De l'utilisation de WordNet pour l'indexation conceptuelle des documents

Fatiha Boubekeur (1), Mohand Boughanem (2), Lynda Tamine (2), Mariam Daoud (2)

amirouchefatiha@mail.ummto.dz / boughane@irit.fr / Lechani@irit.fr / daoud@irit.fr

(1) Université Mouloud Mammeri, 15000 Tizi Ouzou, Algérie

(2) IRIT-SIG/RFI, Université Paul Sabatier, 31062 Toulouse, France

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 un double aspect: d'une part, nous proposons une approche d'identification des concepts en utilisant la base lexicographique WordNet, d'autre part, nous proposons une approche de pondération de ces concepts basée sur une nouvelle notion d'importance.

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

1 Introduction

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. Ce processus intègre deux principales étapes, l'indexation et l'appariement. L'indexation consiste à représenter requêtes et documents par un ensemble, l'index, de termes (généralement des mots simples) pondérés, sensés 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. L'appariement consiste à « *matcher* » les représentations des requête et documents pour sélectionner les documents qui correspondent au mieux à la requête. Une caractéristique clé des systèmes de recherche d'information (SRI) est que l'appariement est impacté par la qualité de la description du besoin en information et par la qualité de l'indexation.

Une problématique fondamentale en RI est l'imprécision du besoin utilisateur (une requête est habituellement une description vague et incomplète du besoin en information de l'utilisateur) et l'ambiguïté de l'indexation. A l'origine de cette problématique est la disparité et l'ambiguïté de la langue naturelle.

- La disparité de la langue naturelle traduit la propriété qu'ont certains termes à être représentés par différentes chaînes de caractères, et associés aux mêmes

sens ou à des sens liés. C'est ainsi par exemple qu'un document sur *Linux* pourtant pertinent pour une requête sur les *systèmes d'exploitation*, ne sera pas retrouvé si les mots *système* et *exploitation* sont absents de ce document. En RI, la disparité des termes implique un silence documentaire.

- L'ambiguïté est divisée en homonymie et polysémie [13]. L'homonymie traduit la propriété qu'ont certains termes à être représentés par une même chaîne de caractères, et associés à différents sens. *Souris* (le mammifère) *vs souris* (du verbe *sourire*) est un exemple d'homonymie. La polysémie est liée à la propriété qu'ont certains termes à exprimer différents sens. *Prendre le large vs prendre un thé* est un exemple de polysémie. Dans les systèmes de recherche d'information (SRI) classiques, l'ambiguïté implique que des documents non pertinents sont retrouvés. Ainsi, un document qui traite de la politique en *France* sera retrouvé comme pertinent pour une requête portant sur *Anatole France* si le mot *France* figure dans le document et dans la requête. L'ambiguïté des termes implique un bruit documentaire.

Les SRI classiques présentent ainsi des insuffisances du fait de leur incapacité à traiter avec l'ambiguïté de la langue et l'imprécision sémantique des mots simples. Pour lever ces problèmes d'ambiguïté et de disparité des mots, de nombreux travaux de recherche en RI se sont orientés vers la prise en compte des sens des mots dans le processus d'indexation. L'indexation sémantique, se base sur les sens des mots (entités sémantiques) plutôt que sur les mots simples (entités lexicales) pour indexer les documents. Pour retrouver les sens des mots dans un contenu donné, l'indexation sémantique se base sur des techniques de désambiguïsation contextuelle des mots dans les documents et requêtes. La désambiguïsation a pour objet de retrouver le sens d'un mot dans un contenu donné. Pour ce faire, la désambiguïsation s'appuie :

(1) sur des corpus d'apprentissage [7], [14], [18] : Une manière d'indexer serait par exemple, d'associer aux mots extraits, des mots du contexte qui aident à déterminer leur sens [31]. Une autre manière serait d'apprendre le sens d'un mot à partir de ses usages possibles du mot à désambiguïser [23], ou à partir de règles d'agencement ou règles de fonctionnement des mots à désambiguïser [28].

(2) sur des ressources linguistiques externes telles que les thésaurus [30], dictionnaires automatisés [10], [15], [26], [29], ontologies [20], [24], et autres Wikipédia [17], qui constituent des sources d'évidence pour les définitions et sens du mot cible. On parle alors d'indexation conceptuelle.

Nous proposons, dans ce papier, une approche d'indexation conceptuelle de documents. Le principe de l'approche consiste à extraire les mots du document, puis à leur associer les sens adéquats correspondants. Nous proposons d'utiliser WordNet [19] comme source d'évidence pour l'identification des sens des mots et pour leur pondération. Les sens des mots correspondent alors à des concepts (ou synsets) de WordNet. L'identification des sens des mots se base sur le calcul d'un score de désambiguïsation. La pondération d'un concept s'appuie sur ses relations sémantiques avec les autres concepts du document en tenant compte de leurs importances.

Le papier est structuré comme suit : en section 1, nous posons la problématique de l'indexation conceptuelle, puis présentons une synthèse des travaux dans le domaine, puis nous situons notre contribution. En section 2 nous présentons notre approche d'indexation conceptuelle. La section 3 présente quelques résultats expérimentaux. La section 4 conclut le papier.

2 Contexte des travaux et problématique

2.1 Problématique

La plupart des approches d'indexation conceptuelle s'appuient en général sur des ontologies pour déterminer les différents sens du mot mais aussi pour désambiguïser les sens des mots. L'indexation conceptuelle se base sur des concepts extraits d'ontologies, de taxonomies, et autres ressources lexicales pour indexer les documents contrairement aux listes de mots simples. Pour ce faire, l'indexation conceptuelle soulève deux principaux problèmes: l'identification des concepts et leur pondération.

(1) L'identification des concepts a pour objectif d'extraire l'ensemble des mots ou collocations de mots du document à indexer, et à leur associer leurs sens correspondant dans le document. L'extraction des mots simples est un problème d'indexation classique. Les approches utilisées se basent le plus souvent sur des techniques linguistiques (tokénisation, lemmatisation, élimination de mots vides) et statistiques pour identifier les mots clés du document. *Etant donnés ces mots clés, le problème crucial de l'indexation sémantique est d'abord d'identifier, pour chacun de ces mots clés, les entrées correspondantes dans l'ontologie, puis de sélectionner parmi ces entrées le sens adéquat du mot clé considéré dans le document : c'est la désambiguïsation des sens des mots (WSD*[1]*).*

(2) La pondération des termes d'indexation a pour objet d'associer à chaque terme d'index son poids 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. Dans les SRI classiques basés sur une indexation par mots clés, la pondération dite *tf*idf* et ses variantes sont largement adoptées. En indexation conceptuelle, le problème est alors de définir une pondération adéquate pour les entités sémantiques que sont les concepts.

*Une fois les termes d'i*ndexation désambiguïsés et pondérés, la représentation des textes indexés se fait soit à partir des seuls sens (ou concepts) identifiés lors de l'étape de désambiguïsation, soit à partir d'une combinaison des mots-clés et sens corrects associés. *Les approches d'indexa*tion *[1], [2], [12], [18], [25], [27] sont basées sur ce principe.* Nos travaux se situent dans ce même contexte, et consistent en l'utilisation de WordNet tant pour l'identification des concepts que pour leur pondération. Dans ce qui suit, nous présentons la base lexicographique WordNet, puis quelques travaux d'indexation conceptuelle principalement basés sur cette ressource. Nous situons enfin notre contribution par rapport à ces derniers.

2.2 Préliminaires: WordNet

WordNet est un réseau lexical électronique qui couvre la majorité des noms, verbes, adjectifs et adverbes de la langue Anglaise, qu'elle 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.

[1] *Word Sense Disambiguation*

- Un terme peut être un mot simple ou une collocation (mot composé).

- 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 mot « dog » est donné dans la table 1.

Noun
UUUUS : (n) 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"*
S : (n) frump, dog (a dull unattractive unpleasant girl or woman) *"she got a reputation as a frump"; "she's a real dog"*
S : (n) dog (informal term for a man) *"you lucky dog"*
S : (n) cad, bounder, blackguard, dog, hound, heel (someone who is morally reprehensible) *"you dirty dog"*
S : (n) 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)
S : (n) 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)
S : (n) andiron, firedog, dog, dog-iron (metal supports for logs in a fireplace) *"the andirons were too hot to touch"*
Verb
S : (v) 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"*

Table1 : Les concepts de WordNet correspondants au concept dog

2.3 Synthèse des travaux sur 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 : (1) l'identification des concepts et (2) leur pondération.

(1) L'identification des concepts : Les termes d'indexation (généralement des mots clés) sont d'abord extraits du document par une approche classique d'indexation (tokenisation, élimination des mots vides, puis lemmatisation) [1], [3], [4], [12], [25], [27]. Ces termes (non vides) sont ensuite projetés sur l'ontologie afin d'identifier les concepts (ou sens) correspondants dans l'ontologie. Un terme ambigu correspond à plusieurs entrées (sens) dans l'ontologie. Il faut le désambiguïser. Pour désambiguïser un mot ambigu, Voorhees [27] classe chaque synset 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. Dans une approche différente, Katz et al [25] 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. Dans l'approche d'indexation de Khan [12], pour désambiguïser un mot à partir des concepts correspondants (dans une ontologie de sport), on détermine le degré de corrélation des concepts sélectionnés, sur la base de leur proximité sémantique. La proximité sémantique de deux concepts est calculée par un score basé sur leur distance minimale mutuelle dans l'ontologie. Les concepts qui ont les plus hauts scores sont alors retenus. Dans une approche similaire, Baziz et al. [1] se basent sur le principe que, parmi les différents sens possibles (dits concepts candidats) d'un terme donné, le plus adéquat est celui qui a le plus de liens sémantiques [15], [16], [21] 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. La désambiguïsation est ici globale contrairement aux approches précédentes. Dans notre approche de désambiguïsation proposée dans [3], [4], ce score est basé sur la somme des valeurs de similarité qu'il a avec les concepts candidats les plus fréquents dans le document.

(2) La pondération des concepts : La pondération des concepts se décline en deux principales tendances : (1) la pondération des concepts en tant qu'entités lexicales et (2) la pondération des concepts en tant qu'entités sémantiques. Dans l'approche de pondération des concepts en tant qu'entités lexicales, les concepts sont considérés à travers les termes qui les représentent. La pondération des concepts consiste alors en la pondération des termes correspondants. Les approches de pondération de Voorhees [27] et de Baziz et al. [1] sont basées sur ce principe. En se basant sur le modèle vectoriel étendu introduit dans [9], dans lequel chaque vecteur est composé d'un ensemble de sous-vecteurs de différents types de concepts (appelés *ctypes*), Voorhees [27] propose de pondérer les concepts en utilisant un schéma de pondération classique *tf*idf* normalisé. L'approche proposée par Baziz et al. [1], étend la pondération *tf*idf* pour tenir compte des termes composés. L'approche proposée dite approche *Cf*idf*, permet de pondérer un terme *t* composé de *n* mots, par la fréquence d'occurrences du terme lui-même, et par celles des sous-termes qui le composent. Dans l'approche de pondération des concepts en tant qu'entités sémantiques, il s'agit d'évaluer l'importance des sens (concepts) dans le contenu du document indexé. L'importance d'un concept dans un document

est évaluée en tenant compte du nombre de ses relations sémantiques avec les autres concepts du document [5], [8], [11]. Ces relations sémantiques sont en outre pondérées dans [11]. Dans l'approche proposée par Boughanem et al. [5], le nombre de relations d'un concept avec les autres concepts définit une mesure dite de centralité du concept. Les auteurs combinent centralité et spécificité pour évaluer l'importance des concepts d'un document. La spécificité du concept définit son degré de « spécialité » (par opposition à généricité). .En combinant pondération sémantique et pondération lexicale des concepts, notre approche définie dans [3], [4], propose de pondérer les termes composés sur la base d'une mesure probabiliste tenant compte des sens possibles du terme par rapport aux sens de ses sous-termes, et de ses sur-termes en tenant compte de leurs fréquences d'occurrences respectives.

2.4 Positionnement de nos travaux

Notre approche proposée dans ce papier est une version revue de notre approche d'indexation conceptuelle dans [3], [4]. L'objectif est de représenter le document par un noyau sémantique composé de concepts pondérés. Les concepts sont extraits de WordNet à l'issue d'une identification-désambiguïsation. Puis les concepts sont pondérés. Dans notre approche d'indexation conceptuelle proposée dans ([3], [4]), les termes d'indexation sont d'abord extraits en se basant sur des étapes d'indexation classiques. Puis chaque mot non vide identifié est projeté sur WordNet. L'objectif est de recenser toutes les entrées de WordNet contenant ce mot. Ces entrées sont utilisées pour définir le contexte local du mot dans le document. Ce contexte permet d'identifier dans le document la collocation (suite de mots) la plus longue correspondant à un synset de WordNet. Lorsqu'un terme est ambigu, la désambiguïsation est appliquée. L'approche se base sur le calcul d'un score, basé sur la somme des valeurs de similarité qu'il a avec les concepts candidats les plus fréquents dans le document. Une approche de pondération des concepts est proposée. La pondération d'un terme *t* est basée sur une mesure probabiliste des sens possibles de *t* (notés *Sens(t)*) par rapport aux sens de ses sous-termes (*Sub(t)*) et de ses sur-termes (*Sur(t)*), en tenant compte de leurs fréquences d'occurrences respectives (*tf*). La probabilité qu'un terme *t* soit un sens possible d'un terme *t'*est mesurée comme le rapport entre le nombre de sens possibles du terme *t* incluant le terme *t'*, sur le nombre de sens possibles du terme *t*. Formellement :

$$P(t \in Sens(t')) = \frac{|\{C \in Sens(t') / t \in C\}|}{|Sens(t')|}$$

Le poids d'un terme *t* est alors défini par la formule suivante, où *N* représente le nombre total de documents dans le corpus et *df(t)* la fréquence documentaire inverse :

$$W_{t,d} = \left(tf(t) + \sum_i tf(Sur_i(t)) + \sum_j \left[P(t \in S(Sub_j(t))) * tf(Sub_j(t)) \right] \right) * \ln\left(\frac{N}{df(t)} \right)$$

Dans ce papier, nous redéfinissons l'approche d'identification des entrées de WordNet correspondant à un mot donné ainsi que l'approche de pondération des concepts.

- L'approche d'identification des concepts est basée sur le degré de recouvrement des entrées de WordNet et du contexte local (la phrase) dans

lequel le mot apparaît dans le document. Contrairement à l'approche proposée dans [3], cette approche présente l'avantage de permettre la détection de collocation de mots indépendamment de leur ordre d'apparition dans le contexte.

- L'approche de pondération des concepts est basée sur une nouvelle mesure de l'importance d'un concept dans un document. Cette mesure tient compte d'une part des proximités sémantiques entre le concept à pondérer et les autres concepts du document, et d'autre part des fréquences d'occurrences de ces concepts. Dans cette approche, l'apport des sous-termes n'est pas considéré. Notre précédente approche dans [3] peut être combinée à la présente approche pour en plus tenir compte de cet apport.

Dans ce qui suit, nous décrivons les différentes étapes de notre approche d'indexation conceptuelle.

3 Indexation conceptuelle des documents

L'indexation conceptuelle vise à représenter un document par un noyau sémantique composé 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 l'ensemble $T(d)= \{t_1, t_2,\ldots, t_n\}$ des termes t_i du document d qui correspondent à des entrées dans WordNet. L'identification des termes se base sur le degré de recouvrement du contexte local du mot analysé avec chaque entrée correspondante dans WordNet. L'entrée qui a le plus haut degré de recouvrement est retenue comme sens possible du mot analysé. Le principe de l'identification des termes est décrit à travers l'algorithme de la Table 2.

Algorithme de détection de concepts
Entrée : document *d*.
Sortie : index *T(d)*
Procédure : Soit mot_i, le prochain mot, non vide, à analyser dans *d*. On appellera contexte ζ_i du mot mot_i dans le document *d* la phrase courante de *d* qui contient le mot mot_i.

1. Calculer $S= \{C^i_1, C^i_2,\ldots C^i_n\}$ l'ensemble des synsets contenant le mot mot_i. *S* est composé de mono et de multi-mots.
2. Ordonner S comme suit : $S= \{C^i_{(1)}, C^i_{(2)},\ldots, C^i_{(n)}\}$ où $(j)_{1..n}$ est une permutation d'indices telle que $|C^i_{(1)}| \geq |C^i_{(2)}| \geq \ldots \geq |C^i_{(2)}|$, où $|C^i_{(j)}|$ est la longueur exprimée en nombre de mots, de la chaîne de caractères représentant le concept $C^i_{(j)}$.
3. Pour chaque $C^i_{(j)}$ dans *S*, faire :
4. Calculer l'intersection des deux chaînes de caractères ζ_i et $C^i_{(j)}$.
5. Si $\left|\zeta_i \cap C^i_{(j)}\right| < \left|C^i_{(j)}\right|$ (le concept $C^i_{(j)}$ n'apparaît pas dans le contexte ζ_i), le concept suivant, $C^i_{(j+1)} \in S$ est analysé,
Si $\left|\zeta_i \cap C^i_{(j)}\right| = \left|C^i_{(j)}\right|$ le concept $C^i_{(j)}$ est identifié comme concept associé au mot mot_i, et le terme représentatif correspondant t_i est ajouté à l'index du document *d* ;

Table 2: Algorithme de détection des termes d'index

3.2 Désambiguïsation des termes

Les termes d'index sont associés à des sens (synsets) correspondants dans l'ontologie. Chaque terme extrait pouvant avoir plusieurs sens possibles, le but de cette étape est de sélectionner le meilleur sens du terme dans le document. L'approche de désambiguïsation utilisée est celle proposée dans [3]. Pour désambiguïser un terme t_i donné, on associe à chacun de ses sens possibles C_j^i un score basé sur:

- les distances sémantiques $Dist(C_j^i, C_k^l)$ entre ce concept et les autres sens possibles associés aux autres termes dans le document,
- les fréquences d'occurrences des termes associés.

Formellement :

$$Score(C_j^i) = \sum_{\substack{l \in [1,..,m] \\ l \neq i}} \sum_{1 \leq k \leq n_l} tf(C_j^i) * tf(C_k^l) * Dist(C_j^i, C_k^l) tf(C^i) \quad (1)$$

Où $Dist(C_j^i, C_k^l)$ est la distance sémantique entre les concepts C_j^i et C_k^l.

Le concept C_j^i ayant le plus grand score est alors retenu comme sens adéquat du terme t_i dans d. L'ensemble des concepts retenus constituera le noyau sémantique $N(d)$ du document 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 fortement corrélé avec les concepts des termes les plus importants (au sens fréquents) du document compte tenu de sa propre importance dans le document, nous proposons de pondérer un concept avec un poids basé sur :

- sur les distances sémantiques entre ce concept et les autres concepts dans le document,
- et sur les fréquences d'occurrences des concepts associés.

Formellement, le poids $W(C^i)$ d'un concept C^i est défini par :

$$W(C^i) = \sum_{i \neq l} tf(C^i) * tf(C^l) * Dist(C^i, C^l) \quad (2)$$

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.

Evaluation expérimentale

L'objectif de ces expérimentations est de mesurer l'efficacité de notre approche de RI sémantique. On présente dans ce qui suit la collection de test et l'approche d'évaluation utilisées.

(1) Collection de test : La collection de test utilisée est la collection Muchmore[2]. Le corpus MuchMore est un corpus parallèle de résumés médicaux scientifiques

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

anglais-allemands obtenus à partir du site web de Springer. Seule la collection des textes anglais non annotée a été utilisée. Cette dernière collection est composée de 7823 documents et de 25 requêtes. Les documents et les requêtes sont composés de textes simples.

(2) Approche d'évaluation : L'approche est évaluée en utilisant le système Mercure [6]. L'évaluation est effectuée selon le protocole TREC. Chaque requête est soumise au système de RI avec les paramètres fixés. Le système renvoie les 1000 premiers documents pour chaque requête. Les valeurs de précision P5, P10, P20 et MAP (précision moyenne) sont calculées. La précision au point *x* (x=5, 10, 20,), *Px*, est le ratio des documents pertinents parmi les *x* premiers documents restitués. R-Prec et MAP sont les précisions exacte et moyenne respectivement. Nous comparons ensuite les résultats obtenus à partir de notre approche à un système de référence (ou baseline).

3.4 Evaluation de l'approche d'indexation par les concepts

Les premières expérimentations menées concernent l'approche d'indexation par les concepts. Il s'agit alors d'évaluer l'impact de la qualité de l'index sémantique du point de vue de l'efficacité de la recherche. Pour atteindre cet objectif, nous comparons plus précisément deux index :

- Le premier constitué par les concepts détectés par notre approche (décrite en section 2.2) et désambiguïsés (approche décrite en section 2.3). C'est l'approche notée *Concepts-TF* sur la Figure 1.

- Le second constitué par les concepts détectés par notre approche et désambiguïsés, combinés aux mots clés. Les mots clés font référence aux mots du document qui n'ont pas d'entrée correspondante dans WordNet. C'est l'approche notée *Concepts-Fusion* sur la Figure 1.

Les résultats de ces deux index sont d'abord comparés par rapport à ceux de deux baselines :

- La première est une baseline classique fondée sur une indexation basée mots clés pondérés par *tf*idf*. Cette approche est désignée par *Classic-TFIDF* sur la Figure 1,

- la seconde est une baseline classique fondée sur une indexation basée mots clés pondérés par la BM25 [22]. Cette approche est désignée par *Classic-OKAPI* sur la Figure 1.

Les résultats de l'évaluation de ces approches sont donnés en Figure 1. De ces résultats, il ressort que :

- l'approche d'indexation *Concepts-TF* par les seuls concepts désambiguïsés est meilleure que la baseline *Classic-TFIDF* avec des taux d'accroissement respectivement de 61% pour P5, de 51% pour P10, de 54% pour P20 et de 51% pour la MAP

- l'approche d'indexation *Concepts-Fusion* est nettement meilleure que l'approche *Concepts-TF* avec des taux d'accroissement respectivement de 20% pour P5, 19% pour P10, 15% pour P20 et de 23% pour la MAP. Ces résultats nous confortent dans l'idée qu'une indexation combinée concepts+mots-clés est plus performante qu'une indexation par les concepts seuls.

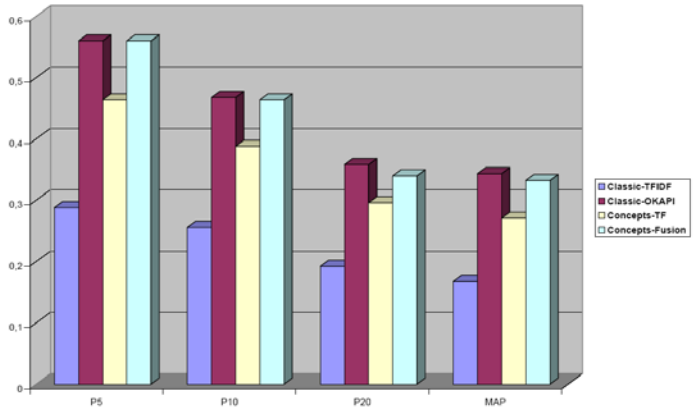

Figure 1 - Résultats d'évaluation de la méthode de détection de concepts par rapport aux baselines

Par ailleurs, notre approche combinée *Concepts-Fusion* présente des résultats nettement meilleurs qu'une baseline *Classic-TF*, avec des taux d'accroissement de 94% pour la P5, de 45% pour la P10, de 77% pour la P20 et de 77% pour la MAP. Néanmoins, comme le montre la Figure 1, l'approche *Concepts-Fusion* présente des résultats moins bons que ceux de la baseline *Classic-OKAPI* avec des taux de décroissement de 0% pour P5, -1% pour P10, -5% pour P20 et de -3% pour la MAP. La cause la plus probable à l'origine de ce problème pourrait être l'imprécision de la désambiguïsation. En effet, dans un contexte de désambiguïsation précise, on s'attend à ce que l'indexation par les concepts apporte au moins autant qu'une indexation classique.

3.5 Evaluation de l'approche de pondération de concepts

La deuxième série d'expérimentations menées concerne l'évaluation de notre approche de pondération des concepts introduite en section 2.3. Concrètement, il s'agit alors d'évaluer l'impact de la qualité de la pondération proposée (en section 2.3) du point de vue de l'efficacité de la recherche. Pour atteindre cet objectif, nous comparons plus précisément deux index :

- le premier est l'index composé des concepts détectés par notre approche proposée en section 2.2, pondérés par la fréquence. Cette approche est notée *Concepts-TF* sur la Figure 2.

- Le second est l'index composé des concepts détectés par notre approche proposée en section 2.2, pondérés par le poids proposé en section 2.3. Cette approche est notée *Concepts-Score* sur la Figure 2.

Les résultats de ces deux index sont comparés mutuellement. L'objectif est de mesurer l'apport de la pondération proposée par rapport à une pondération classique des concepts.

La Figure 2 présente les résultats obtenus. Il apparait que les résultats de la pondération par le poids proposé sont moins bons que ceux basés sur la fréquence des concepts, avec des taux de décroissement de -5% pour la P5, -6% pour P10, -12% pour P20 et -6% pour la MAP. Les résultats obtenus sont bien en deça de ce

qui était attendu. Le problème à l'origine de cette insuffisance vient probablement du score de *ranking*, utilisé par Mercure pour évaluer la correspondance d'un document pour une requête. Ce score est basé sur *tf*idf* (ou une de ses variantes). Ce qui explique que l'approche *Concept-TF* réponde favorablement à cette évaluation. A contrario, dans l'évaluation de l'approche *Concept-Score*, le poids du concept remplace *tf* dans le score de *ranking* et est combine donc à une mesure non corrélée, *idf*, provocant ainsi une baisse de la précision.

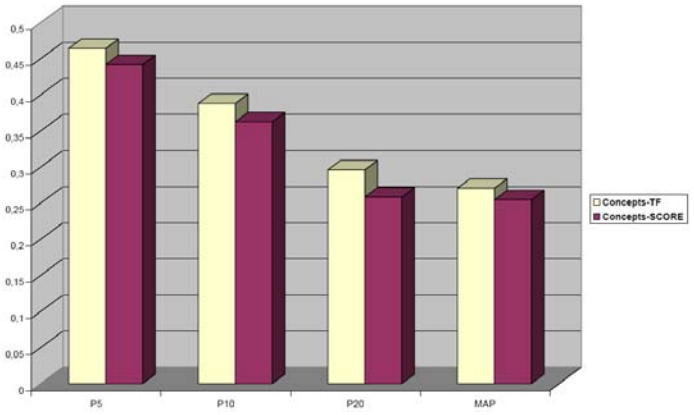

Figure 2 – Résultats d'évaluation de la méthode de pondération des concepts

4 Conclusion

Nous avons présenté dans ce papier, une approche d'indexation conceptuelle basée sur l'utilisation de WordNet. Le formalisme basé concepts est susceptible de résoudre les problèmes de disparité et d'ambiguïté des termes en RI.

En vue de palier à ces problèmes, nous avons présenté une approche en vue d'une indexation conceptuelle des documents. Notre contribution porte sur deux aspects principaux. Le premier consiste en l'indexation conceptuelle basée sur l'ontologie WordNet. L'approche n'est certes pas nouvelle mais nous avons proposé de nouvelles techniques pour identifier les concepts et pour les pondérer. Des résultats préliminaires ont montré que l'approche d'identification des concepts est plus performante qu'une baseline *Classic-TFIDF*, et apporte des taux d'accroissement appréciables par rapport à cette dernière, que les concepts soient utilisés seuls ou combinés aux mots clés. Cependant, l'approche d'indexation par les concepts n'a pas apporté les résultats escomptés par comparaison à une baseline *Classic-OKAPI*, probablement du fait de l'imprécision de la désambiguïsation. Par ailleurs, l'approche de pondération a produit des résultats mitigés. La cause probable de ces insuffisances inattendues serait l'inadéquation du score de *ranking* utilisé par rapport à l'index sémantique. Pour lever ces insuffisances, nous nous proposons, dans un premier temps, de parfaire le score de désambiguïsation, et dans un second temps de réfléchir un schéma de *ranking* pour des index sémantiques qui tienne compte des poids sémantiques des concepts et qui s'affranchit de la mesure classique *idf*. Des réflexions sont en cours dans ce sens.

5 Remerciements

Le présent travail a pu avoir lieu grace au soutien de l'A.U.F (Agence Universitaire de la Francophonie) et de l'U.M.M.T.O. (Université Mouloud Mammeri de Tizi-Ouzou) avec l'aimable collaboration de l'IRIT (Institut de Recherche en Informatique de Toulouse).

6 Bibliographie

[1] 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.

[2] M. Baziz, M.Boughanem, N.Aussenac-Gilles. The Use of Ontology for Semantic Representation of Documents. Dans: The 2nd Semantic Web and Information Retrieval Workshop(SWIR), SIGIR 2004, Sheffield UK, 29 juillet 2004. Ying Ding, Keith van Rijsbergen, Iad Ounis, Joemon Jose (Eds.), pp. 38-45.

[3] F. Boubekeur, M. Boughanem, L. Tamine. Exploiting association rules and ontology for semantic document indexing. Dans: 12th International conference IPMU08, Information Processing and Management of Uncertainty in knowledge-Based Systems, Malaga, 22- 27, june 08, Spain.

[4] F. Boubekeur, M. Boughanem, L.Tamine. Semantic Information Retrieval Based on CP-Nets. Dans : IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2007), London, 23/07/07- 26/07/07, IEEE, (support électronique), juillet 2007.

[5] 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 (à paraître).

[6] M. Boughanem, C. Soulé-Dupuy: A Connexionist Model for Information Retrieval. DEXA 1992: 260-265.

[7] M. Cuadros, JM., Atserias, J., M. Castillo, M., & G. Rigau, G. (2004). Automatic acquisition of sense examples using exretriever. In *IBERAMIA Workshop on Lexical Resources and The Web for Word Sense Disambiguation*. Puebla, Mexico.

[8] D. Dinh, L.Tamine. Vers un modèle d'indexation sémantique adapté aux dossiers médicaux de patients (short paper). Dans : Conférence francophone en Recherche d'Information et Applications (CORIA 2010), Sousse, Tunisie, 18/03/2010-21/03/2010, Hermès, Mars 2010.

[9] E.A. Fox. Extending the boolean and vector space models of information retrieval with p-norm queries and multiple concept types. PhD thesis, Ithaca, NY, USA, 1983.

[10] 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.

[11] B.Y. Kang and S.J. Lee. Document indexing: a concept-based approach to term weight estimation. In Journal of Information Processing & Management. Volume 41, Issue 5, September 2005, Pages 1065-1080

[12] L.R. Khan, D. Mc Leod, E.Hovy. Retrieval effectiveness of an ontology-based model for information selection. The VLDB Journal (2004)13:71–85.

[13] R. Krovetz. Homonymy and polysemy in information retrieval. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (A CL-97}, pages 72-79.

[14] 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.

[15] 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.

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

[17] 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

[18] R. Mihalcea and D. Moldovan. Semantic indexing using WordNet senses. In Proceedings of ACL Workshop on IR & NLP, Hong Kong, October 2000

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

[20] P. Resnik. Disambiguating noun groupings with repect to WordNet senses. 3thWorkshop on Very Large Corpora, 54–68. (1995).

[21] 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).

[22] S.E. Robertson, The probability ranking principle in IR. Journal of Documentation 33, 294-304 (1977). Reprinted in: K. Sparck Jones and P. Willett (eds), *Readings in Information Retrieval*. Morgan Kaufmann, 1997. (pp 281-286).

[23] H. Schütze and J. Pedersen. Information retrieval based on word senses. In Proceedings of the 4th Annual Symposium on Document Analysis and

Information Retrieval, pages 161-175.

[24] 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.

[25] O. Uzuner, B. Katz, D. Yuret: Word Sense Disambiguation for Information Retrieval. AAAI/IAAI 1999 : 985

[26] 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.

[27] 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).

[28] S.F. Weiss. Learning to disambiguate. Information Storage and Retrieval, 9, 33_41. (1973).

[29] Y. Wilks & M. Stevenson. Combining independent knowledge source for word sense disambiguation. Conference « Recent Advances in Natural Language Processing », 1–7.

[30] 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, 454 – 460.

[31] D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods, In 33rd Annual Meeting, Association for Computational Linguistics, Cambridge, Massachusetts, USA , 1995, (p189-196).