

Recherche d'information sémantique dans les documents biomédicaux : approche basée sur le sens précis des concepts

Duy Dinh Lynda Tamine

*Université Paul Sabatier
118 route de Narbonne
31062 Toulouse - Cedex 9
dinh@irit.fr, lechani@irit.fr*

RÉSUMÉ. Ce papier aborde le problème de l'ambiguïté des termes dans les documents, en particulier dans le domaine de la biomédecine. Notre objectif est de proposer une méthode de désambiguïsation des termes ambigus utilisés dans la biomédecine et de l'intégrer dans un modèle d'indexation sémantique et de recherche d'information basé sur le sens des concepts dans les documents ainsi que de la requête. Nous exploitons l'architecture poly-hiérarchique du thésaurus MeSH (Medical Subject Headings) pour désambiguïser les concepts et les indexer avec leur sens le plus adéquat associé dans chaque document. L'évaluation des résultats de nos expérimentations sur la collection de TREC9-FT 2000 montre une amélioration de la performance par rapport aux modèles d'indexation classiques dans le domaine de la Recherche d'Information.

ABSTRACT. This paper discusses the term ambiguity problem for indexing biomedical literature. Our objective is to propose a method of disambiguating biomedical terms and integrate them into a semantic indexing and retrieval model. We exploit the poly-hierarchical structure of the Medical Subject Headings (MeSH) to disambiguate terms having more than one sense. Unambiguous terms are then semantically indexed with their appropriate sense. The experimental evaluation carried out on the TREC9-FT 2000 collection shows that our semantic approach of indexing and retrieval is promising.

MOTS-CLÉS : Indexation sémantique, Recherche d'Information sémantique, Désambiguïsation, Ressources biomédicales

KEYWORDS: Semantic Indexing, Semantic Information Retrieval, Word Sense Disambiguation, Biomedical Resources

1. Introduction

Depuis l'avènement d'Internet, des bibliothèques digitales et de la large accessibilité des médias, les volumes d'informations évoluent de manière significative tant en volume qu'en qualité. Plus précisément, dans le domaine biomédical, les services de production et d'accès à l'information ne cessent de se diversifier. A titre d'exemple, MEDLINE¹ (Medical Literature Analysis and Retrieval System Online) est la base de données bibliographiques de premier ordre, développée par la NLM (US National Library of Medicine), qui contient plus de 19 millions de références d'articles en science de la vie, notamment de la biomédecine. Un trait distinctif de MEDLINE est que les documents sont indexés manuellement ou automatiquement avec les concepts du thésaurus MeSH (Medical Subject Headings). Le portail PubMed² de la NCBI (National Center for BioTechnology Information) fournit un accès aux publications scientifiques de la base d'articles MEDLINE. De manière générale, les informations biomédicales sont exprimées sous forme de langage naturel, ce qui ne présente relativement pas de problème pour l'humain mais demeure encore un grand défi pour les processus automatiques de traitement de l'information. Une des problématiques majeures est l'*ambiguïté* dans le texte, en particulier dans la biomédecine. Par exemple, le terme "has" en anglais indique à la fois un verbe et le nom d'une protéine. Normalement, les sens d'un terme sont définis dans un dictionnaire, une encyclopédie, un thésaurus, ou encore une ontologie, etc. Un terme est dit *ambigu* s'il a plus de deux sens dans les contextes différents. La reconnaissance et l'affectation du sens le plus adéquat aux termes ambigus dans un contexte donné font référence à la désambiguïsation (Word Sense Disambiguation).

Afin de répondre à cette problématique, différentes approches de désambiguïsation ont été proposées. Ces dernières peuvent se subdiviser en quatre catégories : *apprentissage supervisé* (Lee *et al.*, 2004) (Liu *et al.*, 2004), *apprentissage non-supervisé* (Yarowsky, 1995), *apprentissage semi-supervisé* (Abney, 2002) et *approche basée sur la connaissance* (Knowledge-based) (Lesk, 1986), (Gale *et al.*, 1993), (Mihalcea, 2005). Les méthodes basées sur l'apprentissage supervisé comme les arbres de décision, les machines à vecteurs de support, l'entropie maximale, la classification naïve bayésienne, etc. utilisent les corpus d'apprentissage étiquetés pour entraîner les classificateurs. La classification non-supervisée s'applique sur les corpus non-étiquetés pour en extraire plusieurs groupes de textes sémantiquement similaires. L'approche semi-supervisée est basée sur un petit échantillon étiqueté pour entraîner d'abord les classificateurs initiaux, puis sur une plus large collection non-étiquetée pour réentraîner et enrichir les classificateurs au fur et à mesure du processus itératif de l'apprentissage. L'approche basée sur la connaissance du domaine utilise des ressources externes comme les dictionnaires (*Machine Readable Dictionaries - MRDs*), thésaurus, ontologies, etc. comme des ressources sémantiques pour représenter la connaissance du domaine. La plupart des approches dédiées à la désambiguïsation du texte biomédical se base sur l'apprentissage supervisé (Liu *et al.*, 2004), (Leroy *et al.*, 2005), (Joshi *et*

1. <http://www.nlm.nih.gov/pubs/factsheets/medline.html>

2. <http://www.ncbi.nlm.nih.gov/pubmed>

al., 2005), ce qui les rend complètement dépendant des corpus étiquetés et induisent un coût élevé pour annoter manuellement et maintenir la connaissance.

La désambiguïsation des termes biomédicaux est importante et primordiale dans l'indexation sémantique des ressources biomédicales. Les travaux de la recherche d'information biomédicale ont été initiés et intégrés dans le développement de l'outil d'indexation des textes biomédicaux MTI (Medical Text Indexer) (Aronson *et al.*, 2004) aux Etats-Unis ou F-MTI (French-Multiterminology Indexer) (Pereira *et al.*, 2008) en France. L'indexation sémantique des documents de la littérature biomédicale vise à assister les annotateurs dans leurs tâches d'indexation et est vue comme une recommandation automatique de descripteurs (Kim *et al.*, 2001), (Cai *et al.*, 2004) ou de couples de descripteurs/qualificatifs (Névéol *et al.*, 2007). Concernant particulièrement l'indexation des ressources francophones, les travaux intègrent une ou plusieurs terminologies médicales (MeSH, ICD-10, CCAP, TUV) en associant des descripteurs MeSH aux documents dans le catalogue CiSMEF (Névéol *et al.*, 2006), (Pereira *et al.*, 2008).

Ce papier propose une approche d'indexation sémantique basée sur le sens précis de concepts dans le document ainsi que de la requête. Notre approche d'indexation et de recherche d'information sémantique exploite la structure poly-hiérarchique du thésaurus MeSH (Medical Subject Headings) pour désambiguïser les concepts ambigus dans les documents et les requêtes. La suite de cet article est organisée comme suit : La section 2 présente un état de l'art sur l'indexation de la littérature biomédicale et puis positionne notre contribution dans ce cadre. La section 3 décrit notre méthode de désambiguïsation et le processus d'indexation des documents biomédicaux basée sur le sens des concepts biomédicaux. Une évaluation expérimentale est présentée et discutée dans la section 4. La section 5 conclut le papier et annonce des perspectives.

2. Ambiguïté et indexation sémantique de documents biomédicaux

2.1. Problématique de l'ambiguïté

L'ambiguïté est un problème commun dans les textes généraux ainsi que dans les domaines spécifiques comme la biomédecine. A titre d'exemple, une étude de (Schiemann *et al.*, 2008) a montré qu'il existe 175 termes désignant à la fois des espèces et des protéines, 67 termes désignant des médicaments et des protéines, 123 termes désignant des cellules et des tissus. Dans le thésaurus MeSH, un concept peut avoir plusieurs sens. A titre illustratif, le concept "Pain" appartient à quatre branches de trois hiérarchies dont les concepts les plus génériques sont : *Nervous System Disease* (C10); *Pathological Conditions, Signs and Symptoms* (C23); *Psychological Phenomena and Processes* (F02); *Musculoskeletal and Neural Physiological Phenomena* (G11). Des méthodes de désambiguïsation ont été proposées pour associer le sens le plus adéquat aux concepts dans leur contexte d'apparition. Différents travaux ont traité l'ambiguïté des termes issus de l'UMLS (Widdows *et al.*, 2003), des acronymes (Gaudan *et al.*, 2005) et expressions de gènes (Andreopoulos *et al.*, 2008) en proposant des

méthodes basées sur les distributions des fréquences des termes dans les sources d'informations ou en entraînant les différents sens des termes ambigus en se basant sur des méthodes d'apprentissage automatique ou de classification.

La plupart des approches de désambiguïsation dans le texte biomédical sont basées sur l'apprentissage supervisé (Liu *et al.*, 2004), (Gaudan *et al.*, 2005), (Leroy *et al.*, 2005), (Joshi *et al.*, 2005), (Andreopoulos *et al.*, 2008), (Mohammad *et al.*, 2004). Le travail de (Andreopoulos *et al.*, 2008) a utilisé le méta-thésaurus UMLS pour désambiguïser les termes biomédicaux dans les documents en entraînant un classificateur naïf bayésien qui prend en compte la co-occurrence de termes dans les documents. Les abréviations dans les articles de MEDLINE sont résolues par les machines à vecteurs de support (SVMs) qui intègrent un dictionnaire d'abréviations avec leurs formes complètes. Brièvement, les travaux exploitent les caractéristiques linguistiques et lexicales connues dans les textes en général et les appliquent sur les textes biomédicaux comme : l'étiquette grammaticale (*Part-Of-Speech*), les relations sémantiques entre les mots (Leroy *et al.*, 2005), l'unigramme, le bigramme (Joshi *et al.*, 2005), les relations syntaxiques et lexicales (Mohammad *et al.*, 2004). Récemment, le travail de (Stevenson *et al.*, 2008) utilise les identifiants uniques de concepts dans l'UMLS, obtenus par l'outil MetaMap (Aronson, 2001), et les termes de MeSH qui sont manuellement annotés dans les articles de MEDLINE pour construire des vecteurs de caractéristiques en entraînant les classificateurs basés sur les vecteurs à machine de support, les réseaux bayésiens et les modèles vectoriels. Cependant, les méthodes abordées sont basées complètement sur des corpus d'apprentissage qui demandent un coût d'annotation en terme de temps et compétence requise.

L'ambiguïté a un impact important non seulement sur la performance des applications du traitement du langage naturel en général mais aussi sur la performance de la recherche d'information (RI). Par conséquent, les travaux de recherche ont tenté de représenter les documents de manière sémantique en prenant en compte une ou plusieurs ressources sémantiques. Nous synthétisons dans ce qui suit les travaux de représentation et/ou indexation sémantique de la littérature biomédicale.

2.2. Indexation sémantique de documents biomédicaux

Les documents biomédicaux, notamment les publications scientifiques du domaine sous forme d'articles de journaux, d'ouvrages, de rapports ou guides de bonnes pratiques, sont continuellement croissants. Leur accessibilité et indexation sont particulièrement confrontées aux problèmes de la synonymie, polysémie et présence d'acronymes (Hersh, 2008). Il existe deux principales approches d'indexation qui sont l'approche *manuelle* et *(semi)-automatique*. L'indexation manuelle a été initiée par la NLM et a essentiellement servi à l'association de concepts ou descripteurs sémantiques de MeSH aux résumés des documents de la base MEDLINE³. L'accroissement

3. <http://www.nlm.nih.gov/pubs/factsheets/medline.html>

des articles publiés dans MEDLINE a conduit au développement d'outils d'indexation (semi)-automatique tel que MTI (Medical Text Indexer) (Aronson, 2001). L'indexation (semi)-automatique a été plus largement utilisée dans le domaine et vue comme une recommandation automatique de concepts ou descripteurs sémantiques (Kim *et al.*, 2001), (Cai *et al.*, 2004), ou de couples de descripteurs/qualificatifs (Névéol *et al.*, 2007). Récemment, le travail de (Trieschnigg *et al.*, 2009) montre que la recommandation ou l'affectation des descripteurs MeSH dans les documents peut être considérée comme la catégorisation textuelle dans le sens que le classificateur décide lui-même si un descripteur est potentiellement associé à chaque document. Concernant particulièrement l'indexation des ressources francophones, les travaux intègrent une ou plusieurs terminologies médicales (MeSH, ICD-10, CCAP, TUV, ...) en associant des descripteurs MeSH aux documents dans le catalogue CiSMEF (Névéol *et al.*, 2006), (Pereira *et al.*, 2008).

2.3. Objectifs de notre contribution

Dans ce cadre général, notre contribution présentée dans le domaine est résumée à travers les points suivants :

1) Nous proposons une méthode de désambiguïsation qui identifie d'abord les concepts biomédicaux issus de MeSH, et puis associe le sens le plus adéquat à chaque concept dans le contexte du document. Comparativement aux autres méthodes de désambiguïsation, notre approche possède les caractéristiques suivantes : (1) elle est basée sur le contexte local du document sans avoir besoin de corpus d'apprentissage, et (2) exploite l'architecture poly-hiérarchique de MeSH pour identifier le sens correct de concepts ambigus. Cette méthode de désambiguïsation présentée dans (Dinh *et al.*, 2010) a été évaluée sur un corpus de dossiers médicaux de patients. Nous l'adaptions dans ce travail à un corpus de documents de la littérature biomédicale.

2) Nous proposons un modèle d'indexation et de recherche d'information sémantique basé sur le sens local désambiguïté des concepts issus du thésaurus MeSH. A notre connaissance, c'est une première tentative de l'exploitation du sens issu de la poly-hiérarchie de MeSH pour l'indexation et l'appariement de documents biomédicaux.

3. Recherche d'information sémantique basée sur le sens de concepts

Nous proposons par la suite un schéma d'indexation sémantique de documents biomédicaux. Nous présentons dans cette section notre algorithme de désambiguïsation, puis nous détaillons le processus de la recherche d'information sémantique basé sur le sens non ambigu de concepts.

3.1. Algorithme de désambiguïsation de concepts biomédicaux

Notre algorithme de désambiguïsation consiste à sélectionner le sens le plus adéquat d'un concept dans le contexte local du document. Pour cela, nous rappelons les éléments clés de MeSH, puis décrivons notre méthode de désambiguïsation.

Dans le vocabulaire de MeSH, un *terme préféré*, utilisé pour l'indexation, représente le nom d'un concept. Les termes non-préférés sont utilisés pour la recherche d'information. Dans l'architecture poly-hiérarchique de MeSH, un concept est représenté par un noeud appartenant éventuellement à une (non-ambigu) ou plusieurs hiérarchies (ambigu). Chaque hiérarchie correspond à un des seize domaines de MeSH (*A-Anatomie, B-Organismes, C-Maladies, ...*). Notre méthode est basée sur les définitions et notations suivantes (Dinh *et al.*, 2010) :

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

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

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

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

5) **Définition 5** : La relation **is-a** relie les concepts d'une même hiérarchie.

Notre méthode de désambiguïsation est basée sur les hypothèses suivantes :

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

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

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

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

$$\left\{ \begin{array}{l} (s_1, s_2) = \sum_{s_1 \in syn(c_1), s_2 \in syn(c_2)} sim(s_1, s_2) \quad \text{if } k \leq 2 \\ s_k = \arg \max_{s \in syn(c_k)} \left(\sum_{s \in syn(c_k)} sim(s_{k-1}, s) \right) \quad \text{if } k > 2 \end{array} \right. \quad [1]$$

où s_k : le sens du concept c_k ,
 $syn(c_k)$: l'ensemble de sens du concept c_k ,
 $sim(s_1, s_2)$: similarité basée sur les hiérarchies de s_1 and s_2 .

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

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

où $length(s_1, s_2)$ est le chemin le plus court entre s_1 and s_2 , and D est le niveau le plus profond de la hiérarchie.

3.2. Le processus de recherche d'information sémantique

Notre objectif ici est de générer un index sémantique contenant à la fois des concepts identifiés selon l'approche de désambiguïsation précédente et les mots simples qui ne correspondent pas à des entrées de MeSH. Plus précisément, chaque concept est annoté par son sens local désambiguïsé, identifié par son numéro d'arbre dans MeSH. Nous calculons par la suite la similarité entre les documents et la requête selon un schéma sémantique basé sur le sens des concepts dans la requête ainsi que dans les documents. Notre processus d'indexation et de recherche d'information sémantique est décrit par les étapes suivantes :

Etape 1 : Représentation des documents. Etant donné le document initial D_i qui contient à la fois des concepts du thésaurus et des mots simples du vocabulaire, D_i peut être représenté formellement comme suit :

$$\begin{aligned} D_i^s &= \{d_{1i}^s, d_{2i}^s, \dots, d_{mi}^s\} \\ D_i^w &= \{d_{1i}^w, d_{2i}^w, \dots, d_{ni}^w\} \end{aligned} \quad [3]$$

où D_i^s, D_i^w sont respectivement l'ensemble de concepts et mots simples, m et n sont respectivement le nombre de concepts et mots du document D_i , d_{ji}^s est le j -ième concept et d_{ji}^w est le j -ième mot du document D_i .

Etape 2 : Représentation de la requête. Les requêtes sont traitées de la même manière que les documents. Par conséquent, la requête Q peut être représentée formellement comme suit :

$$\begin{aligned} Q^s &= \{q_1^s, q_2^s, \dots, q_u^s\} \\ Q^w &= \{q_1^w, q_2^w, \dots, q_v^w\} \end{aligned} \quad [4]$$

où Q^s, Q^w sont respectivement l'ensemble de concepts et de mots simples, u et v sont respectivement le nombre de concepts et mots de Q , q_k^s est le k -ième concept et q_k^w est le k -ième mot de Q .

Etape 3 : Calcul de la pertinence. La mesure de pertinence du document D_i vis-à-vis de la requête Q considère dans notre cas deux principaux facteurs : (1) l'adéquation du sens de concepts dans le document et de la requête, (2) la spécificité de concepts dans le document. Formellement :

$$RSV(Q, D_i) = RSV(Q, D_i^w) + RSV(Q, D_i^s) \quad [5]$$

où $RSV(Q, D_i^w)$ est la mesure de la similarité *TF-IDF* basée sur les mots et $RSV(Q, D_i^s)$ est la pertinence basée sur le sens du concept du document vis-à-vis de la requête, calculée comme suit :

$$\begin{aligned} RSV(Q, D_i^w) &= \sum_{q_k^w \in Q} TF_i(q_k^w) * IDF(q_k^w) \\ RSV(Q, D_i^s) &= \sum_{q_k^s \in Q} \alpha_k * (1 + h(q_k^s)) * TF_i(q_k^s) * IDF(q_k^s) \end{aligned} \quad [6]$$

où TF_i : la fréquence normalisée du mot q_k^w ou du concept q_k^s dans le document D_i , IDF : la fréquence de document inverse normalisée de q_k^w ou q_k^s dans la collection, α_k : le facteur d'adéquation du sens du concept q_k^s entre D_i et Q , $h(q_k^s)$: la spécificité de q_k^s associée à son propre sens dans la requête, calculée comme suit :

$$h(q_k^s) = \frac{niveau(q_k^s)}{MaxDepth} \quad [7]$$

où $niveau(q_k^s)$: niveau de profondeur de q_k^s , $MaxDepth$: profondeur maximale de la hiérarchie.

$$\alpha_k = \left\{ \begin{array}{ll} 1 & \text{if } sens(q_k^s, Q) = sens(q_k^s, D_i) \\ 1 - \beta & \text{sinon} \end{array} \right\} \quad [8]$$

où $sens(q_k^s, Q)$ (resp. $sens(q_k^s, D_i)$) indique le sens du concept q_k^s dans la requête (resp. dans le document D_i) (cf. définition 4); β est un paramètre expérimental dont la valeur est dans l'intervalle $[0, 1]$. En effet, nous supposons qu'un concept à un niveau de la spécificité plus élevé est plus pertinent pour l'utilisateur. La spécificité dans la formula 7 est considérée afin de privilégier les documents contenant les concepts au niveau de la spécificité plus élevé. Le coefficient sémantique α_k est considéré pour atténuer le poids du document D_i où le sens du concept q_k^s est différent de celui identifié dans la requête.

4. Evaluation expérimentale

L'objectif de notre évaluation expérimentale est d'étudier l'impact de la désambiguïsation de concepts de MeSH sur la performance de la recherche d'information biomédicale. Nous décrivons dans ce qui suit le cadre d'évaluation et présentons, puis discutons les résultats obtenus.

4.1. Cadre d'évaluation

– *Collection test* : Nous utilisons la collection OHSUMED, proposée dans le cadre de la tâche TREC9-Filtering en 2000, qui est constituée des titres et/ou résumés de

270 journaux médicaux publiés entre 1987-1991 (Hersh *et al.*, 1994). Un document contient six champs : *titre* (.T), *résumé* (.W), *concepts indexés de MeSH* (.M), *auteur* (.A), *source* (.S), and *publication* (.P). Quelques caractéristiques statistiques de la collection sont données dans le tableau 1.

Nous avons testé 48 requêtes, chacune est fournie avec un ensemble de documents jugés pertinents par un groupe de médecins. Le champ *titre* indique la *description du patient* (patient description) et le champ *description* annonce *le besoin en information* (information request).

Nombre de documents	293.856
Longueur moyenne du document	100
Nombre de requêtes	48
Longueur moyenne de la requête	6 (TITRE) 12 (TITRE+DESC)
Nombre moyen de concepts/requête	1.50 (TITRE) 3.33 (TITRE+DESC)
Nombre de documents jugés pertinents/requête	50

Tableau 1. *Statistiques de la collection test*

– *Mesures d'évaluation* : Nous utilisons les mesures P@5, P@10 qui sont respectivement la précision moyenne aux 5, 10 premiers documents retournés et MAP (*Mean Average Precision*) sur l'ensemble de 48 requêtes. Pour chaque requête, les 1000 premiers documents sont renvoyés par le système et les précisions moyennes (P@5, P@10, MAP) sont calculées pour mesurer la performance de la RI.

– *Medical Subject Headings* : Nous avons utilisé le thésaurus de référence du domaine biomédical développé par la NLM aux Etats-Unis. Plus précisément, c'est la traduction en français de la version anglaise des termes MeSH établie par l'Institut National de la Santé et de la Recherche Médicale (INSERM) qui est utilisée. Le thésaurus MeSH dans sa version 2009 est composée d'environ 25,186 entrées ; chacune correspond à un concept préféré (main heading) pour l'indexation.

4.2. Résultats expérimentaux

Pour évaluer la performance de notre méthode de désambiguïsation et son impact sur la performance de notre approche d'indexation et d'appariement sémantiques proposée, nous avons réalisé deux séries d'expérimentations : la première est basée sur l'indexation classique de la partie *titre* et *résumé* d'articles de MEDLINE en utilisant la configuration standard sous la plateforme Terrier (<http://ir.dcs.gla.ac.uk/terrier/>) avec le schéma de pondération de référence OKAPI BM25 (Robertson *et al.*, 1998). Cette configuration est utilisée comme la base d'évaluation comparative (baseline), notée *BM25*. La seconde série d'expérimentations concerne notre méthode d'indexation sémantique qui se décline selon deux scénarios :

1) le premier est basé sur la sélection naïve du premier sens du concept trouvé dans le thésaurus, appelé *WSD-0*,

2) le second est basé sur notre méthode de désambiguïsation, appelée *WSD-1*.

Nous utilisons à la fois les termes qui représentent les entrées de MeSH (*concepts* ou *main headings*), et les mots simples du vocabulaire qui ne font pas partie des entrées de ce thésaurus. Dans l'approche d'indexation classique, les documents et les requêtes sont indexés en utilisant la plateforme Terrier.

Dans notre approche basée sur le sens du concept qui intègre l'information sémantique du concept, les documents et les requêtes sont d'abord désambiguïsés et indexés avec les sens appropriés des concepts de MeSH. Puis, le schéma de pondération est appliqué à chaque terme dans la requête en utilisant la formule 6. Des expérimentations préliminaires nous ont permis d'ajuster le paramètre β (voir la formule 8) à 0.15.

Le tableau 2 présente les performances de recherche pour les requêtes courtes (*titre*) et pour les requêtes longues (*titre et description*). La figure 1 montre les taux d'accroissement de notre méthode par rapport à la baseline. Nous avons obtenu les résultats suivants : notre méthode *WSD-1* dépasse la *baseline* au niveau de la performance quelque soit la longueur de la requête. Le taux d'accroissement obtenu de notre méthode *WSD-1* par rapport à la baseline est de 5.61% pour les requêtes courtes et de 7.48% pour les requêtes longues. Cela montre l'intérêt de la prise en compte de la sémantique du document et de la requête en même temps avec la spécificité du document ainsi que de la requête dans le processus de la RI. En plus, les résultats montrent que la sélection naïve du sens de concepts (*WSD-0*) n'améliore pas la performance de la recherche (P@10 de la méthode *WSD-0* se détériore). En revanche, une affectation correcte de sens à chaque concept dans le document permet d'améliorer la performance de la RI. Nous avons observé que la méthode *WSD-1* donne toujours une meilleure précision que *WSD-0* (5.61% vs. 0.7% de la MAP par rapport à la baseline).

Nous avons également testé l'évaluation de la requête avec *titre* et *titre et description* à la fois pour démontrer l'impact de la longueur de la requête sur la performance de la RI. En comparant les résultats obtenus dans les tableaux 2a et 2b, nous nous apercevons que la chaîne sémantique inspirée par notre méthode de désambiguïsation améliore mieux pour les requêtes longues (taux d'accroissement de 7.48%) que pour les requêtes courtes (taux d'accroissement de 5.61%). En effet, pour les requêtes longues, notre méthode identifie mieux le sens de chaque concept qui révèle son propre niveau de la spécificité dans le document. Nous pouvons confirmer que notre méthode *WSD-1* donne une meilleure précision que *WSD-0* quelque soit la longueur de la requête (7.48% vs. 5.12% du taux d'accroissement de la MAP par rapport à la baseline).

Nous avons mené une analyse plus fine au niveau de la requête pour vérifier l'impact de la spécificité en fonction de la longueur de la requête et le nombre de concepts utilisés dans la requête. Comme présenté dans le tableau 3, pour chaque requête, nous

Mesure	BM25	WSD-0	WSD-1
P@5	0.17500	0.1792	0.19170
P@10	0.18540	0.1771	0.18750
MAP	0.10270	0.1034	0.10800

(a) Requêtes sous forme de titre

Mesure	BM25	WSD-0	WSD-1
P@5	0.50420	0.50830	0.52080
P@10	0.45630	0.46040	0.47500
MAP	0.24210	0.25450	0.26110

(b) Requêtes sous forme de titre et description

Tableau 2. Résultats officiels sur la collection test OHSUMED

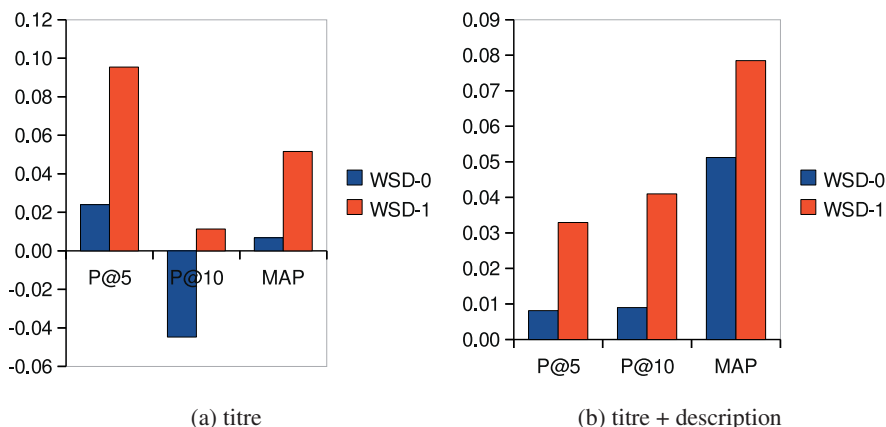


Figure 1. Taux d'accroissement par rapport à la baseline

calculons la spécificité moyenne et nous obtenons des valeurs entre 2 et 6. Les requêtes sont ensuite regroupées selon leur spécificité moyenne qui est la moyenne de la spécificité des concepts de la requête. Pour chaque groupe de requêtes, nous calculons la longueur moyenne, le nombre moyen de concepts utilisés, et le taux moyen d'accroissement. Nous nous apercevons que plus la requête est spécifique, c'est-à-dire, les concepts de la requête sont plus spécifiques, plus le nombre de concepts utilisés dans la requête diminue. Cela pourrait s'expliquer par le fait que quelques-uns des concepts les plus spécifiques couvrent suffisamment le besoin de l'utilisateur alors que plus de concepts génériques sont nécessaires pour mieux exprimer son besoin en information. Dans la plupart des cas, notre approche privilégie les documents contenant des concepts à un niveau de spécificité le plus élevé et montre une amélioration par rapport à la baseline. Toutefois, si la requête est longue mais le nombre de concepts ayant

Spécificité moyenne	Longueur moyenne de la requête	Nombre moyen de concepts	Taux d'accroissement
2	12.00	4.20	4.50
3	11.47	3.82	10.84
4	13.22	3.94	11.41
5	8.83	2.50	11.32
6	11.00	2.50	-13.38

Tableau 3. Analyse de résultats en fonction de la spécificité de la requête

un degré de spécificité élevée dans la requête est moindre, notre approche tend à retourner les premiers documents contenant ces concepts. Cela peut être la cause de la dégradation de la performance lorsque quelques-uns des concepts les plus spécifiques ont un impact important sur d'autres termes dans la requête.

5. Conclusion

Dans ce travail, nous avons proposé et évalué une approche d'indexation sémantique basée sur le sens des concepts. Notre approche s'appuie sur la méthode de désambiguïsation de concepts ambigus issus de MeSH. Cette méthode de désambiguïsation est basée sur le contexte local du document et de la requête où apparaissent les concepts. Le modèle d'indexation et d'appariement sémantique proposé prend en compte l'adéquation du sens des concepts et leur spécificité dans le document et de la requête. L'évaluation de la méthode d'indexation et d'appariement sémantique sur le corpus standard OHSUMED montre que nos résultats sont prometteurs. Dans nos futurs travaux, nous envisageons de procéder à une expansion automatique de la requête à l'aide de concepts désambiguïsés de la hiérarchie indiquée par le sens le plus approprié des concepts.

6. Bibliographie

- Abney S. P., « Bootstrapping », *ACL*, p. 360-367, 2002.
- Andreopoulos B., Alexopoulou D., Schroeder M., « Word Sense Disambiguation in biomedical ontologies with term co-occurrence analysis and document clustering », *Int. J. Data Min. Bioinformatics*, vol. 2, n° 3, p. 193-215, 2008.
- Aronson A. R., « Effective mapping of biomedical text to the UMLS Metathesaurus : the Meta-Map program », *Proceedings AMIA Symposium*, 17-21, 2001.
- Aronson A. R., Mork J., Mork J. G., Gay C., Gay C. W., Humphrey S., Humphrey S. M., Rogers W., Rogers W. J., « The NLM Indexing Initiative's Medical Text Indexer », *In Proceedings of the 11th World Congress on Medical Informatics Demner-Fushman and Lin Answering Clinical Questions (MEDINFO 2004)*, p. 268-272, 2004.

- Avillach P., Joubert M., Fieschi M., « A Model for Indexing Medical Documents Combining Statistical and Symbolic Knowledge », *Proc. AMIA Symp.*, 2007.
- Cai L., Hofmann T., « Hierarchical document categorization with support vector machines », *Proc. CIKM'04*, p. 78-87, 2004.
- Dinh D., Tamine L., « Vers un modèle d'indexation sémantique adapté aux dossiers médicaux de patients », *CORIA 2010, à paraître*, 2010.
- Gale W. A., Church K. W., Yarowsky D., « One sense per discourse », *HLT '91 : Proceedings of the workshop on Speech and Natural Language*, p. 233-237, 1992.
- Gale William ; Church K., Yarowsky D., « A method for disambiguating word senses in a large corpus », *Computers and the Humanities*. 415-439, 1993.
- Gaudan S., Kirsch H., Rebholz-Schuhmann D., « Resolving abbreviations to their senses in Medline », *Bioinformatics*, vol. 21, n° 18, p. 3658-3664, 2005.
- Hersh W., *Information Retrieval : A Health and Biomedical Perspective (Health Informatics)*, 2008.
- Hersh W., Buckley C., Leone T. J., Hickam D., « OHSUMED : an interactive retrieval evaluation and new large test collection for research », *SIGIR'94*, p. 192-201, 1994.
- Joshi M., Pedersen T., Maclin R., « A Comparative Study of Support Vector Machines Applied to the Word Sense Disambiguation Problem for the Medical Domain », *IICAI'05*, p. 3449-3468, 2005.
- Kim W., Aronson A. R., Wilbur W. J., « Automatic MeSH term assignment and quality assessment », *Proc. AMIA Symp.*, 2001.
- Leacock C., Chodorow M., « Combining Local Context and WordNet Similarity for Word Sense Identification », *An Electronic Lexical Database*. 265-283, 1998.
- Lee Y. K., Ng H. T., Chia T. K., « Supervised Word Sense Disambiguation with Support Vector Machines and multiple knowledge sources », *Senseval-3 : Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, p. 137-140, 2004.
- Leroy G., *et al.*, « Effects of information and machine learning algorithms on word sense disambiguation with small datasets », *Medical Informatics*. 573-585, 2005.
- Lesk M., « Automatic sense disambiguation using machine readable dictionaries : how to tell a pine cone from an ice cream cone », *SIGDOC '86*, p. 24-26, 1986.
- Liu H., Teller V., Friedman C., « A Multi-aspect Comparison Study of Supervised Word Sense Disambiguation. », *J Am Med Inform Assoc*, vol. 11, n° 4, p. 320-31, 2004.
- Mihalcea R., « Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling », *HLT'05*, p. 411-418, 2005.
- Mohammad S., Pedersen T., « Combining Lexical and Syntactic Features for Supervised Word Sense Disambiguation », *CoNLL'04*, p. 25-32, 2004.
- Névéol A., Rogozan A., Darmoni S., « Automatic indexing of online health resources for a French quality controlled gateway », *Inf. Process. Manage.*, vol. 42, n° 3, p. 695-709, 2006.
- Névéol A., Shooshan S. E., Humphrey S. M., *et al.*, « Multiple Approaches to Fine-Grained Indexing of the Biomedical Literature », *Pacific Symp. on Biocomputing*, p. 292-303, 2007.
- Pereira S., Neveol A., *et al.*, « Using multi-terminology indexing for the assignment of MeSH descriptors to health resources in a French online catalogue », *Proc. AMIA Symp.*, 2008.

- Robertson S. E., Walker S., Hancock-Beaulieu M., « Okapi at TREC-7 : Automatic Ad Hoc, Filtering, VLC and Interactive », *TREC*, p. 199-210, 1998.
- Schiemann T., Leser U., *et al.*, « Word Sense Disambiguation in Biomedical Applications : A Machine Learning Approach », *Information Retrieval In Biomedicine*, p. 142-161, 2008.
- Schmid H., « Part-of-speech tagging with neural networks », *Proceedings of the 15th conference on Computational linguistics*, p. 172-176, 1994.
- Stevenson M., Guo Y., Gaizauskas R., Martinez D., « Knowledge sources for word sense disambiguation of biomedical text », *BioNLP'08*, p. 80-87, 2008.
- Trieschnigg D., Pezik P., Lee V., de Jong F., Kraaij W., Rebholz-Schuhmann D., « MeSH Up : effective MeSH text classification for improved document retrieval », *Bioinformatics*, vol. 25, n° 11, p. 1412-1418, June, 2009.
- Widdows D., Peters S., *et al.*, « Unsupervised Monolingual and Bilingual Word-Sense Disambiguation of Medical Documents using UMLS », *ACL'03 Workshop*, p. 9-16, 2003.
- Yarowsky D., « One sense per collocation », *HLT'93*, Association for Computational Linguistics, Morristown, NJ, USA, p. 266-271, 1993.
- Yarowsky D., « Unsupervised Word Sense Disambiguation Rivaling Supervised Methods », *ACL'95*, p. 189-196, 1995.