HAL open science

Vers une approche statistique pour l'indexation sémantique des documents multilingues

Farah Harrathi, Catherine Roussey, Loïc Maisonnasse, Sylvie Calabretto

To cite this version:

Farah Harrathi, Catherine Roussey, Loïc Maisonnasse, Sylvie Calabretto. Vers une approche statistique pour l'indexation sémantique des documents multilingues. 28ème congrès INFORSID, May 2010, Marseille, France. p. 127 - p. 143. hal-00495127

HAL Id: hal-00495127

https://hal.science/hal-00495127v1

Submitted on 25 Jun 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Actes du XXVIII congrès INFORSID, Marseille, mai 2010, p.127-143

Vers une approche statistique pour l'indexation sémantique des documents multilingues

Farah HARRATHI, Catherine ROUSSEY, Loïc MAISONNASSE, Sylvie CALABRETTO

Université de Lyon, CNRS, LIRIS UMR 5205, INSA de Lyon, Université Claude Bernard Lyon 1, Campus de la Doua, Bâtiment Blaise Pascal,20, Avenue Albert Einstein 69621 VILLEURBANNE CEDEX
CEMAGREF de Clermont Ferrand, 24, Avenue des Landais 63172 AUBIERE

RÉSUMÉ. *Dans cet article nous présentons une approche statistique d'indexation sémantique des documents multilingues. Cette approche est validée par un ensemble d'expérimentations et une comparaison avec une approche linguistique. Nous montrons ainsi que l'approche statistique obtient des résultats équivalents à l'approche linguistique.*

ABSTRACT. *This article presents a statistical approach of semantic indexing for multilingual documents. This approach is validated by a set of experiments and a comparison with a linguistic approach. The experiments show that the statistical approach obtains results equivalent to the linguistic one.*

MOTS-CLÉS : *recherche d'information, indexation sémantique, ontologie, documents multilingues, analyse linguistique, mesure statistique.*

KEYWORDS: *information retrieval, semantic indexing, ontology, multilingual documents, linguistic analysis, statistical measurement.*

1. Introduction

La numérisation des documents et le développement des technologies internet engendre une augmentation incessante de la masse de documents disponibles; Face à cette masse documentaire, le futur lecteur se sent désorienté et a besoin d'outils pour l'aider à filtrer les documents pour accéder aux documents pertinents. Dans ce but, les Systèmes de Recherche d'Information (SRIs) proposent à l'utilisateur une liste de documents répondant à son besoin d'information formulé sous forme de requête.

Les statistiques présentées dans (INTERNET, 2009) montrent la diversité des langues utilisées dans les documents et dans les requêtes des utilisateurs. Ainsi, les SRIs doivent maintenant répondre à un nouveau défi : proposer à l'utilisateur une liste de documents écrits dans des langues différentes répondant à une requête formulée dans la langue de l'utilisateur : la langue de l'utilisateur peut être différentes des langues des documents. Ces nouveaux systèmes documentaires portent le nom de SRI Translingues (2 langues) ou Multilingues (N Langues).

Notre travail s'intéresse au cas des SRI Multilingues gérant trois langues : français, anglais et allemand. Plus particulièrement, nous travaillons sur la phase d'extraction d'information à partir des textes, phase préliminaire au processus d'indexation des documents. Notre objectif est de surmonter la barrière de la langue dans un SRI en représentant chaque document d'un corpus multilingue par un ensemble de concepts. Les concepts composent un langage pivot de représentation de l'information et sont définis dans une Ressource Sémantique (RS) externe. Nous proposons donc une méthode statistique de détection des concepts. L'extraction des termes qui dénotent ces concepts est effectuée en se basant sur un corpus d'appui pour l'extraction des termes simples et sur une mesure statistique pour extraire les termes complexes. Ces termes sont ensuite transformés en concepts en se basant sur la ressource sémantique.

La section 2 présente un état de l'art sur les méthodes d'indexation utilisant des ressources sémantiques externes. Les étapes de notre méthode ainsi que les bases théorique sont présentées dans la section 3. Dans la section 4 nous présentons une validation expérimentale de la méthode et nous terminons par une conclusion.

2. Etat de l'Art

Plusieurs travaux ont utilisés des Ressources Sémantiques (RS) dans le processus d'indexation et ils ont obtenu de bons résultats. Ils se basent sur l'idée que l'utilisation des ressources sémantiques dans l'indexation d'un corpus multilingue peut améliorer la performance des systèmes de recherche d'information. Dans (MAISONNASSE et al., 2009), (GAUSSIER et al., 2008) (LACOSTE et al., 2006)

les auteurs utilisent UMLS (Unified Medical Language System) pour indexer la collection ImageCLEFmed. De même, le thésaurus MeSH (Medical Subject Heading) est utilisé dans (NEIL et al., 2007) pour indexer les documents de TREC.

Les méthodes d'indexation utilisées dans ces travaux se composent des mêmes étapes : identification de la langue du document, extraction des termes de cette langue par une méthode linguistique adaptée à la langue, détection des concepts par projection des termes extraits sur la ressource sémantique.

Au contraire de ces méthodes, nous voudrions montrer qu'avec une ressource sémantique externe de qualité suffisante, il n'est pas nécessaire d'utiliser des outils linguistiques adaptées à une langue donnée et qu'une méthode d'extraction des termes purement statistique permet d'obtenir des résultats de qualité équivalente. Ainsi, avec cette méthode statistique, nous ne seront pas amenés à changer d'analyseur linguistique à chaque fois que la langue du document change.

3. Notre approche de détection des concepts

Comme le montre la Figure 1, l'approche que nous proposons est composée de deux étapes: extraction des termes et détection des concepts. Chacune de ces étapes est constituée de deux processus distincts. Notre méthode de détection de concept est une étape préliminaire à un processus d'indexation sémantique. En effet, pour construire la représentation finale des documents ou des requêtes, une méthode d'indexation sémantique doit être complétée par exemple par un processus de pondération de concepts si le modèle de RI utilisé est de type vectoriel.

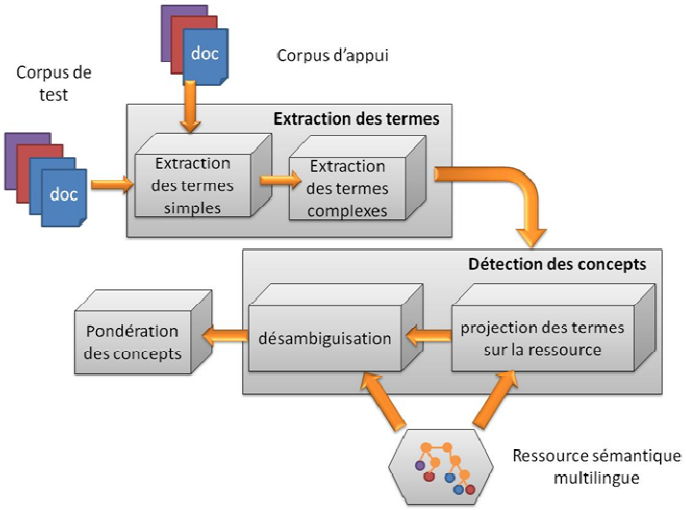

Figure 1 : *Vue d'ensemble de notre méthode de détection des concepts*

3.1. Extraction des termes

Nous distinguons deux types de termes : les termes simples composés d'un seul mot et les termes complexes composés d'une suite de mots. Nous utilisons un corpus d'appui pour extraire les termes simples et une nouvelle mesure statistique pour l'extraction des termes complexes.

3.1.1. Extraction des termes simples

Notre processus d'indexation commence par découper le texte en mots. Un mot est une suite de caractère alphanumérique délimitée par des caractères de ponctuation. Ensuite pour chacun des mots découverts nous le classons dans deux catégories : mots vides ou mots pleins.

Un « mot vide » est un mot qui ne doit pas être utilisé pour indexer un document. Ce type de mot a un pouvoir informatif faible : il peut être un mot grammatical ou un mot lexical non discriminant comme les jours de la semaine (VERGNE, 2004). Ces mots vides sont très fréquents. De ce fait, ils se distribuent de manière identique indépendamment des corpus où ils apparaissent (WIKIPEDIA, 2009). Le principe du moindre effort défini par George K. Zipf a mis en évidence que du fait de leur fréquence élevée ces mots sont généralement courts (ZIPF, 1994).

Au contraire un « mot plein » est un mot qui doit être utilisé pour décrire le contenu d'un document car il possède un pouvoir discriminant fort.

Nous approximons la liste des mots vides en supposant que les mots vides apparaissent dans deux corpus qui couvrent deux domaines disjoints et ont une taille inférieure à un seuil donné. Nous utilisons un corpus d'appui Ca qui couvre un domaine différent du domaine couvert par le corpus à indexer Ci.

$$\{ m \in Mots_Vides | \, taille(m) < Seuil \;\; et \; m \in V_{C_a} \cap V_{C_i} \}$$

Où

- V_{C_a} : Le vocabulaire du corpus d'appui
- V_{C_i} : Le vocabulaire du corpus à indexer
- Le seuil utilisé pour la taille des mots a été fixé à 4 lettres dans nos expérimentations. Cette valeur a été choisie arbitrairement et sera ajustée ultérieurement afin après plusieurs expérimentations.

Nous considérons un terme simple comme un mot non vide c'est-à-dire un mot plein.

$$\{ m \in Termes_Simples \mid m \in V_{C_i} \setminus Mots_Vides \}$$

Cette technique à base de comparaison de corpus est aussi utilisée dans l'outil TermoStat (DROUIN, 2003). TermoStat utilise un corpus non spécialisé pour réduire le nombre de termes simples d'un corpus spécialisé en se basant sur des comparaisons de fréquences.

3.1.2. Extraction des termes complexes

D'après (SMADJA, 1993), une collocation est une combinaison récurrente de mots qui se trouvent ensemble plus souvent que par le simple fait du hasard et qui correspondent à une utilisation arbitraire. Il existe différents type de collocations (en fonction du type grammaticale de ses composants ou de la régularité de la collocation). Certaines collocations représentent des syntagmes nominaux figés qui sont propres à un domaine. L'objet de notre étude porte sur la détection de ces types de collocations que nous appellerons termes complexes. Dans (ROCHE et al., 2008), les auteurs utilisent l'Information Mutuelle pour extraire des collocations pertinentes pour la constitution de lexiques spécifiques. Cette mesure consiste à comparer la probabilité d'apparition des cooccurences de mots à la probabilité d'apparition de ces mots séparément. Cette mesure est donnée par (CHURCH et al, 1990):

$$\mathrm{IM}(\mathrm{m}_1, \mathrm{m}_2) = \log_2\left(\frac{\mathrm{P}(\mathrm{m}_1,\ \mathrm{m}_2)}{\mathrm{P}(\mathrm{m}_1) * \mathrm{P}(\mathrm{m}_2)}\right)$$

Où $\mathrm{P}(\mathrm{m}_1)$ (respectivement $\mathrm{P}(\mathrm{m}_2)$) est une estimation de la probabilité d'apparition du mot m_1 (respectivement m_2) . Cette probabilité est calculée à partir de la fréquence d'apparition du mot m_1 dans le corpus, normalisée par N le nombre de mots contenu dans le corpus. $\mathrm{P}(\mathrm{m}_1, \mathrm{m}_2)$ est une estimation de la probabilité que les deux mots apparaissent ensemble dans une fenêtre de taille donnée. Cette probabilité est estimée par la fréquence d'apparition du couple $\mathrm{m}_1\mathrm{m}_2$ divisé par N.

Les termes complexes sont des séquences de mots qui se suivent et dont la probabilité d'apparition ensemble est plus fréquente que par le simple fait du hasard. l'IM est utilisée pour détecter des composantes des termes complexes, ces composantes sont forcément des mots pleins. Par compte, l'IM ne permet pas de détecter une suite de mot plein suivi de mot vide, car la fréquence des mots vides est très supérieure à la fréquence des mots pleins.

Pour pallier l'inconvénient de la formule de l'IM proposée dans (CHURCH et al, 1990), nous proposons une nouvelle mesure l'Information Mutuelle Adaptée (IMA). Cette nouvelle formule permet de détecter des suites de mots pleins suivi de mots vides, car nous substituons $\mathrm{P}(\mathrm{m}_2)$ par $\mathrm{P}(\mathrm{m}_1)$ si m_2 est un mot vide. L'IMA est donné par :

$$\mathrm{IMA}(\mathrm{m}_1, \mathrm{m}_2)\begin{cases} \log_2\left(\dfrac{\mathrm{P}(\mathrm{m}_1,\ \mathrm{m}_2)}{\mathrm{P}(\mathrm{m}_1)^2}\right) \ si\ \mathrm{m}_2 \text{ est un mot vide} \\ \log_2\left(\dfrac{\mathrm{P}(\mathrm{m}_1,\ \mathrm{m}_2)}{\mathrm{P}(\mathrm{m}_1) * \mathrm{P}(\mathrm{m}_2)}\right) \text{ sinon} \end{cases}$$

Le processus d'extraction des termes complexes est un processus itératif et incrémentale. Nous construisons les termes complexes de longueurs n+1 mots à partir des termes de longueur n mots. Nous partons de la liste des termes simples (de longueur 1 mot), pour chaque suite de mots on calcul la valeur de l'IMA. Les suites de mots dont la valeur de l'IMA est supérieure à un seuil fixé à 15 par expérimentation deviennent des «termes en construction». La valeur du seuil (15) correspond à la valeur de l'IMA à partir de laquelle la valeur de la précision moyenne et la valeur de la précision à 5 documents se stabilisent. Les termes en construction sont ajoutés à liste des termes en entrée d'une nouvelle itération du processus d'extraction. Un terme en construction devient un terme complexe quand il n'est plus possible de lui ajouter un mot. Ainsi le résultat du processus d'extraction des termes complexes contient les suites de mots les plus longues que l'on peut extraire des textes ainsi que les termes simples.

3.2. Détection des concepts à base d'une ressource sémantique

Pour transformer les termes complexes en concepts nous utilisons une ressource sémantique multilingue externe RS telle qu'un thésaurus ou une ontologie. Cette ressource contient un ensemble de concepts C et un ensemble de relations entre concepts R(c1,c2). Comme le montre la ressource de la Figure 2, à chaque concept est associé un ou plusieurs termes par langue avec les propriétés SKOS (MILES et al., 2005) « prefLabel » (preferred label) ou « altLabel » (alternative label). L'ensembles de ces termes forment le vocabulaire de la ressource et sera noté VRs.

```
<rdf:RDF>
<skos:Concept rdf:about="C1">
<skos:prefLabel xml:lang=en> Acquired Immunodeficiency Syndrome
</skos:prefLabel>
<skos:altLabel xml:lang=en > AIDS </skos:altLabel>
<skos:altLabel xml:lang=en > AIDS - HIV-1 stage 6</skos:altLabel>
<skos:prefLabel xml:lang=FRE> syndrome d'immuno-déficience acquise
</skos:prefLabel>
<skos:altLabel xml:lang=FRE > SIDA </skos:altLabel></skos:Concept>
</rdf:RDF>
```

Figure 2. *Exemple d'un concept d'une ressource sémantique au format SKOS*

3.2.1. Projection des termes dans la ressource sémantique

A partir de l'ensemble des termes extrait à l'étape précédente, nous recherchons dans le vocabulaire de la ressource sémantique VRs les termes complexe du document. Pour se faire nous définissons l'opérateur S_c, aussi nommé «conceptualisation des termes». Cet opérateur permet de déterminer le ou les concepts dénotés par un terme.

$$\forall\, t \in V_{Rs}, S_c(t) = \{c \in C / c \text{ est dénoté par } t\}$$

Ainsi pour le concept de l'exemple de la Figure 2 on aura :

$$\begin{cases} S_c(\text{Acquired Immunodeficiency Syndrome }) = \{C1\} \\ S_c(\text{"AIDS "}) = \{C1\} \\ S_c(\text{"AIDS} - \text{ HIV} - 1 \text{ stage 6 "}) = \{C1\} \end{cases}$$

.

3.2.2. Désambigüisation des termes

L'opérateur de conceptualisation de terme S_c consiste à affecter à chaque terme appartenant au vocabulaire de VRs des concepts de C. Cependant certains termes sont ambigus : les termes homographes et les termes polysémiques. Nous distinguons deux situations d'ambigüité: une ambigüité langagière et une ambigüité sémantique.

Ambigüité langagière ou homographie: deux termes appartenant à des langues différentes peuvent avoir la même forme dans un texte : termes homographes. Par exemple le mot « table » existe en français et en anglais. Dans le cas où ti est homographe, nous cherchons dans le document le terme t_k non ambigu le plus proche possible de t_i. Nous définissons une distance intitulé dist_doc(t_k, t_i) qui retourne le nombre de mots qui séparent t_k de t_i dans le document doc.

$$\forall\, t_i \in doc\ tq\ |S_c(t_i)| > 1, \exists t_k \in doc|$$

$$\forall t_j \in doc: |S_c\,(t_k)| = 1\ et\ |S_c(t_j)| = 1\ et\ dist_doc(t_k, t_i) \leq dist_doc(t_j, t_i)$$

Où |X| représente la cardinalité de l'ensemble X.

La langue de t_k définira la langue du terme t_i. Ainsi nous définissons opérateur Smc que nous appelons « désambigüisation langagière» qui permet de déterminer le concept (ou les concepts) dénoté par un terme dans une langue donnée, de la manière suivante :

$$\forall\, t \in V_{Rs}, S_{mc}(t, l) = \{c \in C\ /\ c \text{ est dénoté par t dans la langue l}\}$$

Ambigüité sémantique ou polysémie : un terme peut être associé à plusieurs concepts dans RS : termes polysémiques. Par exemple dans WordNet, le

terme « circuit » est associé à 8 concepts : il possède sept sens en tant que nom et un seul sens en tant que verbe. Dans le cas où ti est polysémique, nous recherchons les termes t_k non ambigus apparaissant dans la même phrase que ti. Si t_k est non ambigu, il existe un unique concept $c_k$dénoté par t_k. Nous privilégions le concept c_i dénoté par t_i qui est utilisé dans le plus grand nombre de relations de RS : R(c_k ,c_i) ou R(c_i , c_k). Pour se faire nous définissons la fonction nbRel:C×C⟶R qui retourne le nombre de relations définies dans RS utilisant un couple de concepts donné. S'il n'existe pas de terme t_k non ambigu dans la phrase de t_i, nous conservons l'ensemble des concepts dénotés par t_i.

$$\forall t_i \in doc\ tq\ |S_{mc}(t_i, l)| = |S_c(t_i, l)| > 1,$$

$$\forall t_k \in doc\ tq\ |S_c(t_k)| = 1\ et\ dist_doc(t_k, t_i) \leq phrase, \exists c_i \in S_{mc}(t_i, l)\ |$$

$$\forall c'_i \in S_{mc}(t_i, l),\ \sum_{c_k = S_c(t_k)} nbRel(c_k, c_i) \geq \sum_{c_k = S_c(t_k)} nbRel(c_k, c'_i)$$

4. Expérimentations et résultats

Dans cette section, nous commençons par la présentation des données de test, ensuite nous présentons les résultats de nos expérimentations.

4.1. Les données de test

Pour nos expérimentations, nous avons testé notre proposition sur des documents médicaux écrits en trois langues : l'anglais, l'allemand et le français. Nous avons utilisé comme corpus multilingue, les parties textuelles de la collection ImageCLEFmed 2007. La ressource sémantique utilisée est le thésaurus médical multilingue UMLS. Notre corpus d'appui a été constitué de deux corpus parallèles du domaine légal.

4.1.1. Le corpus de test : la collection ImageCLEFmed

ImageCLEFmed est constitué d'une collection d'images et d'un jeu de requête pour évaluer les systèmes de RI (MULLER et al., 2007). A chaque image de la collection est associé un diagnostic qui décrit l'image. Un diagnostic est associé à au moins une image. Un diagnostic est écrit en trois langues : l'anglais, l'allemand et le français. La collection 2007 est constituée de 55485 documents. Les requêtes de la collection ImageCLEFmed 2007 sont composées d'une image exemple et d'une partie textuelle. La partie textuelle d'une requête est écrite dans une des trois langues de la collection. Dans nos expérimentations nous avons utilisés 85 requêtes avec les jugements de pertinences associés. Les jugements de pertinence sont faits au niveau des images, nous considérons qu'un diagnostic est pertinent pour une requête donnée s'il correspond à au moins une image pertinente de cette requête.

4.1.2. Le corpus d'appui : rapports du parlement européen

Dans nos expérimentations, nous utilisons le corpus du parlement européen comme corpus d'appui. Ce corpus est un ensemble de 10 corpus parallèles écrits dans 11 langues (PHILIPP, 2005). Le corpus est collecté à partir des rapports du parlement européen. Pour constituer notre corpus, nous avons associé le corpus parallèle anglais-allemand et le corpus parallèle anglais-français. Ainsi, notre corpus d'appui est constitué des documents écrits dans les trois langues de notre corpus de test.

4.1.3. La ressource externe : le méta-thésaurus UMLS

UMLS (Unified Medical Language System) est un méta-thésaurus multilingue qui couvre le domaine médical. Cette ressource a été crée dans le but de faciliter la recherche et l'intégration d'informations provenant des multiples sources d'information biomédicales électroniques (NLM, 2009). Il est la fusion de plusieurs ressources sémantiques écrites telles que MeSH, SNOMEDCT et RXNORM (111 ressources). UMLS contient plus de 1 million de concepts reliés à plus de 5,5 millions de termes dans 17 langues. Les 17 langues ne sont pas couvertes de la même manière dans UMLS. L'anglais est la langue la plus représentée avec 68% du vocabulaire, l'allemand couvre 2,84 % du vocabulaire et le français 2,55%.

4.2 Méthodologie d'évaluation

Pour évaluer notre proposition, nous étudions les performances d'un SRI existant en variant la méthode de détection de concepts. Nous souhaitons comparer l'efficacité de notre méthode statistique de détection de concepts par rapport à une méthode linguistique utilisant différents analyseurs.

4.2.1 Le système de RI de référence

Nous utilisons comme système de référence, le système de Recherche d'Information basé sur une approche modèle de langue développé par L Maisonnasse (MAISONNASSE, 2008). Le modèle de langue appliqué à la RI a été proposé par Ponte et Croft (PONTE et al, 1998). L'idée de base considère chaque document comme un échantillon d'une langue donnée, et l'interrogation comme un processus génératif. Le système de RI, proposé dans (MAISONNASSE et al., 2009), utilise un modèle de langue défini sur des concepts intitulé Model Conceptuel Unigramme. Dans cette approche, une requête q est composée d'un ensemble C de concepts, chaque concept étant indépendant des autres, sachant le modèle de documents $\mathrm{M_d}$.

$$P(C|M_d) = \prod_{c_i \in C} P(c_i|M_d)^{f(c_i,q)}$$

Où $f(c_i, q)$est la fréquence du concept c_i dans la requête q.

La quantité $P(c_i|M_d)$ est calculée par un maximum de vraisemblance combiné à un lissage de Jelinek-Mercer.

$$P(c_i|M_d) = (1-\lambda_u)\left(\frac{f(c_i,\mathrm{d})}{\sum_{c_k\in\mathrm{d}} f(c_k,\mathrm{d})}\right) + \lambda_u\left(\frac{f(c_i,\mathrm{D})}{\sum_{d\ \in\mathrm{D}}\sum_{c_k\in\mathrm{d}} f(c_k,\mathrm{d})}\right)$$

$f(c_i,\mathrm{d})$ (respectivement $f(c_i,\mathrm{D})$) est la fréquence du concept ci dans le document d (respectivement dans la collection D) et λ_u est le lissage de Jelinek-Mercer estimé par apprentissage dans (MAISONNASSE et al., 2009).

4.2.2 La méthode linguistique de détection des concepts

Dans (MAISONNASSE et al., 2009), les auteurs utilisent une méthode linguistique de détection des concepts de UMLS sur la collection ImageCLEFmed 2007. La méthode se décompose en quatre étapes :

1. Analyse morpho syntaxique du document avec lemmatisation des formes fléchies ;
2. Filtrage des mots en fonction de leur catégorie grammaticale : seuls sont conservés les noms, les adjectifs et les abréviations.
3. Repérage des mots ou séquences de mots du document apparaissant dans UMLS. Cette étape est aussi appelée projection des mots du document dans UMLS. Son résultat est une liste de concepts UMLS.
4. Filtrage éventuel des concepts identifiés.

Cette méthode a été utilisée avec trois outils linguistiques différents :

- MetaMap (MM) est un analyseur morphosyntaxique associé à UMLS qui permet d'extraire les concepts à partir des documents. Cet analyseur ne traite que les documents écrits en anglais.
- MiniPar permet d'extraire les termes à partir des documents écrits en anglais.
- TreeTagger est un analyseur morphosyntaxique qui détecte la catégorie grammaticale d'un mot et son lemme. Il existe une version de TreeTagger pour l'anglais, le français et l'allemand.

MetaMap réalise l'ensemble des étapes de la méthode linguistique de détection de concepts UMLS. Ainsi un seul concept est détecté par terme. Lorsque les autres outils sont utilisés, l'étape 4 n'est pas exécutée, c'est-à-dire qu'aucun filtrage n'est réalisé sur les concepts UMLS. Trois versions différentes de cette méthode linguistique ont été utilisées sur le corpus trilingue de ImageCLEFmed 2007.

1. MM : La première version utilise l'analyseur MétaMap sur les documents anglais et TreeTagger sur les documents français et allemands.

2. MP : la seconde version utilise l'outil MiniPar sur les documents anglais et TreeTagger sur les documents français et allemands.
3. TT : la troisième version utilise l'analyseur TreeTagger sur les documents anglais, français et allemands.

4.3 Résultats

Nous évaluons notre méthode statistique d'extraction de concept à travers l'étude des résultats du SRI multilingue. Nous avons calculé la précision moyenne (PM) et la précision à 5 documents (P@5) avec trois ensembles de requêtes différentes appliquées sur le même corpus trilingue: les requêtes anglaises (ENG), les requêtes françaises (FRE), les requêtes allemandes (GER). Le choix de ces deux mesures se justifie par le fait que d'une part, la précision moyenne donne un aperçu général de l'efficacité de notre approche et d'autre, part par le fait que la précision à 5 documents donne un jugement de la performance de cette approche sur les documents les plus consultés par un utilisateur d'un SRI (les 5 premiers documents de la liste retournée par le SRI).

4.3.1 Evaluation de la détection statistique des concepts par langue

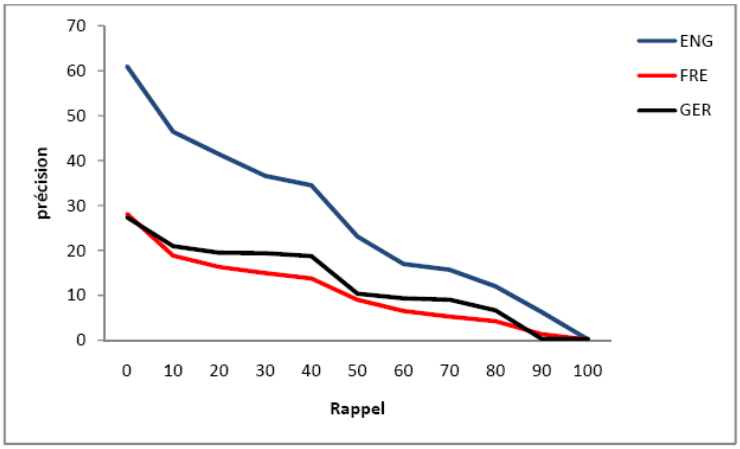

Figure 3. *Résultats de la méthode statistique de détection de concepts sur trois langues.*

La Figure 3 présente les courbes de précision moyenne à 11 points de rappel avec notre méthode statistique de détection de concepts. Les courbes montrent que la couverture de la langue dans la ressource sémantique à un impact direct sur les résultats du système. En effet, nous constatons que la précision obtenue pour les

requêtes écrites en anglais (ENG) est nettement plus importante que ceux écrites en français (FRE) ou en allemand (GER). UMLS couvre mieux la langue anglaise que les autres langues. La précision moyenne obtenue pour les requêtes (GER) et les requêtes (FRE) sont presque similaires avec une légère amélioration pour l'allemand. Ces deux langues sont couvertes de la même manière dans UMLS, et l'allemand a une couverture légèrement supérieure au français.

4.3.2 Comparaison de notre approche statistique avec les approches linguistiques

Nous comparons les résultats obtenus par le SRI utilisant notre méthode statistique de détection de concepts avec les trois versions du SRI basées sur trois outils linguistiques différents. Ces versions sont présentées dans (MAISONNASSE et al, 2009). Comme dans (MAISONNASSE et al, 2009), les auteurs traitent la même collection, la collection Clef médical 2007, cela nous permet de comparer directement nos résultats aux résultats obtenus par des analyses linguistiques. Cela est justifié par le fait que la démarche adoptée dans (MAISONNASSE et al, 2009) a permis aux auteurs d'obtenir la première place dans la campagne d'évaluation Clef médicale 2007 et d'obtenir la troisième place dans la campagne d'évaluation Clef médicale 2008.Les évaluations ont porté sur les 85 requêtes de la collection ImageCLEFmed. Le tableau 2 présente les valeurs obtenues pour la précision moyenne PM et la précision à 5 documents P@5.

Approche	**Analyse**	**PM**	**P@5**	**ΔPM**[1]	**ΔP@5**[2]
Linguistique	MM	0.246	0.357	-0.81%	19.05%
	MP	0.246	0.424	-0.81%	0.24%
	TT	0.258	0.462	-5.43%	-8.01%
Statistique	STAT	0.244	0.425		

Tableau 2. *Comparaison des méthodes de détection de concepts*

MM désigne l'analyse linguistique avec MetaMap, MP désigne l'analyse linguistique avec MiniPar, TT désigne l'analyse linguistique utilisant uniquement TreeTagger et STAT l'analyse statistique. Les résultats de la méthode de détection de concept à l'aide des trois outils linguistiques ont été présentés dans (MAISONNASSE et al., 2009).

Nous constatons qu'en précision moyenne, les méthodes linguistiques sont légèrement meilleures que les méthodes statistiques. En utilisant une approche

[1] $\Delta\mathrm{PM} = \frac{\text{PM obtenue par STAT} - \text{PM obtenue par analyse linguistique}}{\text{PM obtenue par analyse linguistique}} * 100$

[2] $\Delta\mathrm{P@5} = \frac{\text{P@5 obtenue par STAT} - \text{P@5 obtenue par analyse linguistique}}{\text{P@5 obtenue par analyse linguistique}} * 100$

statistique, la valeur de la précision moyenne a diminué de 2.35%. Par contre, en précision à 5 documents notre approche donne des résultats meilleurs. L'augmentation de la valeur de la précision à 5 documents est de 3.76% en moyenne.

En particulier, nous mentionnons que les résultats obtenus par l'analyseur TreeTagger sont meilleurs que les résultats obtenus par notre approche. De même ils sont meilleurs que les résultats obtenus par MetaMap et par MiniPar.

4.3.3 Evaluation du processus d'extraction des termes simples

Pour évaluer les résultats de notre phase d'extraction des termes simples, nous comparons la liste des mots vides extraits à partir de notre corpus de test avec une liste de référence. Cette liste de référence est constituée des listes de mots vides des trois langues de notre étude trouvées sur le web. Les listes des mots vides français, anglais et allemand contiennent respectivement 124, 36 et 127 mots. Ainsi, la liste de référence utilisée contient en total 287 mots vides. Nous avons utilisé les mesures de précision et rappel pour évaluer l'extraction des mots vides:

$$\text{Précision} = \frac{\text{(nombre de mots vides extraits et qui sont présents dans la liste de référence)}}{\text{(le nombre de mots vides extraits)}}$$

$$\text{Rappel} = \frac{\text{(nombre de mots vides extraits et qui sont présents dans la liste de référence)}}{\text{(nombre de mots vides de la liste de référence)}}$$

Ce qui donne

Précision=235/235=100%

Rappel =(235)/287=81.88%

Notre méthode d'extraction des mots vides à partir d'un corpus d'appui est précise et obtient un taux de rappel égal à 81.88%. Cette valeur montre que certains mots vides de la liste de référence ne sont pas extraits par notre processus d'extraction. Par exemple les mots « dedans », « dehors » et « force ». D'une part, ces mots sont absents du corpus de tests et d'autre part ces mots ont une taille supérieure au seuil utilisé fixé à 4 lettres.

4.3.4 Evaluation du processus de désambigüisation

Dans notre méthode de détection des concepts nous traitons deux types d'ambigüité : langagière et sémantique. Nous avons voulu évaluer l'impact de la désambigüisation sur les résultats du SRI. Ainsi, nous comparons les résultats obtenus sans désambigüisation (STA) aux résultats obtenus avec désambigüisation (ATA). Ces résultats sont présentés dans la Figure 4.

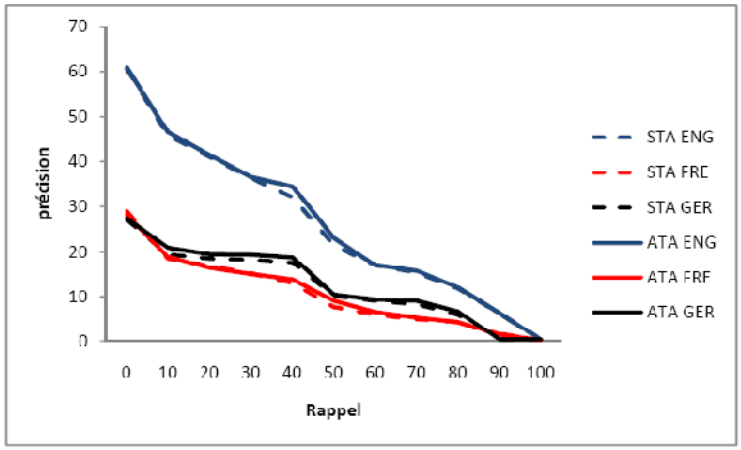

Figure 4. *Evaluation de l'impact de la désambigüisation*

La Figure 4, montre que les courbes de la précision à 11 points de rappel avec désambigüisation sont au dessus des courbes sans désambigüisation. La désambigüisation a amélioré la valeur de la précision moyenne de presque 5% pour les trois langues (voir Tableau 3). Par opposition à la précision moyenne, la précision à 5 documents a régressé. Cette régression est de 3% pour l'anglais, 6 % pour l'allemand et 14% pour le français. Elle s'explique par le fait que notre processus est plus sélectif. Cette sélection est surtout marquée dans les premiers documents retrouvés.

	STA		ATA			
Langue	PM	P@5	PM	P@5	ΔPM	ΔP@5
ENG	0.238	0.439	0.244	0.425	3%	-3%
GER	0.109	0.148	0.115	0.139	6%	-6%
FRE	0.086	0.148	0.089	0.127	3%	-14%

Tableau 3. *Résultats de la désambigüisation sur trois langues*

5. Conclusion

Dans cet article, nous avons proposé une méthode d'indexation sémantique des documents multilingues. Cette méthode se base sur une technique purement

statistique de détection de concepts et exploite une ressource sémantique. Notre méthode n'utilise pas d'analyse linguistique dans le processus d'extraction des termes simples et des termes complexes. Nos expérimentations ont montré que notre approche purement statistique donne des résultats similaires aux méthodes utilisant des techniques linguistiques mais elle est dépendante de la couverture de la langue dans la ressource sémantique. Elle présente l'avantage d'être facilement adaptable à d'autres corpus multilingues. De plus, l'approche statistique est simple à mettre en œuvre contrairement aux approches linguistiques. Cependant, notre approche présente une limite. Elle ne trouve sa performance que sur des corpus volumineux. En effet, notre approche est basée sur des mesures statistiques, ces mesures sont significatives uniquement sur des corpus de grandes tailles.

Dans des travaux futurs nous envisageons d'expérimenter notre approche sur des documents écrits en langue arabe tout en perfectionnant les deux opérateurs de projection sur la ressource sémantique (S_c et S_{mc}). En effet, la projection est stricte et ne prend pas en considération les variations lexicales et syntaxiques des termes

6. Bibliographie

CHURCH et al. (1990). CHURCH K.W and HANKS P., «Word association norms, mutual information and lexicography». *Computational Linguistic*, vol 1, Mars 1990, pp: 22-29 .

DROUIN. (2003). DROUIN P., «Term extraction using non-technical corpora as a point of leverage», In *Terminology*, vol. 9, no 1, p. 99-117.

GAUSSIER et al. (2008). GAUSSIER E, MAISONNASSE L AND CHEVALLET J-P., «Multiplying concept sources for graph modeling», In *CLEF 2007*, LNCS 5152 proceedings.

HERSH et al. (2000). HERSH W.R., PRICE S., DONOHOE L., «Assessing thesaurus-based query expansion using the UMLS Metathesaurus», *Proc AMIA Symp*. 344-8 .

INTERNET. (2009). INTERNET. INTERNET WORD STATS. Internet Usage World Stats - Internet and Population Statistics http://www.internetworldstats.com/.

LACOSTE et al. (2006). LACOSTE C, CHEVALLET J-P ,LIM J-H , WEI X, RACCOCEANU D, HOANG D.L.T, TEODORESCU R ET VUILLENEMOT F., «Ipal knowledge-based medical image retrieval in imageclefmed 2006», In Working Notes for the *CLEF 2006 Workshop*, 20-22 September, Alicante, Spain.

MAISONNASSE et al. (2009). MAISONNASSE Loïc., GAUSSIER Eric.,Chevallet J-P., «Combinaison d'analyses sémantiques pour la recherche d'information médicale », *Dans RISE (Recherche d'Information SEmantique) dans le cadre de la conférence INFORSID'2009*, Toulouse.

MAISONNASSE. (2008). MAISONNASSE L., « Les supports de vocabulaires pour les systèmes de recherche d'information orientés précision : application aux graphes pour la recherche d'information médicale », Université Joseph Fourier – Grenoble I, Thèse de Doctorat en Informatique.

MULLER et al. (2007). MULLER H, DESELAERS T, LEHMANN T, CLOUGH P and HERSH W., « Overview of the ImageCLEFmed 2006 medical retrieval and annotation tasks . Evaluation of Multilingual and Multi-modal Information Retrieval», *Seventh Workshop of the Cross-Language Evaluation Forum* .

NEIL et al. (2007). NEIL S, VETLE T,JIE H, WEI Z AND CLEMENT Y., « Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature», *30th annual international ACM SIGIR conference on Research and development in information retrieva*l.

MILES et al. (2005). MILES A, MATTHEWS B, BECKETT D, BRICKLEY D, WILSONM AND ROGERS N., « SKOS: A language to describe simple knowledge structures for the web, An introduction to SKOS», *2005 XTech Conference* .

NLM. (2009). NLM. Unified Medical Language System Fact Sheet [en ligne]. Disponible sur: http://www.nlm.nih.gov/pubs/factsheets/umls.html. (consulté le 23/04/2009) .

PHILIPP. (2005). PHILIPP K., « Europarl: A Parallel Corpus for Statistical Machine Translation», *MT* Summit 2005.

PONTE et al. (1998). PONTE J M. and CROFT W B., «A Language Modeling Approach to Information Retrieval», *21st annual international ACM SIGIR conference on Research and development in information retrieval*: 75-281.

ROCHE et al. (2008). ROCHE M et PRINCE V, «Evaluation et détermination de la pertinence pour des syntagmes candidats à la collocation», *JADT'08 (Journées internationales d'Analyse statistique des Données Textuelles)*, poster, pages 1009–1020, Volume 2, Lyon, France .

SMADJA. (1993). SMADJA F. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1), pp: 143-177.

VERGNE. (2004). VERGNE J., « Découverte locale des mots vides dans des corpus bruts de langues inconnues, sans aucune ressource », *Journées internationales d'Analyse statistique des données textuelles* 7: 8 pages.

WIKIPEDIA. (2009). WIKIPEDIA. Wikipédia, l'encyclopédie libre, http://fr.wikipedia.org/wiki/Mot_vide/ .

ZIPF. (1949). ZIPF G. K., « *Human Behavior and the Principle of Least Effort* », New York, Harper, réédition 1966.