retrieval, SIGIR'10, p. 234-241. ACM, 2010.

[Coletti et al., 01] M. Coletti, H. Bleich. *Medical subject headings used to search the biomedical literature*. Journal of the American Medical Informatics Association, vol. 8, p. 317–323, 2001.

[Cowie et al., 92] J. Cowie, J. Guthrie, L. Guthrie. *Lexical Disambiguation using Simulated Annealing.* In Proceedings of the 14th International Conference on Computational Linguistics (COLING'92), Nantes, France, p. 359-365, 1992.

[Croft et Harper, 88] W. B Croft, D. J. Harper. *Document retrieval systems*. Chapter, Using probabilistic models of document retrieval without relevance information, p. 161–171, Taylor Graham Publishing, London, UK, UK. 1988.

[Croft et al., 09] W. B. Croft, D. Metzler, T. Strohman. *Search Engines - Information Retrieval in Practice*. Pearson Education, 2009.

[Deerwester et al., 90] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, R. A. Harshman. *Indexing by Latent Semantic Analysis*. In Journal of the American Society of Information Science, Vol. 41, 6, p. 391-407, 1990.

[Dinh et al., 10] D. Dinh, L. Tamine. *Vers un modèle d'indexation sémantique adapté aux dossiers médicaux de patients*. In Conférence francophone en Recherche d'Information et Applications, (CORIA 2010), p. 325-336, 2010.

[Dinh, 12] D. Dinh. *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 Phd, Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier), Septembre 2012.

[Egozi et al., 11] O. Egozi, S. Markovitch, E. Gabrilovich. *Concept-Based Information Retrieval using Explicit Semantic Analysis*. ACM Transactions on Information Systems, Vol. 29 Issue 2, April 2011.

[Ehrig et al., 04] M. Ehrig, P. Haase, M. Hefke, N. Stojanovic. *Similarity for ontology-a comprehensive framework*. In Workshop Enterprise Modelling and Ontology: Ingredients for Interoperability, 2004.

[Fellbaum, 98] C. Fellbaum. *Wordnet – An Electronic Lexical Database*. The MIT Press, Cambridge, Massachusetts, 1998.

[Gale, 91] W. Gale, K. Church, D. Yarowsky. *One sense per discourse*. Speech and Natural Language Workshop, p.233–23, 1991.

[Gaussier et al., 97] E. Gaussier, G. Grefenstette, M. Schulze. *Traitement du langage naturel et recherche d'informations, quelques expériences sur le français*. In Premières Journées Scientifiques et Techniques du Réseau Francophone de l'Ingénierie de la Langue de l'AUPELF-UREF, 1997.

[Gaussier et al., 00] E. Gaussier, G. Grefenstette, D. Hull, C. Roux. *Recherche d'information en français et traitement automatique des langues*. Revue Traitement Automatique des Langues (TAL), 41(2), p.473–494, 2000.

[Gliozzo et al., 04] A. Gliozzo, B. Magnini, C. Strapparava. *Unsupervised domain relevance estimation for word sense disambiguation.* In Proceedings of the 2004 Conference on Empirical Methods in

Natural Language Processing *(*EMNLP, Barcelona, Spain), p.380–387, 2004.

[Gonzalo et al., 99] J. Gonzalo, A. Pefias, F. Verdejo. *Lexical ambiguity and information retrieval revisited*. In Proceedings of EMNLP/VLC, 1999.

[Grolier] Grolier Multimedia Encyclopedia CD-ROM. *Grolier interactive* Inc., 90 Sherman Turnpike, Danbury, CT 06816, USA.

[Guthrie et al., 91] J.A. Guthrie, L. Guthrie, Y. Wilks, H. Aidinejad. *Subject-dependant cooccurrence and word sense disambiguation.* In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkley, p. 146-152, 1991.

[Hammache et al., 09] A. Hammache, M. Boughanem, R. Ahmed-Ouamer. *Introduction de la sémantique d'un document sous le modèle de langage*. Dans la sixième édition de la Conférence francophone en Recherche d'Information et Applications (CORIA 2009), 5-7 mai 2013.

[Hammache et al., 13] A. Hammache, M. Boughanem et R. Ahmed-Ouamer. *Pseudo-réinjection de pertinence basée sur un modèle de langue mixte combinant les termes simples et composés.* Dans la dixième édition de la Conférence francophone en Recherche d'Information et Applications (CORIA 2013). Neuchâtel, Suisse, 2013.

[Harman, 92] D. Harman. *Relevance Feedback Revisited.* In the Proceedings of the ACM SIGIR Conference On Research and Development in Information Retrieval (SIGIR*)*, p.1-10, 1992.

[Harrathi, 10] R. Harrathi. *Recherche d'information conceptuelle dans les documents semi-structurés*. Thèse de Doctorat de L'Institut Nationale des Sciences Appliquées de Lyon. Septembre 2010.

[Harrathi et al., 10] F. Harrathi, C. Roussey, L. Maisonnasse, S. Calabretto. *Vers une approche statistique pour l'indexation sémantique des documents multilingues*, Dans : Actes du XXVIII° congrès INFORSID, Marseille, mai 2010.

[Hersh et al., 92] W.R Hersh, D.H Hickam, T.J. Leone. Words, *concepts or Both: Optimal Indexing Units for Automated Information Retrieval*. Proc 16th Annu Symp Comput Appl Med Care, p.644–819, 1992.

[Ingwersen, 92] P. Ingwersen. *Information retrieval interaction*. London, Taylor Graham, 1992.

[Jacquemin et al., 02] C. Jacquemin, B. Daille, J. Royanté, X. Polanco. *In vitro evaluation of a program for machine-aided indexing*. Inf. Process. Manage. 38, 6 (Nov. 2002), 765-792, 2002.

[Jiang-Conrath, 97] J. Jiang, D. Conrath. *Semantic similarity based on corpus statistics and lexical taxonomy*. In Proceedings of International Conference on Research inComputational Linguistics, Taiwan, 1997.

[Jiang et al., 04] M. Jiang, E. Jensen, S. Beitzel, *Effective use of phrases in language modeling to improve information retrieval*. Symposium on AI & math, Special session on intelligent text processing, Florida, January 2004.

[Kelly et al., 75] E. F. Kelly, P. J. Stone. *Computer recognition of english word senses*. North-Holland Publishing. North-Holland, Amsterdam, 1975.

[Kelly et al., 01] D. Kelly, N.J. Belkin. *Reading time, scrolling and interaction, exploring implicit sources of user preferences for relevance feedback*. In Proceedings of the 24th annual international

ACM SIGIR conference on Research and development in information retrieval, SIGIR '01, p. 408–409, New York, NY, USA. ACM, 2001.

[Khan, 00] L. R. Khan.*Ontology-based Information Selection. Phd Thesis, Faculty of the Graduate School, University of Southern California. August 2000.*

[Khan et al., 04] L. Khan, D. Mc Leod, E. Hovy. *Retrieval effectiveness of an ontology-based model for information selection*. The VLDB Journal, édition13, p.71–85, 2004.

[Knight et al., 94] K. Knight, S. Luk. *Building a large-scale knowledge base for machine translation*. In Proceedings of AAAI'94, 1994.

[Kolte et al., 08] S.G. Kolte, S. G. Bhirud. *Word Sense Disambiguation using WordNetDomains*. In First International Conference on Emerging Trends in Engineering and Technology. IEEE DOI 10.1109/ICETET, 2008.

[Kolte et al., 09] S. G. Kolte, S. G. Bhirud. *WordNet : A Knowledge Source for Word Sense Disambiguation*. International Journal of Recent Trends in Engineering, Vol 2, No. 4, November 2009.

[Krovetz et al., 92] R. Krovetz, W. B. Croft. *Lexical Ambiguity and Information Retrieval*. ACM Transactions on Information Systems, Vol. 10, No 2, pp. 115_141. April 1992.

[Krovetz, 97] R. Krovetz. *Homonymy and polysemy in information retrieval*. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (A CL-97), p. 72-79, 1997.

[Krovetz, 98] R. Krovetz. *More than one sense per discourse*. Association for Computational Linguistics Special Interest Group on the Lexicon (ACL-SIGLEX-1998), 1998.

[Kucera et al., 67] H. Kucera, W.N. Francis. *Computational Analysis of Present-Day American English*. Brown University Press, Providence, RI, 1967.

[Leacock et al., 98] C. Leacock, G.A. Miller, M. Chodorow. *Using corpus statistics and WordNet relations for sense identification*. Comput. Linguist, p..147- 165, 1998.

[Lesk, 86] M.E. Lesk. *Automatic sense disambiguation using machine readable dictionaries : How to tell a pine cone from a nice cream cone. In Proceedings of the SIGDOC Conference*, Toronto, 1986.

[Lin, 98] D. Lin. *An information-theoretic definition of similarity*. In Proceedings of 15th International Conference On Machine Learning, 1998.

[Luhn, 58] H. Luhn. *The automatic creation of literature abstracts*. IBM Journal of Research and Development 24, p.159–165, 1958.

[Magnini et al., 00] B. Magnini, G. Cavagli. *Integrating subject field codes into WordNet*. In Proceedings of LREC-2000, Second International Conference on Language Resources a*nd Evaluation*, Athens, Greece, p. 1413- 1418, 2000.

[Magnini et al., 02] B. Magnini, G. Pezzulo, A. Gliozzo. *The role of domain information in Word Sense Disambiguation. Journal : Naturel Language Engineering, Vol*. 8, p. 359-373, 2002.

[Maisonnasse et al., 09] L. Maisonnasse, E. Gaussier, J-P. Chevallet. *Model Fusion in Conceptual Language Modeling*. ECIR 2009, p.*240*-251, 2009.

[Mallak, 11] I. Mallak. *De nouveaux facteurs pour l'exploitation de la sémantique d'un texte en Recherche d'Information*. Thèse Phd, université Toulouse, 2011.

[Maron et al., 60] M. Maron, J. Kuhns. *On relevance, probabilistic indexing and information retrieval*. Journal of the Association for Computing Machinery 7, p. 216–244. 1960.

[Mihalcea et al, 00] R. Mihalcea, D. Moldovan. *Semantic indexing using WordNet senses*. In Proceedings of ACL Workshop on IR & NLP, Hong Kong, October 2000. http://www.seas.smu.edu/~rada/papers/acl00.nlp_ir.ps.gz

[Miller et al., 93] G. Miller, C. Leacock, R. Tengi, R.T. Bunnker. *A semantic concordance*. In Proceedings of the ARPA Workshop on Human Language Technology, p.303-308, 1993.

[Miller, 95] G. Miller. *WordNet : A Lexical database for English.* Actes de ACM 38, p. 39-41, 1995.

[Mizoguchi, 04] R. Mizoguchi. *Le rôle de l'ingénierie ontologique dans le domaine des EIAH*. Sciences et Technologies de l'Information et de la Communication pour l'Education et la Formation, Vol. 11, 2004.

[Mizzaro, 97] S. Mizzaro. *Relevance, the whole (hi) story*. Journal of the American society for information science, 48(9), p 810–832, 1997.

[Mohammad et al., 06] S. Mohammad, G.Hirst. *Determining word sense dominance using a thesaurus*. In Proceedings of the 11th Conference on European chapter of the Association for Computational Linguistics (EACL, Trento, Italy), p. 121–128, 2006.

[Navigli, 09] R. Navigli. *Word Sense Disambiguation: A survey. ACM Computing Surveys, Vol. 41, No. 2, Article 10, Publication date, February 2009.*

[Nie et al., 99] F. Ren, L.Fan, J-Y. Nie. *SAAK Approach : How to Acquire Knowledge in an Actual application System.* IASTED International Conference on Artificial Intelligence and Soft Computing, Honolulu, p.136-140, 1999.

[Oard et al., 98] D. Oard, J. Kim, *Implicit Feedback for Recommender Systems*. In Proceedings of the AAAI Workshop on Recommender Systems, pages 81–83. 1998.

[Ounis et al., 06] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, C. Lioma. Terrier : *A High Performance and Scalable Information Retrieval Platform*. Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006), 2006.

[Pantel et al., 02] P. Pantel, D. Lin. *Discovering word senses from text.* Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (Edmonton, Alta., Canada), p.613-619, 2002.

[Pirró et al. 10] G. Pirró, J. Euzenat. *A feature and information theoretic framework for semantic similarity and relatedness*. In ISWC 2010, volume 6496 de Lecture Notes in Computer Science, p. 615–630, 2010.

[Ponte et al., 98] J.M. Ponte, W.B. Croft. *A language modeling approach to information retrieval. Research and development in information retrieval.* In Proc. of the International ACM-SIGIR Conference, p. 275–281, 1998.

[Porter, 80] M. Porter. *An algorithm for suffix stripping*. Program, 14(3), p.130-137, July, 1980.

[Proctor, 78] P. Proctor. *Longman Dictionary of Contemporary English*. Longman Group, Harlow, U.K., 1978.

[Rada et al., 89] R. Rada, H. Mili, E. Bichnell, M. Blettner. *Development and application of a metric on semantic nets*. IEEE Transaction on Systems, Man, and Cybernetics, p. 17-30, 1989.

[Resnik, 95] P. Resnik. *Disambiguating noun groupings with repect to WordNet senses*. 3thWorkshop on Very Large Corpora, p.54–68, 1995.

[Resnik, 99] P. Resnik. *Semantic Similarity in a Taxonomy : An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language*. Journal of Artificial Intelligence Research (JAIR), 11, p. 95-130, 1999.

[Robertson et al., 76] S. E. Robertson, K. Sparck Jones. *Relevance weighting of search terms*. Journal of the American Society for Information Science, 27, 129–146, 1976.

[Robertson, 77] S.E. Robertson, *The probability ranking principle in IR*. Journal of Documentation, 33 (4), p.294-304, 1977.

[Robertson, 91] S.E. Robertson. *On term selection for query expansion*. Journal of Documentation. 46(4) , p. 359-364. 1991.

[Robertson, 94a] S.E. Robertson, J. S. Walker, S Jones, M. H.-B. Gatford. *Okapi at 3*. Proceedings of the 3rd Text REtrieval Conference, p. 109-126, 1994.

[Robertson et al., 94b] S.E. Robertson, S. walker. *Some Simple Effective Approximations to the 2-Poisson Model for ProbabilisticWeighted Retrieval.* Proceedings of SIGIR 1994, p. 232-241, 1994.

[Robertson et al., 97] S. E. Robertson, S. Walker. *On relevance weights with little relevance information.* In Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval, p.16–24, ACM Press, 1997.

[Rocchio, 71] J.J. Rocchio. *Relevance feedback in information retrieval.* In The SMART Retrieval System, in Experiments in Automatic Document Processing G.Salton, editor, Prentice-Hall, Englewood Cliffs, p.. 313–323, 1971.

[Roget, 95] Roget. *Roget's II: The new thesaurus*. 3rd ed. Boston , Houghton *Mifflin*, 1995.

[Salton, 68] G. Salton. *Automatic Information Organization and Retrieval.* New York , McGraw. Hill Book Company, 1968.

[Salton, 70] G. Salton. *The SMART retrieval system: Experiments in automatic document processing. Prentice Hall*, 1970.

[Salton, 71] G. Salton. *A comparison between manual and automatic indexing methods.* Journal of American Documentation, Vol. 20, issue1, p. 61-71, 1971.

[Salton et al., 73] G. Salton, C. Yang. *On the specification of term values in automatic indexing*. In Journal of Documentation, 29, p.351–372. 1973.

[Salton et al, 83] G. Salton, E.A. Fox, H. Wu. *Extended Boolean information retrieval system*. CACM 26(11), p. 1022-1036, 1983.

[Saracevic, 75] T. Saracevic. Relevance. *A review of and a framework for the thinking on the notion in information science.* Journal of the American Society for Information Science, 26 (6), 321–343, 1975.

[Schütze et al., 95] H. Schütze, J. Pedersen. *Information retrieval based on word senses.* In Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval, p. 161-175, 1995.

[Schütze, 98] H. Schütze. *Automatic word sense discrimination*. Computational Linguistic: Special Issue on Word Sense Disambiguation, 24 (1), p.97–123, 1998.

[Seco et al., 04] N. Seco, T. Veale, J. Hayes. *An intrinsic information content metric for semantic similarity in wordnet*. In Proceedings of ECAI'2004, p. 1089–1090, Valencia, Spain, 2004.

[Sinclair, 95] J. Sinclair. *Collins Cobuild English Dictionary*. HarperCollins, 1995.

[Singhal et al., 96] A. Singhal, C. Buckley, M. Mitra. *Pivoted document length normalization*. In Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, Zurich, Switzerland, p. 21 - 29, 1996.

[Slimani et al., 07] T. Slimani, B. Ben Yaghlane, K. Mellouli. *Une extension de mesure de similarité entre les concepts d'une ontologie.* SETIT 2007 4rth International Conference: Sciences of Electronic, Technologies of Information and Telecommunications, TUNISIA, March 25-29-2007.

[Smyth et al., 05] B. Smyth, E. Balfe, J. Freyne, P. Briggs, M. Coyle, O. Boydell. *Exploiting Query Repetition and Regularity in an Adaptive Community-Based Web Search Engine*. User Modeling and User-Adapted Interaction, 14(5), 383–423. 2005.

[Soanes et al., 03] C. Soanes, A. Stevenson. *Oxford Dictionary of English.* Oxford University Press, Oxford, U.K, 2003.

[Sussna, 93] M. Sussna. *Word sense disambiguation for free-text indexing using a massive semantic network.* 2nd International Conference on Information and Knowledge Management (CIKM-1993), p.67–74, 1993.

[Tchechmedjiev, 12] A. Tchechmedjiev. *État de l'art : mesures de similarité sémantique locales et algorithmes globaux pour la désambiguïsation lexicale à base de connaissances*. 14e Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, RECITAL'12, Grenoble, France, 2012.

[Uzener et al., 98] O. Uzuner, B. Katz, D. Yuret. *Word Sense Disambiguation for Information Retrieval*. AAAI/IAAI, 1999.

[Van Riejsbergen, 79] C.J. Van Riejsbergen. *Information retrieval*. London, Butterworth, 1979.

[Vázquez et al., 04] S. Vázquez, A. Montoyo, G. Rigau. *Using Relevant Domains Ressource for Word Sense Disambiguation.* Proceeding of international conference on Artificial intelligence, (IC-AI'04), Nivada, 2004.

[Véronis et al., 90] J. Véronis, N. Ide. *Word sense disambiguation with very large neural networks extracted from machine readable dictionaries*. 13th International Conference on Computational Linguistics (COLING-1990), Vol.2, p.389–394. 1990.

[Voorhees, 93] E. Voorhees. *Using WordNet to Disambiguate Word Senses for Text Retrieval*. Proceedings of the 16thAnnual Conference on Research and Development in Information Retrieval, SIGIR'93, Pittsburgh, PA, 1993.

[Vossen, 98] P. Vossen. *EuroWordNet : A Multilingual Database with Lexical Semantic Networks*. *Kluwer* Academic Publishers, Dordrecht, The Netherlands, 1998.

[Weiss, 73] S. F. Weiss. *Learning to disambiguate*. Information Storage and Retrieval, p. 33-41, 1973.

[White et al., 02] R. White, I. Ruthven, J.M. Jose. *The use of implicit evidence for relevance feedback in web retrieval.* In Proceedings of the 24th BCS-IRSG European Colloquium on IR Research, Advances in Information Retrieval, p. 93–109, London, UK, Springer-Verlag, 2002.

[Wilks et al., 90] Y. Wilks, D. Fass, C. Guo, J.E. Mcdonald, T. *Plate, B.M. Slator*. Providing Machine Tractable Dictionary Tools. In Machine Translation, Vol.5, p.99-154. 1990.

[Wilks et al., 97] Y. Wilks, M. Stevenson. *Combining independent knowledge source for word sense disambiguation.* Conference Recent Advances in Natural Language Processing, p.1–7, 1997.

[Wong et al., 85] S. Wong, W. Ziarko, P.Wong. *Generalized vector spaces model in information retrieval*. In Proc. of the 8th ACM-SIGIR conference, p. 18-25. Montreal, Quebec, 1985.

[Woods, 97] W.A. Woods. *Conceptual indexing : A better way to organize knowledge.* Technical Report SMLI TR-97-61, Sun Microsystems Laboratories, Mountain View, CA, April. ww.sun.com/research/techrep/1997/abstract-61.html, 1997.

[Wu-Palmer, 94] Z. Wu, M. Palmer. *Verb semantics and Lexical selection*. Proceedings of the 32th Annual Meetings of the Association for Computational Linguistics, p. 133-138, 1994.

[Yarowsky, 92] D. Yarowsky. *Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora*. Proceedings of the 14th International Conference on Computational Linguistics (COLING-92). Nantes, France, August, p.454 – 460, 1992.

[Yarowsky, 00] D. Yarowsky. *Hierarchical decision lists for word sense disambiguation*. Journal Computers and the Humanities, Vol.34 (1-2), p179-186, 2000.

[Zadeh, 65] L.A. Zadeh. *Fuzzy sets.* Information and control, 8, p.338-353, 1965.

[Zhou et al., 06] X. Zhou, X. Zhang et X. Hu. *MaxMatcher : Biological Concept Extraction Using Approximate Dictionary Lookup*. In PRICAI, volume 4099, p. 1145–1149, 2006.

Annexe

A. Exemples de ressources linguistiques les plus exploitées en RI

Les ressources linguistiques jouent un rôle important dans le traitement automatique de l'ambiguïté de l'information recherchée par un SRI. Elles servent comme outils de compréhension des mots de la langue utilisés dans l'indexation des documents et requêtes, en exploitant dans la désambiguïsation de ces mots les relations sémantiques qui les relient ainsi que les définitions de leurs sens telles que définies dans un dictionnaire informatisé, ou un thésaurus ou une ontologie. L'intérêt principal étant d'améliorer la représentation des documents et requêtes traités par un système de recherche d'information afin de retrouver uniquement l'information sémantiquement pertinente au besoin d'un utilisateur. Parmi les ressources terminologiques les plus utilisées en RI, nous retrouvons le lexique WordNet [Miller, 95 ; Fellbaum, 98] et son extension aux domaines WordNetDomains [Magnini et al., 00], et le thésaurus de référence du domaine biomédical MeSh [Coletti et al., 01]. Un bref aperçu de ces ressources est donné dans ce qui suit.

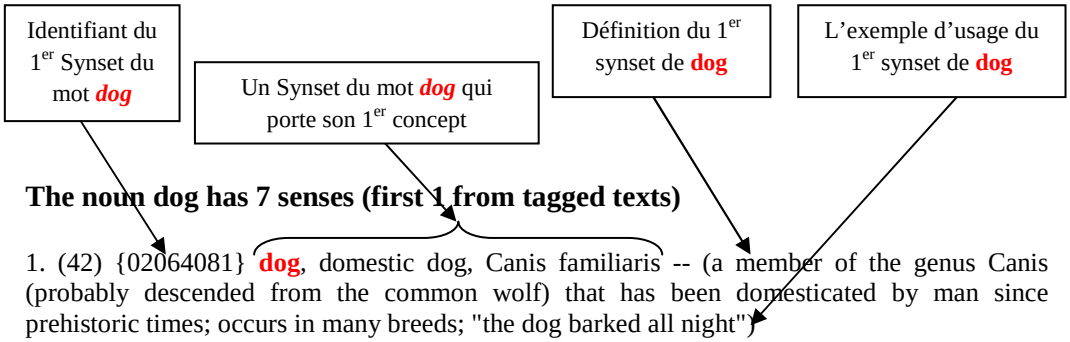
A.1 WordNet

WordNet [Miller et al, 95] est une base de données lexicale électronique, développée par des linguistes du laboratoire des sciences cognitives de l'université de Princeton, dans le but de répertorier, classifier et mettre en relations les mots simples ou composés (collocations de mots) de la langue anglaise. Ces mots de différentes catégories syntaxiques (noms, verbes, adjectifs et adverbes) sont représentés par un réseau de nœuds et des arcs qui les relient.

- Un nœud appelé *synset* (*synonym set*), défini un ensemble de mots synonymes désignant un sens ou un concept particulier. Chaque synset est identifié par un numéro, et possède une définition ou un glossaire (*gloss*) qui décrit le concept porté par ce synset en utilisant des commentaires et/ou des exemples de son usage. Un exemple de synsets de WordNet associés au mot *dog* est donné dans le tableau A.1.
- Un arc représente la relation sémantique entre deux synsets. Généralement, les relations sémantiques existantes dans la ressource linguistique WordNet relient les synsets des termes de même partie de discours (i.e des termes de même forme syntaxique). Parmi ces relations, on trouve :
  - la relation de *synonymie*, qui représente la relation de base reliant les termes appartenant au même synset ;
  - la relation de subsomption *is-a* (dite relation d'*Hyperonymie/Hyponymie*) des noms et verbes, qui relie un concept (synset) général (l'*hyperony*me) à un autre concept plus spécifique (son *hyponyme*). Par exemple : {*canine*} a pour *hyponymes* {*wolf, wild dog,*

dog}, selon l'extrait de la hiérarchie *is*-a des noms dans WordNet, illustré à travers la figure A.1.

- la relation *part-of* (dite relation *Méronymie/Holonymie*) des noms, associe à un concept X sa partie constituante du concept Y. Par exemple : {*wheel, engine*} sont des parties (ou *méronymes*) de {*car*}, et inversement {*car*} est un *holonome* de {*wheel, engine*}.



The noun dog has 7 senses (first 1 from tagged texts)

1. (42) {02064081} **dog**, domestic dog, Canis familiaris -- (a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds; "the dog barked all night")
2. {09963790} frump, **dog** -- (a dull unattractive unpleasant girl or woman; "she got a reputation as a frump"; "she's a real dog")
3. {09874274} **dog** -- (informal term for a man; "you lucky dog")
4. {09740502} cad, bounder, blackguard, dog, hound, heel -- (someone who is morally reprehensible; "you dirty dog")
5. {07571296} frank, frankfurter, hotdog, hot dog, **dog**, wiener, wienerwurst, weenie -- (a smooth-textured sausage of minced beef or pork usually smoked; often served on a bread roll)
6. {03857710} pawl, detent, click, **dog** -- (a hinged catch that fits into a notch of a ratchet to move a wheel forward or prevent it from moving backward)
7. {02685221} andiron, firedog, **dog**, dog-iron -- (metal supports for logs in a fireplace; "the andirons were too hot to touch")

The verb dog has 1 sense (first 1 from tagged texts

1. (2) {01983615} chase, chase after, trail, tail, tag, give chase, **dog**, go after, track -- (go after with the intent to catch; "The policeman chased the mugger down the alley"; "the dog chased the rabbit")

Tableau A.1 : Les synsets (concepts) de WordNet correspondants au mot *dog*.

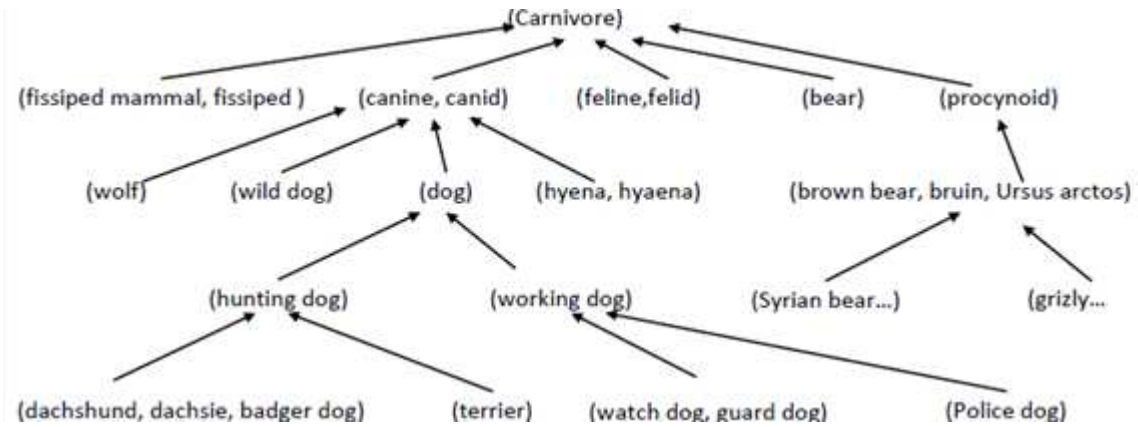


Figure A.1 : Extrait de la hiérarchie *is-a* des noms dans WordNet correspondant au synset *dog*.

A.2 WordNetDomains

WordNetDomains est une ressource lexicale développé par Magnini [Magnini et al., 00] pour étendre la ressource linguistique WordNet, en étiquetant ses synsets (concepts des termes) par des labels de domaines sémantiques (tels que les les domaines : *Sport*, *Politic*, *Medicine*, *Economic*, ...etc). Ces domaines sont organisés selon une hiérarchie définissant la relation de spécialisation/généralisation entre eux. Par exemple, le domaine *Tennis* est plus spécifique que le domaine *Sport*, et le domaine *Architecture* est plus général que le domaine *Buildings*. Un extrait de cette hiérarchie est donné dans le tableau A.2. Le domaine *Top-Level* représente la racine de cette hiérarchie. Par opposition le domaine *Factotum*, indépendant de cette hiérarchie, est un domaine fonctionnel regroupant tous les synsets des termes, dans WordNet, qui n'appartiennent à aucun domaine particulier mais qui peuvent apparaître avec des termes associés à d'autres domaines. Ce domaine (*Factotum*) est relié à des sous-domaines tels que : *Quality, Color, Number, …etc.*

Top_level	Humanities	History		
		Linguistics	Grammar	
		Literature	Philology	
		Philosophy	Psychoanalysis	
		Art	Music	
			Plastic_Arts	Jewellery
				Sculpture
			Theatre	
			Cinema	
		Paranormal		
		…		
	…			
	Pure_Science			
		Biology	Anatomy	
		Animals		
		Earth	Geology	
			Geography	
		Mathematics		
		Physics	Acoustics	
		…		
	Social_Science	Economy	Finance	Money
		Politics		
		Fashion		
		Military		
		…		
Factotum	Quality			
	Number			
	…			

Tableau A.2 : Extrait de la hiérarchie généralisation/spécialisation de WordNetDomains.

Un syset de WordNet est annoté dans WordNetDomains par un seul ou plusieurs domaines décrivant son (ou ses) domaines d'usage. Par exemple, les domaines associés aux synsets du mot *bank* est donné à travers le tableau A.3.

Les senses (synsets) du mot *bank*	**Domaines**
1. (883) depository financial institution, **bank**, banking concern, banking company -- (a financial institution that accepts deposits and channels the money into lending activities; "he cashed a check at the bank"; "that bank holds the mortgage on my home")	***ECONOMY,***
2. (99) **bank** -- (sloping land (especially the slope beside a body of water); "they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents")	***GEOGRAPHY, GEOLOGY***
3. (76) **bank** -- (a supply or stock held in reserve for future use (especially in emergencies))	***ECONOMY***
4. (54) **bank**, bank building -- (a building in which the business of banking transacted; "the bank is on the corner of Nassau and Witherspoon")	***FACTOTUM, ECONOMY***
5. (7) **bank** -- (an arrangement of similar objects in a row or in tiers; "he operated a bank of switches")	***FACTOTUM***
6. (6) savings bank, coin bank, money box, **bank** -- (a container (usually with a slot in the top) for keeping money at home; "the coin bank was empty")	***ECONOMY***
7. (3) **bank** -- (a long ridge or pile; "a huge bank of earth")	***GEOGRAPHY, GEOLOGY***
8. (1) **bank** -- (the funds held by a gambling house or the dealer in some gambling games; "he tried to break the bank at Monte Carlo")	***ACONOMY, PLAY***
9. (1) **bank**, cant, camber -- (a slope in the turn of a road or track; the outside is higher than the inside in order to reduce the effects of centrifugal force)	***ARCHITECTURE***
10. **bank** -- (a flight maneuver; aircraft tips laterally about its longitudinal axis (especially in turning); "the plane went into a steep bank")	***TRANSPORT***

Tableau A.3 : Domaines de WordNetDomains associés aux synsets du mot *bank*.

A.3 MeSh (MEdical Subject Headings)

Le thésaurus MeSh [Coletti et al., 01] est un vocabulaire normalisé utilisé dans l'indexation des documents biomédicaux qui sont issus des bases de données de la National Library of Medicine (NLM[34]).

[34] http://www.nlm.nih.gov/

Un terme dans le vocabulaire MeSh est composé d'un seul mot ou d'un ensemble de mots synonymes représentant un concept biomédical. Chaque concept dans ce thésaurus est désigné par un terme préféré (dit concept préféré) représentant le sens le plus large de ce concept. A titre d'exemple, le concept *Pain* désigne le concept préféré des termes {*Pain ; Ache ; Pain, Burning ; Pain, Crustring ; Pain, Migratory ; Pain, Radiating ; Pain, Spliting ; Suffring, Physical*}.

Les concepts qui sont proches sémantiquement sont représentés par un descripteur qui porte le nom du concept le plus préféré. L'ensemble des descripteurs MeSH sont répartis en 16 catégories (domaines) recouvrant la biologie, la médecine et les domaines connexes [Dinh, 12]. La figure A.2, illustre les différents domaines de ce thésaurus.

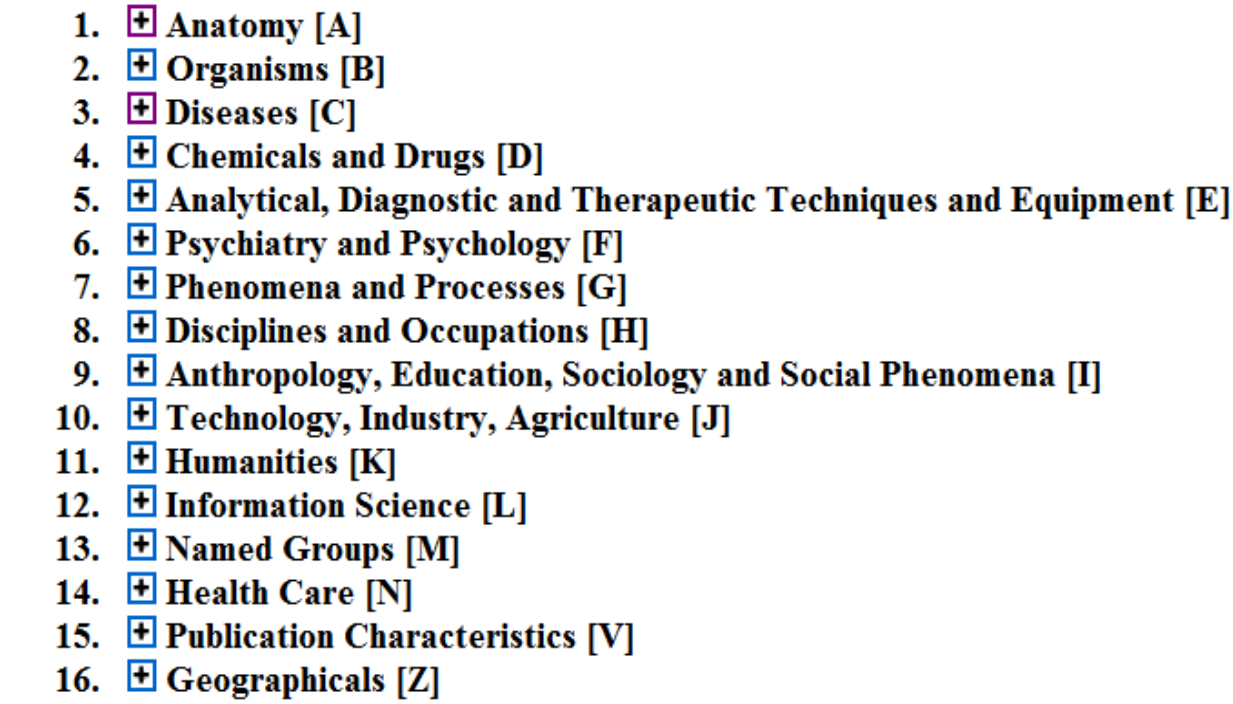


Figure A.2: Les différents domaines du thésaurus MeSh.

Une catégorie (domaine) dans MeSh est structurée en arborescence hiérarchique de descripteurs. Chaque descripteur dans cette arborescence est représenté par un code alphanumérique, où une lettre indique de son domaine d'appartenance et une chaine numérique définie sa localisation dans ce domaine.

Certains descripteurs possèdent plusieurs localisations au sein la même catégorie ou de catégories différentes. A titre d'exemple, le descripteur *Pain* appartient à plusieurs domaines (*Diseases*, *Psychiatry and Psychology*, *Phenomena and Processes*), dont ses localisations sont : *C10.597.617*, *C23.888.592.612*, *C23.888.646*, *F02.830.816.444* et *G11.561.600.810.444*. La figure A.3 décrit le descripteur *Pain* dans le domaine *Diseases*, dont sa localisation est *C10.597.617*.

Figure A.3 : Le descripteur *Pain* (*C10.597.617*) dans l'arborescence du domaine *Diseases* du thésaurus MeSh.