Indexation sémantique des documents multilingues

Farah Harrathi[1,4], Catherine Roussey[2], Sylvie Calabretto[1], Loïc Maisonnasse[1,3], Mohamed Mohsen Gammoudi[4]

[1]Université de Lyon, CNRS, LIRIS UMR 5205, INSA, Campus de la Doua, bâtiment Blaise Pascal (501), 20, avenue Albert Einstein 69621 VILLEURBANNE CEDEX
[2] Université de Lyon, CNRS, LIRIS UMR 5205, Université Claude Bernard Lyon 1, Bâtiment Nautibus (710), 43, Boulevard du 11 Novembre 1918 69622 VILLEURBANNE CEDEX
[3]Equipe MRIM, Laboratoire LIG 38 041 Grenoble cedex 9, France.
prenom.nom@liris.cnrs.fr
[4]URPAH, Faculté des sciences de Tunis, Campus universitaire, Tunis ElManar 1090, Tunis, Tunisie
Mohamed.Gammoudi@fst.rnu.tn

Résumé : Dans cet article, nous décrivons une méthode d'indexation sémantique adaptée aux documents multilingues. Nous proposons une démarche d'extraction des concepts et des relations entre les concepts. L'idée centrale de notre travail est que l'utilisation de ressources sémantiques externes telle que les ontologies et les thésaurus peut améliorer l'efficacité des processus d'indexation. La méthode proposée peut s'appliquer à plusieurs langues car elle construit ses ressources linguistiques directement à partir du corpus multilingue.

Mots-clés : corpus multilingues, indexation sémantique, extraction de concepts, mesures statistiques, ontologie.

1 Introduction

Une consultation des statistiques relatives aux langues utilisées sur le web, montre que la langue anglaise était la langue de rédaction des documents de la plupart des utilisateurs d'internet jusqu'à l'année 2000[1]. Cette proportion est descendue à 30.4% en Mars 2008. Actuellement, la proportion d'internautes qui utilisent une langue européenne autre que l'anglais est d'environs de 27.1 %. Par contre, la proportion d'internautes utilisant une langue autre que les langues européennes est de 44.5%. Cependant, l'aspect multilingue ne se limite pas à Internet. Dans des pays multilingues tels que la Belgique, le Canada ou la Suisse, les textes réglementaires sont rédigés dans plus d'une langue. Pour le Canada, ces textes sont écrits en français et en anglais, produisant ainsi un corpus bilingue. De plus, les organisations internationales comme l'UNESCO (Organisation des Nations Unies pour l'Education, la Science et la Culture), l'ONU (Organisation des Nations Unies) l'OMC (Organisation Mondiale du Commerce) et le Parlement européen produisent quotidiennement des documents écrits dans plusieurs langues. Les entreprises

[1] Source, http://www.internetworldstats.com/stats7.htm

transnationales visant une clientèle dans différents pays produisent des documents rédigés dans plusieurs langues, comme les manuels d'utilisation de produits, les bons de commande, les affiches publicitaires, etc. Ainsi, ces documents constituent des corpus multilingues. Dans ces corpus chaque document est écrit dans une seule langue. Mais, il arrive souvent qu'un document dans une langue contient des passages écrits dans une ou plusieurs autres langues.

Avec l'augmentation incessante de ces documents multilingues et bilingues, il est devenu difficile de les gérer et de les exploiter. Cette difficulté est étroitement liée à l'aspect multilingue de ces documents. En effet, une exploitation manuelle de ces documents est possible et donne de bons résultats. Mais, une procédure manuelle est non envisageable avec des corpus de grande taille. De plus même avec des corpus de petite taille le travail est énorme et nécessite une compétence humaine souvent rare. Pour exploiter des documents d'un corpus multilingue où les documents sont écrits dans n langues, il faut n compétences, une dans chaque langue.

Actuellement, il est indispensable de proposer des méthodes et des approches qui permettent de gérer et d'exploiter ces documents : les Systèmes de Recherche d'Information MultiLingue (SRIML) sont donc devenus une nécessité.

Un système de recherche d'information est composé de deux processus : un processus de représentation et un processus de recherche. Dans une première étape, les documents et la requête sont représentés par des descripteurs regroupés pour constituer un index. Ces descripteurs reflètent au mieux le contenu des documents. Cette étape est appelée l'indexation. La deuxième étape est une étape de recherche. Dans l'étape de recherche une fonction de correspondance ou ranking est utilisée, afin de déterminer les documents qui répondent à la requête. Elle consiste à comparer la représentions des documents à la représentation de la requête. Cette fonction est notée RSV (Retrieval Status Value). Dans la plupart des processus d'indexation un poids est affecté à chaque descripteur. Ce poids permet de déterminer le pouvoir discriminant du descripteur dans le document où il est présent.

Un processus de Recherche d'Information (RI) consiste à indexer ou interpréter la requête (I_q), indexer chaque document (I_d) et comparer la représentation de la requête à la représentation de chaque document (RSV). Ce qui se traduit formellement par :

$$I_q : Q \to E \qquad (1)$$
$$q \mapsto I_q(q)$$
$$I_d : D \to E \qquad (2)$$
$$d \mapsto I_d(d)$$
$$RSV : E \times E \to \mathbb{R}^+ \qquad (3)$$
$$(I_q(q), I_d(d)) \mapsto RSV(I_q(q), I_d(d))$$

Avec

- Q est l'ensemble des requêtes,
- D *est l'ensemble des documents,*
- E *est l'ensemble des descripteurs.*

Le but de notre travail est de rendre automatique le processus d'indexation. C'est-à-dire d'extraire d'une manière automatique les descripteurs de chaque document. Dans ce papier nous proposons une méthode d'indexation sémantique des documents multilingues. La méthode d'indexation ainsi proposée consiste à :

1. extraire les concepts d'un document à indexer,
2. extraire les relations entre les concepts issus de l'étape précédente.

Dans la section suivante nous présentons notre méthode d'indexation sémantique, dans la section 3 nous présentons une prévision des tests et expérimentations que nous comptons faire prochainement et nous concluons à la section 4.

2 Indexation des documents multilingues

La méthode d'indexation que nous proposons consiste à extraire les concepts et les relations sémantiques entre les concepts à partir des corpus multilingues. Dans une première étape, nous identifions les concepts en repérant les termes qui les dénotent dans les documents et en projetant ces termes sur l'ontologie (mapping). Dans une seconde étape, nous détectons les relations qui résident entre ces concepts. Dans la suite nous présentons ces étapes en détail.

2.1 Extraction des concepts

La méthode que nous proposons pour l'extraction des concepts comprend deux grandes étapes :

1. une étape d'extraction des termes, qui permet d'associer à chaque document un ensemble de termes pertinents,
2. une étape de transformation de la représentation termes à la représentation concept.

2.1.1 Extraction des termes

Dans la littérature, les différents travaux d'extraction des termes à partir des corpus textuels utilisent deux approches : l'analyse statistique ou numérique [8] [18] [11] et l'analyse linguistique ou structurelle [10] [4]. L'analyse statistique se base sur l'étude des contextes d'utilisation et les distributions des termes dans les documents. L'analyse linguistique exploite des connaissances linguistiques, telles que les structures morphologiques ou syntaxiques des termes. D'autres travaux [28] [12] couplent ces deux approches et constituent une approche dite «approche hybride ou mixte».
L'approche que nous présentons permet d'extraire dans une première étape les termes simples et dans une deuxième étape les termes composés. Elle se base sur la définition d'un corpus spécialisé et sur des mesures statistiques telles que la loi de Zipf et l'information mutuelle.

Extraction des mots pleins : les candidats termes simples

Nous partons de la définition suivante d'un corpus spécialisé : «*un corpus spécialisé est un corpus limité à une situation de communication, ou à un domaine. Il s'intéresse aux langages de spécialité et aux sous-langages. Selon Harris, ces sous langages se caractérisent par un lexique limité et un nombre fini de schémas syntaxiques*» [21].

Nous proposons de découvrir automatiquement l'ensemble des mots vides d'une langue à partir de deux corpus spécialisés de cette langue couvrant deux domaines différents. Soit deux corpus spécialisés, C_A et C_B de deux domaines disjoints D_A et D_B, l'intersection de leurs vocabulaires contient les mots vides. Si V_A et V_B sont les vocabulaires des corpus C_A et C_B alors :

- $V_A \backslash V_A \cap V_B$ *: est le vocabulaire de spécialité du corpus* C_A*,*
- $V_B \backslash V_A \cap V_B$ *: est le vocabulaire de spécialité du corpus* C_B*,*
- $V_A \cap V_B$ *: est le vocabulaire qui n'est pas de spécialité (les domaines de* C_A *et* C_B *sont disjoints) et donc c'est un vocabulaire d'usage général et grammatical c'est-à-dire des mots vides.*

Nous signalons que dans cette étape nous n'avons pas traité le problème de ressemblance graphique. Il s'agit du cas des mots qui s'écrivent de la même manière dans des langues différentes, tels que les mots « science », « habit » et « caution » qui existent en français et en anglais. Nous avons jugé que ce problème n'a pas de grand effet sur les calculs. En effet, d'une part ces termes sont minoritaires et d'autre part seule la mesure de pondération sera erronée.

Validation de candidats termes simples: Extraction des termes simples

A l'étape précédente nous avons déterminé les mots pleins par différenciation des mots vides : les mots qui apparaissent dans l'intersection des vocabulaires des deux corpus. Cependant, l'intersection des deux vocabulaires ne contient pas seulement les mots vides, mais on peut trouver aussi des mots de spécialité (des mots pleins). En effet, deux domaines disjoints peuvent partager des mots ayant une sémantique différente dans chaque domaine. Ainsi, un mot peut être utilisé dans différents contextes ou différents domaines. Par exemple, le mot « Laser » est utilisé dans le domaine de la médecine et dans le domaine de l'informatique. C'est pourquoi nous passons par l'étape de validation des mots vides. L'objectif de la validation est d'éliminer les parasites, résultat des partages des mots de spécialité entre des domaines disjoints. Nous vérifions si un mot vide candidat est un mot vide ou un mot de spécialité commun (mot plein). Dans cette étape nous utilisons la loi de Zipf [33] et la conjecture de H. Luhn [19].

Dans [32], J. Vergne confirme qu'il est possible de construire la liste des mots vides en se basant sur la loi de Zipf. Cette loi énoncée par G. K. Zipf [33] considère que plus un mot est fréquent plus il est court[2]. La liste des mots vides est construite en se basant sur les longueurs et les fréquences des mots. Nous considérons comme mots vides les mots qui sont à la fois courts et fréquents. Ce test est effectué sur les mots situés dans l'intersection des vocabulaires des deux corpus. A l'issue de l'étape de validation des mots vides, un ensemble de termes simples est obtenu. Cet ensemble est formé par : l'ensemble des mots présents dans le corpus à indexer et absent dans le deuxième corpus (corpus d'appui), et l'ensemble des mots qui apparaissent dans les deux corpus et qui ne vérifient pas la loi de Zipf.

[2] « the length of a morpheme tends to bear an inverse ratio to its relative frequency of occurrence »

Pondération des termes simples

La pondération des termes consiste à affecter à chaque terme un poids qui représente son degré de pertinence dans le document où il apparait. Ce poids permet de distinguer les documents entre eux. En effet, un terme ne représente d'une manière adéquate le document que si son poids dans ce document est assez significatif. Un terme qui apparait dans tous les documents n'est pas discriminant c'est-à-dire qu'il ne permet pas de distinguer un document des autres documents. Un poids faible sera affecté à ce terme. Cette mesure n'a pas l'objectif d'éliminer des termes simples qui ont été déjà validés dans l'étape précédente. Mais, elle permet de trier ces termes par ordre d'importance. Cette mesure est souvent notée TF*IDF. Plusieurs formules de calcul de cette mesure ont été proposées dans [27] [23] [29] [30]. Dans notre approche nous avons adapté la formule de pondération normalisée utilisée dans le système Inquery [5] qui nous a semblé la plus adéquate.

$$TF * IDF_{i,j} = 0.4 + 0.6 * \left(\frac{tf_{i,j}}{tf_{i,j} + 0.5 + 1.5 * \frac{dl_j}{\Delta l}} \right) * \left(\frac{\log\left(\frac{N + 0.5}{n_i}\right)}{\log\left(N + 1\right)} \right)$$

- N : est le nombre total des documents dans le corpus.
- n_i : est le nombre de documents contenant terme i,
- $tf_{i,j}$: est la pondération locale du terme i dans le document j,
- dl_j : est la longueur du document j en nombre de mots,
- Δl : est la moyenne des longueurs des documents du corpus en nombre de mots.

Extraction des termes composés

L'extraction des termes composés est une étape que l'on retrouve dans les systèmes : LEXTER de D. Bourigault [4], le prototype ACABIT [9] et le système XTRACT [28]. Ces systèmes sont basés sur des approches linguistiques ou statistiques. Dans un contexte multilingue, il nous semble difficile d'adopter une approche linguistique. En effet, malgré que les résultats obtenus par les méthodes linguistiques sont jugés pertinents leur utilisation nécessite une maîtrise complète des langues des corpus étudiés. L'extraction des termes simples et des termes composés nécessite une connaissance parfaite des règles syntaxiques de dérivation dans la langue du corpus. Ces méthodes linguistiques sont basées sur des propriétés linguistiques de la langue naturelle. Ces propriétés sont intrinsèques à la langue du corpus d'étude. Elles ne sont pas, de ce fait, généralisables à d'autres langues. Ces propriétés et ces règles utilisées dans ces méthodes sont issues d'un traitement manuel du corpus d'étude. Ces éléments sont difficiles à dégager à partir des corpus volumineux. En effet, pour dégager une règle il est indispensable de feuilleter la quasi-totalité du corpus d'étude. Cette tâche n'est pas aisée dans le cas où le corpus est de grande taille. Les approches linguistiques trouvent leurs performances dans des corpus bien spécifiques sur lesquels une étude linguistique détaillée a été réalisé. Dans notre approche nous rejoignons les travaux de F. SMADJA.

Dans [28], F. SMADJA utilise l'approche statistique et propose le système XTRACT. XTRACT procède en deux étapes pour extraire les termes composés. Dans une première étape, XTRACT extrait l'ensemble des séquences de mots de longueur deux

dont une mesure statistique dépasse une valeur seuil notée VS. Cette valeur seuil VS est déterminée par expérimentation et paramétrée par l'utilisateur. La mesure statistique utilisée dans XTRACT est une fonction des fréquences d'apparition des mots et de la manière avec laquelle ces mots se distribuent les uns par rapport aux autres. Dans une deuxième étape, XTRACT étudie le contexte de chaque séquence de mots de longueur deux retenue. Il repère les séquences de mots de longueur trois dont la probabilité de cooccurrence des 3 mots consécutifs est supérieure à la valeur seuil Vs. Le processus est itératif et se termine lorsqu'aucun nouveau terme composé n'est repéré. XTRACT présente une faiblesse majeure due à l'utilisation d'une seule valeur seuil globale. En effet, l'identification d'un terme composé de longueur n+1 dépend largement de l'identification des termes complexes de longueur n. Par exemple l'identification du terme « laboratoire de recherche » dépend de l'identification du terme « laboratoire de ». Ce dernier terme possède une mesure très faible du fait de la forte fréquence du mot « de » dans l'ensemble des documents et il ne sera probablement pas retenu ce qui empêche par la découverte de tous les termes composés commençant par « laboratoire de ».

Dans notre approche nous adoptons la démarche de F. SMADJA [28] et nous proposons une technique statistique qui permet d'identifier les termes composés à partir d'un corpus de documents textuels multilingues. Cette approche se base sur une variante de l'information mutuelle. Afin de résoudre le problème de la construction des termes composés de longueur n+1 à partir des termes composés de longueur n, nous proposons de ne pas prendre en compte la fréquence d'un mot vide durant la construction. Par exemple pour le terme « laboratoire de» la fréquence du mot vide « de » ne sera pas prise en compte et elle sera substituée par la valeur de la fréquence du terme « laboratoire ». Durant le processus d'extraction des termes composés, le terme « laboratoire de» est marqué comme étant un « terme de construction ». Ce terme est supprimé à l'itération suivante. Ainsi, nous définissons une nouvelle mesure : l'information mutuelle adaptée. Pour un couple de termes (t_i, t_j) cette mesure est calculée de la manière suivante :

$$IMA(t_i, t_j) = \begin{cases} \log_2\left(\frac{f(t_i, t_j)}{f(t_i) * f(t_i)}\right) & si\ t_j\ est\ un\ mot\ vide \\ \log_2\left(\frac{f(t_i, t_j)}{f(t_i) * f(t_j)}\right) & si\ non \end{cases}$$

Pondération des termes composés

A cette étape nous cherchons à affecter à chaque terme composé extrait dans l'étape précédente une pondération qui reflète son importance dans le document. Dans [1] [2], l'auteur affirme que les termes composés ont en général un seul sens même si les termes qui les composent ont plus qu'un seul sens. Par la suite, ces termes ne requièrent pas de désambiguïsation sémantique. Ils sont sémantiquement plus riches que les termes simples qui les composent. Ainsi, nous proposons une nouvelle mesure de pondération qui favorise les termes composés, que nous appelons $CTF * IDF_{i,j}$(CTF pour Compound Term Frequency). Nous pensons que plus le terme composé est long, plus il est expressif et non ambigü. La pondération d'un terme composé dans un document dépend de quatre facteurs: la fréquence du terme dans ce document, la

fréquence du terme dans le corpus, les pondérations des termes simples qui le composent et la longueur du terme en nombre de termes simples. Dans la mesure que nous proposons, nous prenons en compte ces quatre facteurs. Les trois premiers facteurs sont représentés par la mesure classique$\mathrm{TF} * \mathrm{IDF}_{\mathrm{i,j}}$. La pondération d'un terme composé doit augmenter avec sa longueur ?. Nous augmentons la valeur de cette pondération par$\left(1 - \frac{1}{\mathrm{longueur(i)}}\right)$. La mesure $\mathrm{CTF} * \mathrm{IDF}_{\mathrm{i,j}}$ est donc exprimée en fonctions de ces facteurs de la manière suivante :

$$CTF * IDF_{i,j} = \left(1 - \frac{1}{longueur(i)}\right) + TF * IDF_{i,j} + \left(\frac{1}{longueur(i)}\right) * \sum_{k \in i} TF * IDF_{k,j}$$

Où

- i : un terme composé,
- j *:* un document,
- k *:* un terme simple entrant dans la composition de i et différent de i,
- $TF * IDF_{k,j}$: la pondération du terme k dans le document j,
- $longueur(i)$: le nombre de terme simples qui participe dans la construction du terme composé i,
- $TF * IDF_{i,j}$: la pondération du terme i dans le document j,

Dans le cas où i est un terme simple nous retrouvons la valeur de la mesure $\mathrm{TF} * \mathrm{IDF}_{\mathrm{i,j}}$. En effet, $\mathrm{longueur(i)} = 1$ et i ne contient pas d'autres termes simples.

$$CTF * IDF_{i,j} = \left(1 - \frac{1}{longueur(i)}\right) + TF * IDF_{i,j} + \left(\frac{1}{longueur(i)}\right) * \sum_{k \in i} TF * IDF_{k,j}$$

$$\Rightarrow CTF * IDF_{i,j} = \left(1 - \frac{1}{1}\right) + TF * IDF_{i,j} + \left(\frac{1}{1}\right) * 0$$

$$\Rightarrow CTF * IDF_{i,j} = TF * IDF_{i,j}$$

Par exemple la pondération pour le terme composé « ministère des affaires étrangères » est calculée comme suit :

CTF*IDF (« ministère des affaires étrangères ») = $\left(1 - \frac{1}{3}\right)$+TF*IDF (« ministère des affaires étrangères ») + $\frac{1}{3}$ (TF*IDF (« ministère») + TF*IDF (« affaires») + TF*IDF (« étrangères »).

2.1.2 Transformation de la représentation termes à la représentation concept: Extraction des concepts

Le but de cette étape est d'extraire les concepts à partir des documents multilingues. Ces concepts sont dénotés dans les documents textuels par des termes simples ou composés. Ces termes ont été extraits pendant la phase précédente. Il reste à faire la correspondance entre les termes et les concepts. Pour ce faire nous nous basons sur une ressource sémantique externe: une ontologie multilingue de domaine.

Dans le cadre de notre travail, nous considérons qu'une ontologie est composée d'un ensemble de concepts et un ensemble de relations entre les concepts. Formellement une ontologie O est définie de la manière suivante :

$$O = \{C, R, \leq^{C}, \sigma_{R}, V_{O}\}$$

- *C : l'ensemble des concepts de l'ontologie,*
- *R : l'ensemble des relations entre les concepts de l'ontologie,*
- *$\leq^c : C \times C$ est un ordre partiel sur C, il définit la hiérarchie de concepts, $\leq^c (c_1, c_2)$ signifie que c_1 subsume c_2 (relation orientée)*
- *$\sigma_R : R \rightarrow C \times C$ est la signature d'une relation*
- V_o.*est le vocabulaire de l'ontologie*

Dans l'ontologie multilingue un identifiant unique est attribué à chaque concept. Dans chaque langue, un concept possède un label qui est un terme associé à ce concept. Cependant il arrive qu'un concept possède plusieurs labels. Pour les distinguer un seul label est marqué « préféré ». Les autres sont marqués « alternatifs ». Les labels alternatifs sont en général des synonymes du label préféré.
Par exemple le concept « `concepts#shrubs`» représenté dans la **Fig. 1** possède deux labels anglais et deux labels français :

- En anglais, ce concept possède un label préféré : le terme « `shrubs`» et un label alternatif : le terme « `bushes`».
- En français, ce concept possède un label préféré : le terme « `arbuste`» et un label alternatif : le terme « `buisson`».

```
  <rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#">

    <skos:Concept
rdf:about="http://www.example.com/concepts#shrubs">
      <skos:prefLabel xml:lang="en">shrubs</skos:prefLabel>
      <skos:altLabel xml:lang="en">bushes</skos:altLabel>
      <skos:prefLabel xml:lang="fr">arbuste</skos:prefLabel>
      <skos:altLabel xml:lang="fr">buisson</skos:altLabel>
    </skos:Concept>

  </rdf:RDF>
```

Fig. 1. Exemple de concept d'une ontologie multilingue

Dans l'ontologie un ensemble de termes est utilisé afin de labéliser les concepts et les relations entre les concepts. Cet ensemble forme le vocabulaire de l'ontologie et sera noté V_o. $V_o = \{V_{OC}, V_{OR}\}$

- *V_{OC} : l'ensemble des termes utilisés pour labélliser les concepts de l'ontologie,*
- *V_{OR} : l'ensemble des termes utilisés pour labélliser les relations de l'ontologie.*

Sur l'ensemble V_{OC} on définit les relations de référence L_{mc} , L_c, S_{mc} et S_c de la manière suivante :

$$\begin{cases} \forall\, c \in C, L_{mc}(c, l_i) = \{t \ \in V_{OC}\ tel\ que\ \ t\ \ dénote\ c\ \ dans\ la\ langue\ l_i\} \\ et \\ \forall\, t \in V_{OC}, S_{mc}(t, l_i) = \{c\ \ \in C\ \ tel\ que\ c\ \ est\ labelisé\ par\ t\ dans\ la\ langue\ l_i\} \\ et \\ \forall\, c \in C, L_c(c) = \{t\ \ \in V_{OC}\ tel\ que\ \ t\ \ dénote\ c\ \ \} \\ et \\ \forall\, t \in V_{OC}, S_c(t,) = \{c\ \ \in C\ \ tel\ que\ c\ \ est\ labelisé\ par\ t\ \} \end{cases}$$

Ainsi pour le concept « `concepts#shrubs`» de l'exemple précédent on aura :

— L_{mc}("concepts#shrubs", "en") = {"shrubs", "$bushes$"}
— L_{mc}("concepts#shrubs", "fr") = {"arbuste", "$buisson$"}
— S_{mc}("shrubs", "en") = {"$concepts\#shrubs$"}
— S_{mc}("$bushes$", "en") = {"$concepts\#shrubs$"}
— S_{mc}("arbuste", "fr") = {"$concepts\#shrubs$"}
— S_{mc}("$buisson$", "fr") = {"$concepts\#shrubs$"}
— L_c("concepts#shrubs") = {"shrubs", "$bushes$""arbuste", "$buisson$", }
— S_c("shrubs") = {"$concepts\#shrubs$"}
— S_c("$bushes$") = {"$concepts\#shrubs$"}
— S_c("arbuste") = {"$concepts\#shrubs$"}
— S_c("$buisson$") = {"$concepts\#shrubs$"}

La méthode que nous proposons pour l'extraction des concepts à partir des documents multilingue consiste à affecter à chaque terme d'un document le concept associé. Afin d'identifier le concept associé à chaque terme nous utilisons la relation S_{mc} définie précédemment. Cependant il arrive souvent que le terme à traiter soit ambigu. Nous distinguons deux situations d'ambigüité: une ambigüité langagière et une ambigüité sémantique :

- **Ambigüité langagière** : deux termes appartenant à des langues différentes peuvent avoir la même forme dans un texte, cette relation peut être vue comme une relation d'homonymie multilingue. Par exemple le mot « table » existe en français et en anglais. Dans ce cas, nous cherchons dans le document le terme le plus proche non ambigu du point de vue langue. La langue de ce terme situé à proximité du terme ambigu définiera la langue du terme ambigu. Si un tel terme n'existe pas, on prend toutes les langues du terme ambigu.
- **Ambigüité sémantique** ou polysémie : cas où plusieurs concepts sont dénotés par le même terme c'est-à-dire qu'un même terme peut être le label de plusieurs concepts dans l'ontologie. Ainsi ce terme renvoi à des concepts différents. Par exemple le terme « circuit » possède dans WordNet[3] sept sens comme nom et un sens comme verbe. Il peut donc renvoyer à huit concepts différents. Dans le cas de la polysémie, nous procédons de la manière suivante. Pour un terme ambigu t_i nous cherchons dans le document un label d'un concept c en relation, dans l'ontologie, avec un concept c' dénoté par le terme ambigu t_i. Si c existe on prend c' comme étant le concept dénoté par ce terme. Si non, on prend tous les concepts dénotés par le terme considéré.

[3] www.wordnet.princeton.edu/

Dans le processus d'extraction des concepts nous effectuons deux passes. Dans la première, nous ne traitons que les termes non ambigus. Cela nous permet de les utiliser pour désambigüiser les termes ambigus dans la deuxième passe. L'algorithme de la méthode est le suivant :

Algorithme Extraction des concepts

Entrée: LT : *liste des termes du corpus* à *indexer*

Sortie: LC : *liste des concepts avec leurs pondérations associés* à *chaque document de* C

pour tout *treme* $t_i \in LT$ **faire**
  $L \leftarrow S_c(t_i)$
  si $(|L| = 1)$ *alors*
    pour tout *document* $d_j \in Ce$ *et contenant le terme* t_i **faire**
      Ajouter $(L[0], d_j, pondération(t_i, d_j))$;
    fin pour
  fin si
fin pour
pour tout *terme* $t_i \in LT$ **faire**
  $L \leftarrow S_c(t_i)$
  si $(|L| > 1)$ *alors*
    pour tout *document* $d_j \in Ce$ *et contenant le terme* t_i **faire**
      si $(d_j$ *contient un terme* t_k *non ambigu* $)$ *alors*
        $L \leftarrow S_{mc}(t_i, langue(t_k))$
        si $(|L| = 1)$ *alors*
          Ajouter $(L[0], d_j, pondération(t_i, d_j))$;
        si non
          si $(d_j$ *contient un concept* C *en relation avec un* $C_p \in L)$ *alors*
            Ajouter $(C_p, d_j, pondération(t_i, d_j))$
          fin si
        si non
          Ajouter $(L, d_j, pondération(t_i, d_j))$
          fin si
      fin pour
    fin si
fin pour

2.2 Extraction des relations entre concepts

Dans cette section nous présentons un état de l'art sur les travaux qui se sont intéressés à l'extraction des relations entre concepts, puis nous décrivons notre approche.

2.2.1 Travaux similaires

Dans la littérature, peu de travaux se sont intéressés à l'extraction des relations entre concepts. La majorité de ces travaux visent la construction ou l'enrichissement des ressources sémantiques, comme les thésaurus et les ontologies.
Certains de ces travaux se basent sur des patrons syntaxiques ou lexico-syntaxiques. Dans un premier temps, un ensemble de patrons lexico-syntaxiques est défini (un pour chaque relation). Dans un deuxième temps, ces patrons seront projetés sur le corpus

de texte afin de repérer les instances des relations. La construction des patrons lexico-syntaxiques est alors une étape préliminaire afin de découvrir les relations dans un corpus. Précisément, il s'agit d'une acquisition des marqueurs de relations à partir du corpus étudié. Dans [14], pour l'extraction des liens d'hyperonymie à partir de textes M. Hearst propose la méthode itérative suivante :

1. sélectionner la relation R,
2. établir une liste de termes pour lesquels on a identifié cette relation,
3. trouver dans le corpus des phrases où les termes reliés sont co-occurrents,
4. trouver les régularités dans ces phrases et faire l'hypothèse que ces phrases sont la base de formules ou patrons qui indiquent la relation étudiée,
5. si un nouveau patron a été repéré et validé, utiliser ce patron pour trouver d'autres couples en relation et revenir en (2).

Exemple:

PS.N°	Patron Syntaxique	Relation d' hyperonymie ($\forall\ NP_i\ 1 \leq i \leq n$)
1	$NP_0\ such\ as\ \{NP_0, \ldots, (and \mid or)\} NP_n$	hyperonymie (NP_i , NP_n)
2	$such\ NP_0\ as\ \{NP_i^*, (and \mid or)\} NP_n$	hyperonymie (NP_i , NP_n)
3	$NP_1\ as\ \{, NP_i\}^*\{, \}(or \mid and)\}\ other NP_{n+1}$	hyperonymie (NP_i , NP_{n+1})
4	$NP_0\ \{, \}\ (including \mid especially)\ \{NP_i, \}^*\ (or \mid and)\ NP_n$	hyperonymie (NP_i , NP_0)

Fig. 2. Les patrons utilisés par Hearst pour l'extraction de l'hyperonymie

La phrase: «The bow lute, such as the Bambara ndang, is plucked and has an individual curved neck for each string», satisfait le patron 1 du tableau 2. Dans cette phrase, NP0 correspond à «bow lute» et NPn correspond à «Bambara ndang». La relation ainsi extraite est :

Hyperonymie (« Bambara ndang », « bow lute »)

La méthode, présentée par M. Hearst, fournit des résultats jugés pertinents pour la relation d' hyperonymie. Cependant, l'auteur signale les difficultés pour la généralisation de ce type de méthode à d'autres relations comme la relation de méronymie et souligne qu'elle obtient de bons résultats pour l'identification de relations spécifiques.

La méthode présentée par M. Hearst a été reprise dans de nombreux travaux d'extraction des relations à partir de corpus [24] [20] [26] [7]. Ces travaux partent du même principe : la découverte de schémas lexico-syntaxiques dans un corpus. Ils effectuent une recherche itérative dans le corpus textuel des marqueurs d'une relation donnée et des couples de termes qui entrent dans cette relation. On trouve l'utilisation des patrons syntaxiques dans [3] [13] [31]. Dans ces travaux les auteurs ont été intéressés par l'extraction de la relation « partie de ». Ces travaux se basent tous sur les patrons syntaxiques. Ils diffèrent par la manière avec laquelle s'effectue l'extraction des patrons. Cependant il est à signaler que dans les approches qui se basent sur les patrons syntaxiques, les relations à extraire sont connues à priori.

D'autres travaux se basent sur l'idée que les relations sont marquées par les verbes. De ce fait, ils utilisent des patrons de relations et des schémas syntaxiques afin de repérer les relations entre des paires de concepts prédéfinies. Dans ces patrons, les marqueurs de relations sont les verbes. Dans [25] [6], les auteurs utilisent des patrons afin de repérer les relations entre concepts. Ils procèdent à une extraction des couples

de concepts en relation. Ces couples co-occurrents dans les mêmes phrases. Par la suite, ils utilisent une mesure statistique, la mesure χ^2 pour vérifier le degré d'association de ces couples avec le marqueur de la relation, le verbe. Une fois la relation validée par la valeur de χ^2, ils construisent le patron de la relation. Ce patron sera projeté dans le texte afin de repérer de nouvelles occurrences de cette relation. L'approche présentée dans [16] est fondée sur le même principe. Dans cette approche les auteurs utilisent une variante de l'information mutuelle appelée AE. Ils mesurent le degré d'association d'un couple de concepts (C_1, C_2) en présence d'un verbe V. La mesure AE utilisée consiste à comparer la probabilité d'apparition du couple (C_1, C_2) avec le verbe V à la probabilité d'apparition de chacun des concepts seul avec le verbe. Cette mesure est donnée par la formule suivante :

$$AE\big((C_1 \hat{} C_2)\big|V\big) = \frac{P\big((C_1 \hat{} C_2)\big|V\big)}{P(C_1|V) * P(C_2|V)}$$

Dans [22], l'auteur signale que cette dernière mesure ne permet pas d'identifier la direction de relation. C'est-à-dire qu'avec la mesure AE on ne peut pas distinguer le cas où C_1 apparait comme premier argument, (C_1, R_k, C_2), du cas où C_1 apparait comme deuxième argument, (C_2, R_k, C_1). Dans l'approche proposée dans [22], l'auteur prend en considération la direction de la relation et distingue ainsi les deux cas mentionnés. Il différencie le cas où C_1 est un sujet du cas où C_1 est un objet dans les phrases du texte où C_1apparait. Pour un couple (C_1, C_2) de concepts, les deux concepts ne sont reliés par la relation R_k que s'il existe une phrase dans le texte dans laquelle, C_1 est le sujet et C_2 est l'objet. Pour extraire les labels des relations, l'auteur ne considère que les verbes spécifiques au domaine et qui apparaissent souvent avec les couples de concepts. Afin de déterminer les verbes spécifiques au domaine, l'auteur utilise une forme de la mesure statistique de pondération, $TF * IDF$ il la nomme $VF * ICF$. L'auteur se base sur l'idée que les verbes qui occurrent avec peu de concepts sont plus pertinents que les verbes qui occurrent avec tous les concepts. Pour un verbe V, la mesure $VF * ICF$ est calculée de la manière suivante :

$$VF * ICF(V) = \big(1 + \log\big(VF(V)\big)\big) * \log\left(\frac{|C|}{CF(V)}\right)$$

Dans cette équation, $|C|$ est le nombre total des concepts, $VF(V)$ est la fréquence du verbe V et $CF(V)$ est le nombre de fois où le verbe V apparait avec un concept de C.

En 2000, F. Le Priol [17] a présenté le système SEEK-JAVA. Ce système est conçu pour être le successeur de SEEK de C. Jouis [15]. SEEK-JAVA, un peu comme SEEK, se base sur des variantes de patrons syntaxiques afin d'extraire des occurrences de la relation « Partie-de », la relation de « localité », la relation « est un attribut de » et la relation « est un ». L'auteur définit un ensemble de règles syntaxiques pour chaque relation à extraire. Ces règles sont de type « si condition alors ». Par la suite il repère dans le texte, les phrases qui vérifient ces règles. Par exemple, pour la relation de « localité » la règle est la suivante :

SI on trouve une occurrence du verbe être
et
SI on trouve un élément de la liste L
ALORS on a une relation de localisation
L = {(à l'intérieur de, dans…), (hors de, à l'extérieur de …), (sur, au dessus ,)…}

Ainsi, l'auteur identifie la présence de la relation dans le texte en repérant les phrases satisfaisant la règle de la relation. Par exemple dans la phrase « Je suis dans la chambre de mon héros » la relation « localisation » est identifiée.
Dans la majorité des travaux, les corpus sont monolingues. Ces corpus sont lemmatisés et étiquetés. Une catégorie grammaticale est attribuée à chaque terme afin de distinguer les verbes, des noms. Ces verbes seront utilisés comme marqueur ou déclencheur de relation. Les travaux visant la construction ou l'enrichissement d'une ressource linguistique, utilisent des mesures statistiques pour détecter des hypothèses de relations entre couples de concepts. Ces mesures calculent le degré d'association entre des couples de concepts. Une fois détectée une hypothèse de relation par une mesure statistique, un patron ou une règle est construit pour représenter cette nouvelle relation. A l'opposé des travaux d'enrichissement de ressources, d'autres travaux se focalisent sur l'extraction d'un petit nombre de relations. Ces relations sont connues à priori et ne nécessitent pas une phase de détection des hypothèses de relations. Ainsi, les patrons et les règles sont construits au préalable sans avoir recours aux mesures statistiques comme dans SEEK-JAVA.
Dans le cadre de notre travail, l'utilisation de l'ontologie présente un avantage majeur. En effet, les relations à extraire sont déjà connues en avance. Il s'agit des relations qui existent dans l'ontologie. Ainsi, dans notre approche nous ne sommes pas amenés à utiliser des mesures statistiques pour détecter des hypothèses de relations. Contrairement aux travaux cités, les documents du corpus sont multilingues. Cet aspect rend difficile la construction des patrons de relation. En effet la construction des patrons syntaxiques nécessite une maîtrise parfaite de chaque langue du corpus.

2.2.2 Notre proposition pour l'extraction des relations

Nous dissocions, comme dans la majorité des travaux, le problème d'extraction des relations entre les concepts en deux sous problèmes. Le premier est l'identification des couples de concepts arguments des relations. Le deuxième est l'identification de la relation ou des relations entre ces couples de concepts. Dans les travaux cités, le premier problème est résolu en utilisant une mesure statistique : Le deuxième en utilisant les verbes qui occurrent souvent avec les couples comme labels ou identifiants de relation. Ainsi, la démarche que nous proposons comprend deux étapes :

1. Identifier les couples de concepts (C_i, C_j) qui peuvent être reliés par une relation,
2. Identifier la relation $R_k$qui relie chaque couple (C_i, C_j) .

Identification des couples de concepts :

Nous considérons un document $d_p$du corpus décrit par un ensemble de concepts. Soit C_i et C_j deux concepts appartenant à l'ensemble des concepts de l'ontologie, une relation R_k entre ces concepts est repérée dans ce document, si les deux conditions suivantes sont satisfaites :

1. Dans l'ontologie, il existe une relation R_k entre les concepts C_i et C_j ,

2. Dans le document d_p, il existe une phrase où les concepts C_i et C_j cooccurrent.

Par exemple dans le document suivant :

Patient âgé de 58 ans qui a constitué il y a un an et demi un infarctus du myocarde ancien inaugural.
…….. sont négatives à 120 watts. La coronarographie montre cependant des lésions tritronculaires sévères et une altération marquée de la fonction ventriculaire gauche.
……
A la recherche d'une ischémie silencieuse……

Figure 3: Exemple de document

Les deux concepts « coronarographie » et «lésions » co-occurrent dans la phrase « La coronarographie montre cependant des lésions tritronculaires sévères et une altération marquée de la fonction ventriculaire gauche ». Si dans l'ontologie il existe une relation entre les concepts « coronarographie » et «lésions ». Alors une relation entre ces concepts est repérée dans ce document.

Identification de la relation

En général, dans le processus de construction des ontologies multilingues un ensemble de labels est affecté à chaque relation, un label pour chaque langue. Ces labels sont des verbes, chaque verbe est dans une langue de l'ontologie. L'ensemble de ces labels, c'est-à-dire les termes utilisés pour labéliser les relations forment le sous-vocabulaire V_{OR} de l'ontologie, défini précédemment. Dans l'ensemble V_{OR} on définit la relation L_{mr} qui permet de trouver le label d'une relation de l'ontologie dans une langue donnée et la relation inverse S_{mr}, de la manière suivante

$$\begin{cases} \forall\, r \in R, L_{mc}(r, l_i) = \{t \in V_{Or} \,/\, t \text{ est le label de } r \text{ dans la langue } l_i\} \\ \text{et} \\ \forall\, t \in V_{Or}, S_{mc}(t, l_i) = \{r \in \mathrm{R} \,/\, r \text{ est labelisée par } t \text{ dans la langue } l_i\} \end{cases}$$

Ainsi, pour une relation d'identifiant « r#001 » et de label « montrer » dans la langue française « fr » et de label « show » dans la langue anglaise « en », on aura :

- $\mathrm{L_{mr}}(\text{"r\#001"}, \text{"fr"}) = \{\text{"montrer"}\}$
- $\mathrm{L_{mr}}(\text{"r\#001"}, \text{"en"}) = \{\text{"show"}\}$
- $\mathrm{S_{mr}}(\text{"montrer"}, \text{"fr"}) = \{\text{"r\#001"}\}$
- $\mathrm{S_{mr}}(\text{"show"}, \text{"en"}) = \{\text{"r\#001"}\}$

Pour la phrase « La coronarographie montre cependant des lésions tritronculaires sévères et une altération marquée de la fonction ventriculaire gauche », supposons que dans l'ontologie nous avons la relation "$r\#001$" de label « montrer ». Cette relation est repérée en utilisant l'opérateur S_{mr}.

3 Projet d'évaluation

A court terme nous proposons de valider notre approche d'indexation multilingue. Nous utilisons dans nos expérimentations les données médicales de la campagne

d'évaluation CLEF. Ces données sont constituées d'un corpus et des requêtes. Ce corpus est composé de diagnostics médicaux écrits dans trois langues : le français, l'anglais et l'allemand. Durant notre travail d'expérimentation nous utilisons le méta-thésaurus UMLS[4].

4 Conclusion

Dans cet article nous avons présenté une méthode d'indexation sémantique des documents multilingues. Elle permet d'extraire les concepts et les relations entre les concepts. Notre méthode est fondée sur des mesures statistiques et sur une ressource sémantique externe, l'ontologie multilingue du domaine.
Pour extraire les concepts nous identifions dans une première étape les termes simples et les termes composés, dans une deuxième étape nous utilisons une fonction de correspondance pour passer de la représentation termes à la représentation concepts.
Ainsi, nous avons proposé une méthode de catégorisation des mots en mots vides et mots pleins : les termes simples. Une nouvelle mesure de degré d'association entre les termes est introduite, l'information mutuelle adaptée (IMA). Cette mesure est utilisée pour l'extraction des termes composés. Comparée à la mesure traditionnelle, l'information mutuelle (IM), l'information mutuelle adaptée permet l'extraction des termes composés de longueur supérieure à deux. Une pondération est affectée à chaque terme d'un document donné. Cette pondération est basée sur la mesure CTF*IDF (CTF pour Compound Term Frequency). A l'opposé de la mesure statistique classique, TF*IDF, la mesure introduite, CTF*IDF permet de déterminer la pondération d'un terme composé (de longueur plus que un) dans un document donné.
Aussi nous avons présenté une approche pour décrire les documents par des concepts. Nous avons défini les relations de référence, L_c et L_{mc} , ainsi que leurs relations inverses S_c et S_{mc}. Ces relations sont utilisées pour passer de représentation termes à la représentation en concepts. Durant cette dernière étape nous avons utilisé une ontologie multilingue du domaine. Au sujet de l'ambigüité des termes, nous avons proposé une démarche de désambigüisation. Cette démarche consiste à examiner les termes ambigus dans le contexte où ils apparaissent, le document. Deux types d'ambigüité ont été traités : l'ambigüité langagière et la polysémie.
Pour extraire les relations entre les concepts nous identifions dans une première étape les couples arguments d'une relation et dans une deuxième étape la relation entre ce couple.
Nous signalons que tout au long du processus d'indexation les langues de documents ne sont pas diagnostiquées. Ce processus n'utilise aucune connaissance spécifique à une langue du corpus. La démarche est entièrement automatique et ne nécessite pas d'intervention de l'utilisateur. Ainsi, nous qualifions notre approche de multilingue et d'endogène.

[4] http://www.nlm.nih.gov/research/umls/umlsmain.html

5 References

1.BAZIZ M. Indexation conceptuelle/sémantique guidée par ontologie pour la recherche d'information, Thèse de Doctorat en informatique effectuée à l'Institut de Recherche en Informatique de Toulouse (IRIT), 2005.
2.BAZIZ M., BOUGHANEM M., PASI G., PRADE H. An Information Retrieval Driven by Ontology from Query to Document Expansion . Proceedings of the 8th Conference on Large-Scale Semantic Access to Content (Text, Image, Video and Sound), RIAO 2007 .
3.BERLAND M., CHARNIAK E. Finding parts in very large corpora. In Annual meeting of Association of Computational Linguisitcs, 1999.
4.BOURIGAULT D. LEXTER, a Natural Language Processing tool for terminology extraction. Proceedings of the 7th EURALEX International Congress, Goteborg, 1996 .
5.CALLAN J. P., CROFT W. B., HARDING S. M. The INQUERY Retrieval System. DEXA 1992: 78-83.
6.CIARAMITA M., GANGEMI A., RATSCH E., SARIC J., ROJAS I. . Unsupervised learning of semantic relations between concepts of a molecular biology ontology. In International Joint Conference on Artificial Intelligence, 2005.
7.CONDAMINES A, REBEYROLLES J. Construction d'une base de connaissances terminologiques à partir de textes : expérimentation et définition d'une méthode. In CHARLET J, ZACKLAD M., KASSEL G. & BOURIGAULT D. éds. Ingénierie des connaissanc, 2000 .
8.DAILLE B. Approche mixte pour l'extraction de terminologie : statistiquel exicale et filtres linguistiques. Rapport interne, Université de Paris 7. Thèse de Doctorat en Informatique Fondamentale, 1994.
9.DAILLE B. Study and implementation of combined techniques for automatic extraction of terminology. In J. KLAVANS & P. RESNICK, Eds., The Balancing Act :Combining Symbolic and Statistical Approaches to Language, p. 49–66. MIT Press, 1994.
10.DAVID S., PLANTE P. De la nécessité d'une approche morpho-syntaxique en analyse de textes, dans Intelligence Artificielle et Sciences Cognitives au Québec, vol. 2, no 3, septembre 1990, p. 140-155.
11.ENGUEHARD C. Automatic natural acquisition of a terminology. In Proceedings of the 2nd International Conference of Quantitative Linguistics (QUALICO'94), p. 83–88, Moscow,1995 .
12.FRANTZI K. T., ANANIADOU S., TSUJII J. Classifying Technical Terms , dans Proceedings Third ICCC/IFIP Conference on Electronic Publishing, Ronneby, p. 144-155.
13.GIRJU R., BADULESCU A., MOLDOVAN D. Learning semantic constraints for the automatic discovery of part-whole relations. In Human Language Technologies and North Ameircan Association of Computational Linguisitcs, pages 80-87, 2003.
14.HEARST M. Automatic acquisition of hyponyms from large text corpora. In 14sup th International Conference on Computational Linguistics, 1992.
15.JOUIS C. « Contribution à la conceptualisation et à la modélisation des connaissances à partir d'une analyse linguistique de textes, Réalisation d'un prototype : le système SEEK », Thèse de Doctorat, EHESS, Paris, 1993 .
16.KAVALEC M., MAEDCHE A., SVATEK V. Discovery of lexical entries for non-taxonomic relations in ontology learning. In SOFSEM, 2004.
17.LE PRIOL F . « Extraction et capitalisation automatique de connaissances à partir de documents textuels. SEEK-JAVA : identification et interprétation de relations entre concepts. », Thèse de Doctorat en Informatique, Université Paris-Sorbonne, 2000.
18.LEBART L., SALEM A. Analyse statistique des données textuelles : questions ouvertes et lexicométrie. Paris: Dunod, 1994.
19.LUHN H. The automatic creation of literature abstracts. IBM Journal of Research and Development, Vol 2, N° 2, pp :159–165, 1958.

20. MORIN E. Des patrons lexico-syntaxiques pour aider au dépouillement terminologiques, Traitement Automatique des Langues, volume 40, Numéro 1, pages 143-166, 1999 .
21. Observatoire du Traitement Informatique des Langues et de l'Inforoute, « C - Lexique de l'inforoute et du traitement informatique des langues », http://www.owil.org/lexique/c.htm, consulté en Décembre 2006.
22. PUNURU J. R. Knowledge-Based Methods for Automatic Extraction of Domain-Specific Ontologies. Phd thesis, Louisiana State University, degree of Doctor of Philosophy, 2007.
23. ROBERTSON S. E, WALKER S. . On relevance weights with little relevance information. In Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval, pages 16–24. ACM Press, 1997.
24. ROUSSELOT F., FRATH P. et OUESLATI R. Extracting concepts and relations from corpora, Proceedings workshop on Corpus-Oriented Semantic Analysis, Proceddings of the 12th European Conference on Artificial Intelligence (ECAI'96), 1996.
25. SCHUTZ A., BUITELAAR P. . Relext: A tool for relation extraction from text in ontology extension. In Fourth International Semantic Web Conference, 2005.
26. SEGUELA P., AUSSENAC-GILLES N. Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine, Actes de la conférence Ingénierie des Connaissances (IC'99), pp 79-88, Paris, 1996.
27. SINGHAL A., MITRA M., BUCKLEY C. Learning routing queries in a query zone. In Proceedings of the 20th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Philadelphia, Pennsylvania, United States, July 27 - 31, 1997) .
28. SMADJA F. Retrieving collocations from text: Xtract. Computational Linguistics, 19(1):143-177, 1993 .
29. SPARCK JONES K. Automatic keywords classification for information retrieval. 1971.
30. SPARCK JONES K., VAN RIJSBERGEN C.J. Progress in documentation Journal of Documentation, Vol. 32, Num. 1, Pages 59-75, 1976.
31. TURNEY P. D. Expressing implicit semantic relations without supervision. In 21sup st international conference on computational linguistics, pages 313-320, 2006.
32. VERGNE J. Une méthode indépendante des langues pour indexer les documents de l'internet par extraction de termes de structure contrôlée. In Actes de la Conférence Internationale sur le Document Électronique (CIDE 8), Beyrouth, Liban , 2005.
33. ZIPF G. K. The Psycho-biology of Language. An Introduction to Dynamic Philology. The M.I.T. Press, Cambridge, second paperback printing (first edition : 1935) edition 1968.