Indexation sémantique de documents textuels

Fatiha BOUBEKEUR
Université Mouloud Mammeri, Algérie

Wassila AZZOUG
Université M'hamed Bouguerra, Algérie

Sarah CHIOUT
Université Mouloud Mammeri, Algérie

Mohand BOUGHANEM
IRIT-SIG, Université Paul Sabatier de Toulouse

Résumé : Ce papier décrit une approche d'indexation sémantique des documents. Nous proposons d'utiliser WordNet comme ressource linguistique afin de retrouver les concepts représentatifs du contenu d'un document. Notre contribution porte sur trois aspects: nous proposons (1) une approche d'identification des concepts en utilisant la base lexicographique WordNet, (2) une approche de désambiguïsation à deux niveaux, basée sur l'utilisation conjointe de WordNetDomains et de WordNet, et (3) une approche de pondération des concepts basée sur une nouvelle notion d'importance.

Mots-clés : Recherche d'information, indexation sémantique, désambiguïsation, WordNet, WordNetDomains.

Abstract : This paper describes a document semantic indexing approach. We propose to use WordNet as linguistic resource for retrieving the representative concepts of a document. Our contribution in this paper is threefold: we propose (1) an approach for identifying concepts using WordNet lexical database, (2) a disambiguation approach based on the joint use of WordNet and WordNetDomains, and (3) a concept weighting approach based on a novel definition of concept importance.

Keywords : Information retrieval, semantic indexing, disambiguation, WordNet, WordNetDomains.

Introduction et problématique

Un processus de recherche d'information (RI) a pour but de sélectionner l'information pertinente pour un besoin en information exprimé par l’utilisateur sous forme de requête. Une étape clé dans ce processus de RI, est l’indexation. L’indexation consiste à représenter requêtes et documents par un ensemble de termes (généralement des mots simples) pondérés, sensés définir au mieux leurs contenus sémantiques. Les termes sont automatiquement extraits ou manuellement assignés aux documents et aux requêtes, puis pondérés par des valeurs numériques qui traduisent leur importance dans le document. De la qualité de l’indexation dépend en grande partie la qualité de la recherche.

Un facteur clé impactant la qualité de l’indexation concerne la capacité du système à traiter avec l’ambiguïté de la langue naturelle et à comprendre les sens des mots dans les documents. Il ne s’agit plus alors de représenter le document par de simples chaines de caractères (entités lexicales), mais bien par des entités véhiculant des sens (entités sémantiques): les concepts. L’indexation sémantique, se base sur les concepts plutôt que sur les mots pour indexer les documents. Pour ce faire, les approches d’indexation sémantiques se basent globalement sur trois étapes : (1) une première étape d’identification des termes à l’issue de laquelle les mots (simples ou composés) contenus dans le document sont identifiés. Cette étape se base sur des techniques linguistiques classiques (tokénisation, lemmatisation, élimination de mots vides) et sur quelques techniques plus avancées d’identification des collocations de mots, (2) une seconde étape de désambiguïsation des sens des mots qui a pour objet de retrouver le sens correct d’un mot dans un contexte donné. Pour ce faire, les approches de désambiguïsation s’appuient sur des ressources linguistiques telles que les corpus d’apprentissage [5], [8], [12], dictionnaires automatisés [6], [9], [18], [20], ou encore les ontologies [14], [16], et autres Wikipédia [11], qui constituent des sources d’évidence pour les définitions et sens d’un mot. Le principe de la désambiguïsation consiste en général à associer un score de désambiguïsation aux différents sens possibles d’un mot (fournis par les dictionnaires et autres ressources …). La précision de la désambiguïsation dépend non seulement de la ressource linguistique utilisée mais aussi en grande partie du score de désambiguïsation établi. (3) Dans la troisième étape, il s’agit de pondérer les concepts identifiés à l’issue de l’étape précédente. La pondération a pour objet d’associer à chaque concept un poids numérique représentant son degré d’importance dans le document. La pondération est un problème crucial en RI. La qualité de la recherche dépend de la qualité de la pondération adoptée.

Ce papier présente la formalisation d’une approche d'indexation sémantique de documents. Dans cette approche, nous proposons d’utiliser WordNet [13] et son extension WordNetDomains comme

source d'évidence pour l'identification des sens des mots et pour leur pondération. Les mots sont alors désambiguïsés par rapport à leurs domaines associés dans WordNetDomains. La pondération d'un concept s'appuie sur une notion revisitée de l'importance d'un concept.
Le papier est structuré comme suit : Après une introduction, nous présentons en section 1 une synthèse des travaux dans le domaine, puis nous situons notre contribution. En section 2, nous donnons quelques notions préliminaires sur WordNet et WordNetDomains, puis des définitions utilisées dans la suite du papier. Notre approche d'indexation sémantique est détaillée en section 3. La section 4 présente une illustration. La section 5 conclut le papier.

1. Etat des lieux de l'indexation conceptuelle

L'indexation conceptuelle représente les documents par des concepts. Ces concepts sont extraits d'ontologies et autres ressources linguistiques. Pour ce faire, le processus d'indexation s'appuie en générale sur deux étapes : l'identification des concepts et leur pondération. Le processus clé dans l'étape d'identification des concepts concerne la désambiguïsation des sens des mots. De nombreuses approches existantes se basent sur WordNet comme source d'évidence pour la désambiguïsation. C'est ainsi que pour désambiguïser un mot ambigu, Voorhees [19] classe chaque synset (sens correspondant à une entrée de WordNet) de ce mot en se basant sur le nombre de mots co-occurrents entre un voisinage (Voorhees l'a appelé *hood*) de ce synset et le contexte local (la phrase où l'occurrence du mot apparaît) du mot ambigu correspondant. Le synset le mieux classé est alors considéré comme sens adéquat de l'occurrence analysée du mot ambigu. Les concepts sont ensuite pondérés en utilisant un schéma de pondération classique *tf*idf* normalisé. Dans une approche différente, Katz et al [17] proposent aussi une approche basée sur le contexte local. Le contexte local d'un mot est défini comme étant la liste ordonnée des mots démarrant du mot utile le plus proche du voisinage gauche ou droit jusqu'au mot cible. L'hypothèse de Katz et al. est que des mots utilisés dans le même contexte local (appelés *sélecteurs*) ont souvent des sens proches. Les sélecteurs des mots d'entrée sont extraits des contextes locaux gauche et droit, puis l'ensemble *S* de tous les sélecteurs obtenus est comparé avec les synsets de WordNet. Le synset qui a le plus de mots en commun avec *S* est sélectionné comme sens adéquat du mot cible. Ce principe est repris dans l'approche d'indexation de Baziz et al. [1]. Les auteurs considèrent ainsi que parmi les différents sens possibles (concepts candidats) d'un terme donné, le plus adéquat est celui qui a le plus de liens sémantiques [9], [10], [15] avec les autres concepts du même document. L'approche consiste à affecter un score à chaque concept

candidat d'un terme d'indexation donné. Le score d'un concept candidat est obtenu en sommant les valeurs de similarité qu'il a avec les autres concepts candidats correspondant aux différents sens des autres termes du document. Le concept candidat ayant le plus haut score est alors retenu comme sens adéquat du terme d'indexation associé. Les concepts sont ensuite pondérés sur la base d'un schéma de pondération dit *Cf*idf*, qui étend la pondération *tf*idf* pour tenir compte des termes composés. L'approche de Baziz et al. a été reprise dans Boughanem et al. [4], avec une nouvelle définition de la pondération. En effet, dans [4], les auteurs introduisent les notions de centralité et de spécificité d'un concept. La centralité définit le nombre de relations de ce concept avec les autres concepts du document. Sa spécificité définit son degré de « spécialité ». Le schéma de pondération utilisé est basé sur la combinaison de ces deux mesures. Dans notre approche d'indexation sémantique proposée dans [2], [3], le choix du concept correct dans un contexte s'appuie sur un score basé sur la somme des valeurs de similarité que le concept cible a avec les concepts les plus fréquents dans le document. Les concepts sont alors pondérés sur la base d'une mesure de leur importance dans le document, quantifiée au travers de leur proximité sémantique aux autres concepts du document. Dans une approche plus récente [7] les auteurs proposent une approche intéressante de désambiguïsation à deux niveaux : d'abord retrouver le domaine correct d'un mot dans le document, puis désambiguïser ce mot dans le domaine ainsi identifié. Le domaine correct d'un mot est celui qui maximise ses occurrences dans le contexte local du mot cible. Les auteurs utilisent WordNetDomains, qui permet de classifier les différentes entrées de WordNet dans des domaines prédéfinis.

Positionnement de notre proposition

Notre approche proposée dans ce papier tente de combiner notre approche d'indexation conceptuelle dans [2], [3] et l'approche de désambiguïsation par les domaines au sein d'un paradigme unifié. L'objectif est de représenter de manière précise le document par un noyau sémantique composé de concepts pondérés. Dans notre proposition, les termes d'indexation sont d'abord extraits en se basant sur des étapes d'indexation classiques. Cette étape inclut en outre une nouvelle proposition pour la détection des collocations à partir d'une liste pré-établie des collocations de WordNet; A l'issue de cette étape, trois listes sont construites : la liste des collocations, la liste des mots simples ayant des entrées correspondantes dans WordNet, et la liste des mots simples n'ayant pas d'entrée dans WordNet (ces mots seront dits des mots orphelins). Puis chaque mot non vide identifié dans WordNet est désambiguïsé dans son contexte global dans le document. La désambiguïsation d'un mot se base d'abord sur sa désambiguïsation de

domaine (ie. trouver le domaine correcte du mot dans le document), puis sa désambiguïsation sémantique dans le domaine choisi.
- La désambiguïsation du domaine se base sur une mesure de probabilité d'appartenance du mot cible et des mots du contexte à un domaine donné. Le domaine qui maximise cette mesure est retenu comme domaine correct du terme dans le document.
- La désambiguïsation des concepts est basée sur l'intuition que les concepts appartenant à un même domaine sont fortement (sémantiquement) liés et doivent renforcer la désambiguïsation sémantique du concept cible par rapport à des concepts qui ne sont pas dans le même domaine que lui.
L'approche de pondération des concepts est basée sur une nouvelle définition de la notion de centralité d'un concept. La centralité dépend de la pertinence d'un concept dans le document et sur sa fréquence. Ces notions seront définies en section suivante.
Dans ce qui suit, nous décrivons les différentes étapes de notre approche d'indexation sémantique. Mais d'abord, nous présentons quelques notions préliminaires sur WordNet, ainsi que des définitions et notations utilisées dans la suit de notre papier.

2. Préliminaires

2.1. WordNet

WordNet est un réseau lexical électronique qui couvre la majorité des noms, verbes, adjectifs et adverbes de la langue Anglaise, qu'il structure en un réseau de noeuds et de liens.
* Les noeuds sont constitués par des ensembles de termes synonymes appelés synsets.
- Un synset représente un concept.
- Un concept est une entité sémantique, lexicalement représentée par un terme.
- Un terme peut être un mot simple ou une collocation de mots
* Les liens représentent des relations sémantiques entre concepts, dont par exemple les relations d'hyponymie-hyperonymie suivantes :
- la relation de subsumption entre noms, (relation is-a) qui permet d'associer un concept classe (l'hypernyme) à un concept sous-classe (l'hyponyme). Par exemple, le nom tower#1 a pour hyponymes silo, minaret, pylon…Cette relation permet d'organiser les concepts de WordNet en une hiérarchie.
- la relation d'instanciation (instance) qui permet d'associer un concept et son instance. Par exemple, le nom tower#1 a pour instance hyponyme tour Eiffel.
Un exemple de hiérarchie de synsets correspondant au nom « bank » est donné en Table 1.

The noun bank has 10 senses (first 9 from tagged texts)

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")
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")
3. (76) bank -- (a supply or stock held in reserve for future use (especially in emergencies))
4. (54) bank, bank building -- (a building in which the business of banking transacted; "the bank is on the corner of Nassau and Witherspoon")
5. (7) bank -- (an arrangement of similar objects in a row or in tiers; "he operated a bank of switches")
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")
7. (3) bank -- (a long ridge or pile; "a huge bank of earth")
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")
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)
10. bank -- (a flight maneuver; aircraft tips laterally about its longitudinal axis (especially in turning); "the plane went into a steep bank")

Table 1 : Les concepts de WordNet correspondants au mot bank

WordNetDomains est une extension de WordNet dans laquelle les synsets ont été étiquetés par des labels de domaines. Un exemple de domaines associés au synsets du mot *bank* est donné en Table 2.
Ces domaines sont organisés selon une hiérarchie définissant la relation de spécialisation/généralisation entre les domaines. Par exemple, le domaine *Tennis* est plus spécifique que le domaine *Sport*, et le domaine *Architecture* est plus général que le domaine *Buildings*. Une partie de la hiérarchie de WordNetDomains est donnée en Table3. Le domaine *Top-Level* représente la racine de cette hiérarchie. Le domaine *Factotum* de WordNetDomains est un domaine fonctionnel (par opposition à sémantique) qui regroupe tous les sens des mots qui n'appartiennent à aucun domaine particulier mais qui peuvent apparaître avec des termes associés à d'autres domaines. *Factotum* constitue un domaine particulier, indépendant du domaine *Top-Level* et de sa hiérarchie.

Les senses	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")	ECONOMY, 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)	ARCHITECT-URE
10. bank -- (a flight maneuver; aircraft tips laterally about its longitudinal axis (especially in turning); "the plane went into a steep bank")	TRANSPORT

Table 2 : Les domaines associés dans WordNetDomains, aux synsets du mot bank

2.2. Définitions et notations

Soit m_i un mot d'un texte à analyser.

a. On appelle occurrence de m_i, toute instance de m_i dans le texte.

b. Une instance de m_i apparaît dans une seule phrase. L'ensemble des mots de cette phrase constitue son contexte local. On note le contexte local de l'instance m_i par ζ_i

c. On appelle contexte local droit de l'instance m_i, l'ensemble des mots à droite de m_i jusqu'à la prochaine ponctuation. Le contexte local droit de l'instance m_i sera noté : ζ_{di}

d. On appelle expression locale de l'instance m_i, la chaîne de caractères concaténant, par le biais du souligné (_), le mot m_i avec les mots de contexte local droit successivement. La taille d'une expression locale est le nombre de mots qui la composent.

e. On appelle contexte global du mot m_i, l'union de tous ses contextes locaux dans le texte du document. Le contexte global du mot sera noté par : ζ_{Gi}

Top_level	Humanities	History		
		Linguistics	Grammar	
		Literature	Philology	
		Philosophy	Psychoanalysis	
		Art	Music	
			Plastic_Arts	Jewellery
				Sculpture
			Theatre	
			Cinema	
		Paranormal		
	Free_Time	Play		
		Sport		
			Tennis	
		...		
	Applied_Science	Agriculture		
		Architecture	Buildings	
	Pure_Science			
		Biology	Anatomy	
		Animals		
		Earth	Geology	
			Geography	
		Mathematics		
		Physics	Acoustics	
		...		
	Social_Science	Economy	Finance	Money
		Politics		
		Fashion		
		Military		
Factotum	Quality			
	Number			
	...			

Table 3 : Extrait de la hiérarchie de WordNetDomains

3. Indexation sémantique des documents

L’indexation sémantique vise à représenter un document par un ensemble de concepts pondérés qui décrivent au mieux son contenu.
Le processus d'indexation du document s’effectue en trois étapes : (1) l’identification des termes d’index, (2) la désambiguïsation des termes d’index et (3) la pondération des concepts.

3.1. Identification des termes d’index

Le but de cette étape est d'identifier :

1. l’ensemble ξ_{Expres} des expressions du document, correspondant aux collocations de WordNet.

2. l’ensemble $\xi_{Simples}$ des mots simples ayant une entrée dans WordNet,

3. l’ensemble ξ_{Orphel} des mots orphelins (mots simples n’ayant pas d’entrée dans WordNet).

Cette étape débute par l’identification des expressions. Pour cela, nous avons d’abord construit la liste φ_{Coloc} de toutes les collocations

existantes dans WordNet. Puis, pour une occurrence de mot à analyser, on extrait de φ_{Coloc} l'ensemble ς_i de toutes les collocations qui commencent par m_i. On ordonne ς_i par tailles décroissantes de ses éléments, puis on projette chaque élément de ς_i sur des expressions locales E_i de m_i. Si une expression locale s'apparie avec une collocation, elle est retenue comme expression et insérée dans l'ensemble ξ_{Expres}. Si aucune collocation de ς_i ne s'apparie avec une expression locale de m_i, alors m_i est un mot simple. Si m_i possède une entrée dans WordNet, il sera inséré dans l'ensemble $\xi_{Simples}$; Sinon il sera mis dans l'ensemble des orphelins ξ_{Orphel}.

Le principe de l'identification des termes est décrit à travers l'algorithme de la Table 4.

Algorithme de détection des termes
Entrée : document d.
Sortie : ξ_{Expres}, $\xi_{Simples}$, ξ_{Orphel}
Procédure : Soit m_i le prochain mot à analyser dans d.
Début
1. Calculer $\varsigma_i = \{C^i_1, C^i_2, \ldots C^i_n\}$ l'ensemble 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. Pour chaque $C^i_{(j)}$ dans ς_i et bool=faux, faire :
5. Calculer l'expression locale E_i de taille $|C^i_{(j)}|$
6. Si $E_i = C_{(j)}$ alors bool <- vrai ; fait ;
7. Si (bool) mettre E_i dans ξ_{Expres}
   Sinon Si m_i mot non vide alors
   Début Si m_i possède une entrée dans WordNet alors mettre m_i dans $\xi_{Simples}$
   Sinon mettre m_i dans ξ_{Orphel} . fin ;

Fin.

Table 4: Algorithme de détection des termes d'index

3.2. Désambiguïsation des termes

Les collocations étant des expressions quasiment désambiguïsées, l'étape d'indexation concernera uniquement les mots simples ayant des entrées dans WordNet, soit donc l'ensemble des termes de $\xi_{Simples}$.

Chaque terme de $\xi_{Simples}$ peut avoir plusieurs sens possibles. Le but de cette étape est de sélectionner le meilleur sens du terme dans le document. L'approche de désambiguïsation proposée est une approche à trois niveaux :
(1) dans le premier niveau, il s'agit de déterminer la forme grammaticale (nom, verbe, …) du mot m_i dans le document, en utilisant le Stanford POS Tagger.
(2) le second niveau, permet d'identifier le domaine d'usage du mot dans le document. L'identification des domaines s'appuie sur l'utilisation de *WordNetDomains*. Ce niveau de désambiguïsation permettra de limiter le nombre de sens du terme qui seront examinés dans le niveau suivant de désambiguïsation. (3) le troisième niveau de désambiguïsation consiste alors à sélectionner parmi les sens possibles du terme dans le domaine sélectionné, celui qui est sensé le définir au mieux dans le document.

3.2.1. Identification de la forme grammaticale des mots simples
Les sens d'un mot m_i dans WordNet sont classés selon ses différentes catégories grammaticales possibles. Ainsi, nous utilisons le Stanford POS Tagger pour identifier la catégorie grammaticale du mot m_i dans le document afin de déterminer les sens appartenant à cette forme grammaticale. Cette étape permet de limiter le nombre de sens du terme qui seront utilisés dans la désambiguïsation et de récupérer les domaines qui correspondent à ses sens dans WordNetDomains.

3.2.2. Désambiguïsation au niveau des domaines
Chaque mot dans $\xi_{Simples}$ possède plusieurs sens dans WordNet. Les sens de WordNet sont étiquetés dans *WordNetDomains* par des labels de domaines. Ainsi, un sens peut appartenir à un ou plusieurs domaines.
On note :

S_i l'ensemble de tous les synsets associés au mot m_i,

D l'ensemble, non redondant, de tous les domaines associés aux éléments de S_i,

$S_{i(j)}$ est l'ensemble des synsets de S_i appartenant au domaine D_j,

$S_{i(j)}[k]$ le kième élément de l'ensemble $S_{i(j)}$

Partant de l'hypothèse que le domaine probable d'un mot est celui qui maximise sa similarité avec les autres domaines des autres mots du même contexte, nous attribuons à chaque domaine D_j associé à un sens du mot m_i, un score basé sur la somme de ses similarités avec les différents

domaines associés aux sens des autres termes t_k ($\left(t_k \in \left\{ \xi_{Simples} \cup \xi_{Expres} \right\} \right)$) d'index. Le domaine D_j ayant le plus grand score est sélectionné comme domaine adéquat pour le mot m_i dans le document.

Formellement : $Score(D_j) = \arg\max_j \left(\sum_{t_k \in \xi_{Gi}} \sum_{k \in [1..n]} Sim(Dj, D_k) \right)$

Où :

$Sim(Dj, D_k)$ désigne la similarité entre les domaines D_j et D_k.

Pour mesurer la similarité entre les domaines D_j et D_k, nous utilisons et adaptons la formule de Wu-Palmer [21] à la hiérarchie Top-Level de WordNetDomains, ce qui donne :

$$Sim(D_j, D_k) = \frac{2 * prof\ ondeur(D^*)}{prof\ ondeur(D_j) + prof\ ondeur(D_k)}$$

Où :

D^* : est le domaine le plus spécifique qui subsume D_j et D_k dans la hiérarchie de WordNetDomains.

$prof\ ondeur(D^*)$: est le nombre d'arcs entre la racine de WordNetDomains et le domaine D^*.

$prof\ ondeur(D_j)$: est le nombre d'arcs entre la racine de WordNetDomains et le domaine D_j en passant par le domaine D^*.

Remarque :
La formule de similarité est appliquée aux seuls domaines de la hiérarchie Top-Level. Le domaine factotum, indépendant de cette hiérarchie est un domaine fonctionnel (non sémantiquement informatif). Il ne sera pas considéré dans cette désambiguïsation.

3.2.3. Désambiguïsation des sens des mots

A l'issue de l'étape précédente, tout mot m_i de $\xi_{Simples}$ est associé à un seul domaine D_j dans le document. Deux cas peuvent se présenter :

- soit m_i possède un seul sens dans D_j , dans ce cas il est désambiguïsé.

- soit m_i possède plusieurs sens dans D_j, dans ce cas il est ambigu. Il faut le désambiguïser.
Nous proposons une désambiguïsation sur les seuls sens appartenant à D_j, soit donc aux seuls éléments $S_{i(j)}[k]$ de l'ensemble $S_{i(j)}$. L'objectif est alors de sélectionner parmi ces sens le sens correct pour le mot m_i dans le document.

Pour désambiguïser le mot m_i dans son domaine, on associe à chacun de ses sens $S_{i(j)}[k]$ de l'ensemble $S_{i(j)}$, un score basé sur sa proximité sémantique avec les autres sens associés aux mots de son contexte global dans leurs domaines respectifs. Le concept $S_{i(j)}[k]$ ayant le plus grand score est alors retenu comme sens adéquat pour le mot m_i dans *d*.
Formellement :

$$S_{i(j)}[k] = Arg\max\left(\sum_{\substack{l \,/\, m_l \in \xi_{G_i} \\ l \neq i}} \sum_{1 \leq n \leq \left|S_{l(m)}\right|_l} sim\left(S_{i(j)}[k], S_{l(m)}[n]\right)\right)$$

Où $sim\left(S_{i(j)}[k], S_{l(m)}[n]\right)$ est la similarité sémantique entre les concepts $S_{i(j)}[k]$ et $S_{l(m)}[n]$.
L'ensemble des concepts retenus constituera le noyau sémantique *N(d)* du document *d*. Pour des raisons de simplification, on utilisera la notation C^i pour désigner le ième élément de *N(d)*.

3.3. Pondération des concepts

Partant de l'idée qu'un concept est d'autant plus représentatif du contenu du document qu'il est fréquent et pertinent dans ce document, nous proposons de pondérer un concept avec un poids basé sur :
Sa pertinence, que nous définissons sa proximité sémantique aux autres concepts du document,
Sa fréquence dans le document.
Formellement, le poids $W(C^i)$ du concept C^i est défini par :

$$W(C^i) = \alpha * tf(C^i) + (1-\alpha)\sum_{i \neq l} Dist(C^i, C^l)$$

Où α est un facteur de pondération qui permet de balancer la fréquence par rapport à la pertinence. Ce facteur pourra être fixé expérimentalement.
Le schéma de pondération que nous proposons permet outre la pondération des concepts, la pondération des collocations et des termes

orphelins. Dans ce dernier cas, seule la fréquence est considérée, les proximités sémantiques inexistantes, sont initialisées à zéro.
Le noyau sémantique de *d* est alors construit en gardant seulement les concepts dont les poids sont plus grands qu'un seuil fixé. Nous proposons, dans un premier temps, de garder tous les concepts dont le poids est différent de zéro.

4. Illustration

Dans le paragraphe suivant, nous montrons la faisabilité de notre approche d'indexation sémantique en l'appliquant sur un exemple. Nous focalisons en particulier sur la désambiguïsation puisque de sa précision dépend en grande partie la précision de l'indexation.
Etant donné le texte suivant (extrait du document *Arthroskopie.00130003.eng.abstr* de la collection Muchmore[1])
"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. The posterior cruciate ligament consists of multiple small fiber bundles."

4.1. Détection des concepts

Notre algorithme de détection des concepts retourne les trois ensembles : ξ_{Expres}, $\xi_{Simples}$, ξ_{Orphel} suivants :

$$\xi_{Expres} = \{\text{human_knee, knee_joint, fiber_bundle}\}$$

$$\xi_{Orphel} = \{\text{PCL, intercondylar}\}$$

$$\xi_{Simples} = \begin{cases} \text{posterior, cruciate, ligament, strong, origin, lateral, wall, medial, fem} \\ \text{condyle, insertion, locate, part, area, consist, multiple, small} \end{cases}$$

4.2. Désambiguïsation des termes simples

4.2.1. Identification de la forme grammaticale
En utilisant le Stanford POS Tagger, on obtient la forme grammaticale de chaque mot simple dans le document comme suit :

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

4.2.2. Désambiguïsation au niveau des domaines
Nous retrouvons pour chaque terme d'index l'ensemble des domaines associés à ses différents sens. Puis nous désambiguïsons au niveau des

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

domaines. La désambiguïsation au niveau des domaines permet d’associer les domaines adéquats aux termes d’index. Les numéros des sens et les domaines associés aux termes d’index sont présentés dans les tableaux de la figure 1 suivante. Les résultats de la désambiguïsation des domaines sont récapitulés dans les tableaux de la figure 2.

Posterior	00136931 : animals	Wall	04369872 : buildings	Consist	02574255 : factotum
cruciate	02290409 : factotum		03899176 : military		02670036 : factotum
ligament	04990644 : anatomy		08876684 : factotum		02578919 :factotum
	03528343 : factotum		04370607 :factotum		02554853 : factotum
strong	02239657 : quality		05284169 :anatomy	multiple	02140712 : factotum
	02238035 : factotum		08876934 :geology	Small	01343705 :factotum
	01766728 : factotum		04370836 :factotum		01366545 : factotum
	01764165 : quality		137331862 :factotum		02259250 : factotum
	02436955 : factotum		13731862 : factotum		01597253 : factotum
	02193170 : factotum	Femoral	02614670 : anatomy		01419050 : factotum
	01901958 : Grammar	Human knee	05254826 :anatomy		01481659 : factotum
	01118836 : factotum	Insertion	06308340 :factotum		01406192 : acoustics
	01023736 : factotum		00306609 :factotum		01503651 : factotum
	00807924 : factotum	Locate	02220173 : factotum		00845424 : factotum
Origin	07989929 :factotum		02613751 :factotum		01416987 : factotum
	04675241 :biology		02266215 :factotum		02155033 : factotum
	06874664 : factotum		00401714 :politics	condyle	05157880 : anatomy
	05654262: mathematics	area	07980485 :geography	Knee joint	05254826 : anatomy
	07610417 : factotum		13687487 :factotum	Fiber bundle	05161310 : anatomy
lateral	02353094 : factotum		05646624 : factotum		
	00746517 :factotum		02641332 : factotum		
medial	00746822 :factotum		04921220 :anatomy		
	00326753 :factotum		04842376 : factotum		

Figure 1 *: Sens et domaines associés aux termes d’index*

4.2.3. Désambiguïsation des sens des mots

A l’issue de l’étape précédente, chaque terme d’index est associé aux seuls sens liés au domaine sélectionné. Seuls ces sens sont désambiguïsés. Pour le calcul du score de désambiguïsation des sens des mots, nous nous basons sur la mesure de similarité de Lesk. Dans les tableaux de la figure 2 suivante, nous représentons pour chaque terme d’index, son domaine sélectionné ainsi que ses sens associés dans ce domaine. Le sens grisé dans le tableau représente le sens désambiguïsé (sens adéquat du terme dans le document). Un examen rapide des résultats, appuyé par une désambiguïsation manuelle, nous permet de voir que :

- La désambiguïsation au niveau des domaines a permis d’associer les domaines adéquats aux termes d’index. C’est ainsi par exemple que les termes d’index wall, area et ligament se sont vus assigner le domaine anatomy qui est le domaine le plus probable du texte indexé.
- La désambiguïsation au niveau des sens donne aussi des sens corrects dans le document. Pour vérifier cela, nous avons comparé nos résultats avec ceux obtenus par l’approche de Baziz et al. [1] pour le même texte. Les résultats obtenus montrent que nous retrouvons plus de sens

corrects que dans [1]. A titre d'exemple, le mot wall est désambiguïsé par wall#a#1 (définit dans WordNet par : an architectural partition …) dans [1], alors que notre approche nous retourne wall#a#5 (défini par : (anatomy) a layer (a lining or membrane)…) plus proche sémantiquement de la thématique (médicale) du document. Les résultats de notre approche sont encourageants mais doivent néanmoins être vérifiés sur une collection de taille réelle.

Mot simple	Domaine désambiguïsé	Les sens du mot dans le domaine désambiguïsé
posterior	Animals	00136931 : posterior#a#1
cruciate	Factotum	022904409 : cruciate#a#1
ligament	Anatomy	04990644 : ligament#n#1
Strong	Factotum Grammar Quality	02238035 : strong#a#1
		02238035: strong#a#2
		01462766: strong#a#3
		01766728: strong#a#4
		01764165: strong#a#5
		02436955: strong#a#6
		02193170 : strong#a#7
		01901958: strong#a#8
		01118836 : strong#a#9
		01023736 : strong#a#10
		00807924 : strong#a#10
Origin	Factotum Biology	07989929 : origin#n#1
		04675241: origin#n#2
		06874664 : origin#n#3
		07610417 : origin#a#5
		02353094 : lateral#a#1
lateral	anatomy	00746517 : lateral#a#2
wall	Factotum	05284169: wall#n#5
medial	Factotum	00746822 : medial#a#1
		00326753 : medial#a#2
femoral	Anatomy	02614670 : femoral#a#1
condyle	Anatomy	05157880 : condyle#1
insertion	Factotum	06308340 : insertion#n#1
		00306609 : insertion#n#2
multiple	Factotum	02140712 : multiple#a#1
area	Anatomy	04921220 : area#n#5

Mot simple	Domaine désambiguïsé	Les sens du mot dans le domaine désambiguïsé
locate	Factotum	02220173 : locate#v#1
		02613751 : locate#v#2
		02266215 : locate#v#3
		00401714 : locate#v#4
part	Factotum Theatre	13028617 : part#n#1
		08103697 : part#n#2
		05344775 : part#n#3
		03746082 : part#n#4
		05526011 : part#n#5
		00678112 : part#n#6
		08797461 : part#n#7
		12529266 : part#n#9
		00738005 : part#n#10
consist	Factotum	02574255 : consist#v#1
		02670036 : consist#v#2
		02578919 : consist#v#3
		02554853 : consist#v#4
		013433705 : small#a#1
small	Factotum Acoustics	01366545: small#a#2
		0229250 : small#a#3
		01597253 : small#a#4
		01419050: small#a#5
		01481659: small#a#6
		01503651 : small#a#8
		00845424: small#a#9
		01416987 : small#a#10
		02155033 : small#a#11

Figure 2 : Présentation des domaines et des sens désambiguïsés

5. Conclusion

Nous avons présenté dans ce papier, les fondements théoriques d'une nouvelle approche d'indexation sémantique basée sur l'utilisation conjointe de WordNet et de WordNetDomains. Notre contribution porte sur les trois aspects de l'indexation sémantique : la détection des termes d'index, la désambiguïsation des termes et la pondération des concepts. En particulier, nous avons proposé une nouvelle approche de détection des concepts incluant la détection des collocations, une approche de désambiguïsation par les domaines et dans les domaines, et enfin un nouveau schéma de pondération des concepts. Nous avons montré la faisabilité de notre approche en la déroulant sur un exemple. Des travaux sont en cours en vue de sa validation expérimentale.

Bibliographie

M. BAZIZ, M. BOUGHANEM, N. AUSSENAC-GILLES. A Conceptual Indexing Approach based on Document Content Representation. Dans : CoLIS5 : Fifth International Conference on Conceptions of Libraries and Information Science, Glasgow, UK, 4 juin 8 juin 2005. F. Crestani,
I. Ruthven (Eds.), Lecture Notes in Computer Science LNCS Volume 3507/2005, Springer-Verlag, Berlin Heidelberg, p. 171-186.

F. BOUBEKEUR, M. BOUGHANEM, L.TAMINE, M. DAOUD. De l’utilisation de WordNet pour l’indexation conceptuelle des documents. 13 ème Colloque International sur le Document Electronique.16-17 Décembre 2010, INHA, Paris.

F. BOUBEKEUR, M. BOUGHANEM, L.TAMINE, M. DAOUD. Using WordNet for Concept-based document indexing in information retrieval. Dans: Fourth International Conference on Semantic Processing (SEMAPRO 2010), Florence, Italy, Octobre 2010.

M. BOUGHANEM, I. MALLAK, H. PRADE. A new factor for computing the relevance of a document to a query (regular paper). Dans : IEEE World Congress on Computational Intelligence (WCCI 2010), Barcelone, 18/07/2010-23/07/2010, 2010.

M. CUADROS, JM., ATSERIAS, J., M. CASTILLO, M., & G. RIGAU, (2004). Automatic acquisition of sense examples using exretriever. In IBERAMIA Workshop on Lexical Resources and The Web for Word Sense Disambiguation. Puebla, Mexico.

J.A GUTHRIE, L. GUTHRIE, Y. WILKS, H. AIDINEJAD (1991). Subject-dependant cooccurrence and word sense disambiguation. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkley, CA. 146-152.

S. G. KOLTE, S. G. BHIRUD. Word Sense Disambiguation using WordNetDomains. In First International Conference on Emerging Trends in Engineering and Technology. 2008 IEEE DOI 10.1109/ICETET.2008.231

C. LEACOCK, G.A. MILLER, and M. CHODOROW. Using corpus statistics and WordNet relations for sense identification. Comput. Linguist. 24, 1 (Mar. 1998), 147-165.

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.

D. LIN. (1998) An information-theoretic definition of similarity. In Proceedings of 15th International Conference On Machine Learning, 1998.

O. MEDELYAN ; D. MILNE ; C. LEGG ; I.H. WITTEN. Mining meaning from Wikipedia. In International Journal of Human-Computer Studies archive, Volume 67 , Issue 9 (September 2009). Pages: 716-754. Year of Publication: 2009. ISSN: 1071-5819

R. MIHALCEA and D. MOLDOVAN. Semantic indexing using WordNet senses. In Proceedings of ACL Workshop on IR & NLP, Hong Kong, October 2000

G. MILLER (1995) WordNet: A Lexical database for English. Actes de ACM 38, pp. 39-41.

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, 1999, (p. 95-130).

M. SUSSNA. Word sense disambiguation for free-text indexing using a massive semantic network. 2nd International Conference on Information and Knowledge Management (CIKM-1993), 67–74.

O. UZUNER, B. Katz, D. Yuret. Word Sense Disambiguation for Information Retrieval. AAAI/IAAI 1999 : 985

J. VÉRONIS and N. IDE. Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. 13th International Conference on Computational Linguistics (COLING-1990), 2, 389–394. 1990.

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

Y. WILKS & M. STEVENSON. Combining independent knowledge source for word sense disambiguation. Conference « Recent Advances in Natural Language Processing », 1–7.

Z. WU & M. PALMER. Verb semantics and Lexical selection. Proceedings of the 32th Annual Meetings of the Association for Computational Linguistics, pp. 133-138. 1994.