Table des matières

Introduction Générale **6**

Chapitre I **8**
La Recherche D'information **8**
1. Introduction 9
2. Définition d'un Système de Recherche d'Information 9
3. Bref Historique de la RI 10
4. Processus de RI 10
4.1. L'indexation 11
4.1.1. Les approches d'indexation 12
4.1.2. Les étapes de l'indexation automatique 12
4.1.3. Le résultat de l'indexation 16
4.2. L'appariement Document/ Requête 16
4.3. La reformulation de la requête 16
4.3.1. La reformulation manuelle 16
4.3.2. La reformulation semi-automatique 16
4.3.3. La reformulation automatique 17
5. Modèles de RI 17
5.1. Les modèles ensemblistes 17
5.1.1. Le modèle booléen 17
5.1.2. Modèle Booléen basé sur des ensembles flous 18
5.2. Les modèles algébriques 19
5.2.1. Le modèle vectoriel 19
5.2.2. Le modèle vectoriel généralisé 21
5.3. Les modèles probabilistes 22
5.3.1. Le modèle de base 22
5.3.2. Le modèle de réseau inférentiel bayésien 23
6. Evaluation d'un système 24
6.1. Corpus de test 25
6.1.1. Les collections TREC 25
6.2. Mesures d'évaluation 26
6.3. Comparaison de systèmes et Précision moyenne 27
7. Relations avec d'autres domaines 27
7.1. La RI et les Bases de Données 27
7.2. La RI et les systèmes question\réponse 28
8. Difficultés de la RI 28
9. Conclusion 29

Chapitre II **30**
La recherche d'information par croisement de langues **30**
1. Introduction 31
2. Quesque la recherche d'information multilingue 31
3. Les différentes approches de l'indexation multilingue 32
3.1. Approche basée sur un vocabulaire contrôlé 33
3.2. Les différentes approches de la traduction (texte libre) 33
3.2.1. Approche basée sur la traduction de la requête 33
3.2.2. Approche basée sur la traduction des documents 34

3.2.3. Approche basée sur le langage pivot 34
4. Les ressources multilingues 35
4.1. Les traducteurs automatiques 35
4.1.1. Les problèmes 35
4.1.2. Exemples de systèmes de traductions 36
4.2. Les dictionnaires 36
4.2.1. Les dictionnaires bilingues 36
4.2.2. Les dictionnaires multilingues 37
4.3. Les Corpus Alignés 38
4.3.1. Exemples de Corpus Alignés 38
4.3.2. Les techniques d'alignement 39
4.4. Les thesaurus 39
4.4.1. Exemples de Thésaurus 39
5. Etat des recherches dans le domaine de la RI multilingue 40
5.1. Approches basées sur le langage pivot 40
5.2. Approches basées sur les traducteurs automatiques 40
5.3. Approches basées sur les dictionnaires 42
5.4. Approches pour la désambigüisation des requêtes 43
5.5. Approches basées sur les Corpus alignés 45
5.6. Approches basées sur le vocabulaire prédéfini 47
6. Les problèmes de la recherche d'information multilingue 49
7. Conclusion 49

Chapitre III 50
Expansion de requête pour un Système de Recherche d'Information par croisement de langues 50
1. Introduction 51
2. Problématique 52
3. Description de l'approche suivie 52
3.1. Prétraitement 53
3.2. Désambigüisation et expansion de la requête 54
3.2.1. La désambiguïsation 54
3.2.2. Expansion de la requête 55
3.3. Traduction de la requête désambiguïsée et étendue 57
3.4. Indexation 57
3.5. Appariement documents-requête 58
4. Expérimentation et évaluation 59
4.1. Environnement d'expérimentation 59
4.1.1. Présentation de WordNet 59
4.1.2. Présentation de NetBeans 59
4.1.3. La base de test 60
4.2. Evaluation 61
4.2.1. Evaluation des stratégies 61
4.2.2. Evaluation finale 63
5. Conclusion 64

Conclusion Générale 66

References bibliographiques 68

Liste des figures

Figure I.1 Processus en U de recherche d'information …….. 11

Figure I.2 Les étapes de l'indexation automatique……. 13

Figure I.3 Courbe générale de précision/rappel…….. 26

Figure II.1 Les différentes approches de l'indexation multilingue……. 33

Figure III.1 Schémas synoptique de la stratégie suivie…… 53

Rapport-gratuit.com
LE NUMERO 1 MONDIAL DU MÉMOIRES

Liste des tableaux

Tableau III.1 Récapitulatif des observations menées …… 57

Tableau III.2 Nombre de mots et de concepts dans la base lexicographique WordNet 2.0…… 59

Tableau III.3 Description de la base de test utilisée…….. 61

Tableau III.4 Comparaison entre l'expansion faite à l'aide des hyperonymes et l'expansion faite à l'aide des hyponymes en termes de précision et de rappel…….. 62

Tableau III.5 Comparaison entre l'expansion faite à l'aide des hyperonymes (niveau 1) et de l'expansion faite à l'aide des synonymes……. 63

Tableau III.6 Impact de la désambigüisation et de l'expansion faite à l'aide des synonymes…… 63

Liste des abréviations

BDD	Bases de données
IA	Intelligence Artificielle
P	Précision
Q/R	Question/Réponse
R	Rappel
RI	Recherche d'Information
RIM	Recherche d'Information Multilingue
SRI	Système de Recherche d'Information
TAL	Traitement Automatique de la Langue

Introduction Générale

L'apparition d'Internet a rendu accessible au public des services variés comme le courrier électronique, la messagerie instantanée et le World Wide Web, ces derniers ont profondément transformé les moyens de communication, notamment en facilitant les échanges de documents entre les pays. Dès lors, les collections de documents se sont enrichies par des documents écrits dans différentes langues. Les systèmes de recherche d'information (SRI) ont dû s'adapter à cette révolution technique pour devenir des systèmes capables de gérer des collections multilingues de documents. La recherche devenant donc multilingue, il faut retrouver tous les documents relatifs à une requête donnée quelque soit leur langue.

Par ailleurs, un SRI multilingue doit aussi faire face au problème de la représentation du contenu des documents ainsi qu'au problème de l'évaluation de la pertinence. Cette évaluation est plus difficile que dans un SRI monolingue, en effet il est difficile de construire une fonction de correspondance avec différents langages pour les documents et la requête.

Problématique : Pourquoi le multilingue ?

Si l'utilisateur ne dispose que de SRI monolingues et qu'il veut récupérer les documents écrits dans d'autres langues, il doit traduire sa requête. Il doit ainsi soumettre autant de requêtes que de langues qu'il souhaite prendre en compte. Cette manipulation est lourde et n'est pas à la portée des usagers qui ne connaissent que leur langue maternelle; de plus la traduction peut conduire à une perte de sens.

L'organisation retenue pour la présentation de nos travaux, s'articule en trois chapitres :

Le premier chapitre traite des généralités concernant le domaine de recherche d'information. Nous présentons en détail les principales approches qui ont fait preuve de performance en recherche d'information ainsi que les modèles universellement appliqués dans la mise en œuvre de ces systèmes. Au final nous avons expliqué la façon d'évaluer ce genre de système.

Le second chapitre présente le principe de la recherche d'information par croisement de langues, nous avons défini en premier lieu la recherche d'information multilingue et la notion de multilinguisme, ensuite nous avons expliqué les différentes approches existantes dans ce domaine puis nous avons présenté un état d'art concernant les travaux élaborés. Nous en avons tiré au final les apports de ce genre de système et les difficultés rencontrés dans ce domaine.

Le troisième chapitre est consacré à la présentation de notre approche qui consiste en la mise en œuvre d'une technique permettant d'améliorer la qualité de la traduction dans un processus de recherche par croisement de langues. L'objectif de notre travail est d'explorer les différentes méthodes utilisées dans la recherche documentaire multilingue afin d'améliorer la description sémantique des requêtes pour que la traduction n'entraîne pas une perte de sens et que les documents retournés par le système répondent aux vrais besoin de l'utilisateur.

Au final, nous dressons un bilan de nos travaux et nous présentons ensuite les perspectives d'évolution de ces travaux.

Rapport-Gratuit.com

Chapitre I

La Recherche D’information

1. Introduction

La recherche d'information est un domaine historiquement lié aux sciences de l'information et à la bibliothéconomie. Elle traite l'information dans la manière de l'organiser et de la façon de la sélectionner, elle peut être définie comme une activité qui dans le but de répondre à une question vise à localiser et à traiter une ou plusieurs informations au sein d'un environnement documentaire complexe.

Le traitement de cet environnement ne peut pas être effectué manuellement et donc l'objectif de la recherche d'information est d'extraire les informations pertinentes vis-à-vis d'une requête pour un utilisateur donné à travers l'utilisation d'un ensemble de programmes informatiques appelés systèmes de recherche d'information.

Dans ce premier chapitre, nous allons définir les concepts de base de la recherche d'information et les systèmes de recherche d'information (SRI).

2. Définition d'un Système de Recherche d'Information

La recherche d'information est la science qui étudie la manière de répondre pertinemment à une requête en retrouvant de l'information dans un corpus.

On définit un SRI comme étant un système permettant de retrouver les documents pertinents à une requête d'utilisateur écrite dans un langage libre, à partir d'une base de documents volumineuse. Dans cette définition, il y a trois notions clés les documents, la requête et la pertinence :

- ***Documents:*** « Le document est l'élément centrale du SRI, c'est un objet complexe sans cesse en évolution car il est lié aux développements des technologies de la communication ».

Un document peut être un texte, un morceau de texte, une page WEB, une image, une bande vidéo, etc. On appelle document toute unité qui peut constituer une réponse à une requête d'utilisateur.

- ***Requête:*** une requête est une façon d'exprimer un besoin en information de l'utilisateur par un ensemble de mots clés, ce besoin est traduit à l'aide d'un langage naturel ou booléen.

- ***Pertinence:*** La pertinence est une notion complexe et un peu floue qui dépend de l'utilisateur et de la requête mais de façon générale, le but de la RI est de retrouver

Rapport-gratuit.com
LE NUMERO 1 MONDIAL DU MÉMOIRES

seulement les documents pertinents et un document pertinent doit contenir l'information que l'utilisateur recherche. C'est sur cette notion que les SRI sont jugés. [Jian-Yun, 01]

3. Bref Historique de la RI

-1940: Apparition des SRI, focalisation de la RI sur les applications dans des bibliothèques.

-1950: Apparition du modèle booléen et l'élaboration de petites expérimentations sur des petites collections de documents.

-1960 et 1970: Apparition du système SMART, Développement d'une méthodologie d'évaluation de système et conception de corpus de test(CACM).

-1980: Développement de l'intelligence artificielle, ainsi on tentait d'intégrer des techniques de l'IA en RI (système expert).

-1990 et 1995: L'apparition d'internet, la RI a été modifié et sa problématique plus élargie.

4. Processus de RI

Le processus de recherche d'informations est fourni par SRI, ce dernier met en correspondance les représentations des informations contenues dans un fond documentaire et des besoins d'utilisateur exprimés par une requête.

Cette notion de pertinence peut être appréhendée à deux niveaux :

- *Niveau utilisateur*: la pertinence correspond à la satisfaction de l'utilisateur par apport à l'ensemble des documents restitués par le SRI.

- *Niveau système:* le système mesure un degré de pertinence, une valeur de similitude entre un document et une requête.

Le but de tout SRI est de rapprocher la pertinence système de la pertinence utilisateur. Pour effectuer de façon efficace cette fonction, le SRI doit réaliser l'indexation des documents, la formulation de la requête, la comparaison requête-documents et enfin la reformulation de la requête (processus non toujours présent mais important). Nous pouvons représenter schématiquement un SRI, comme illustré par la figure I.1, par ce qui est appelé communément le processus en U de recherche d'information.

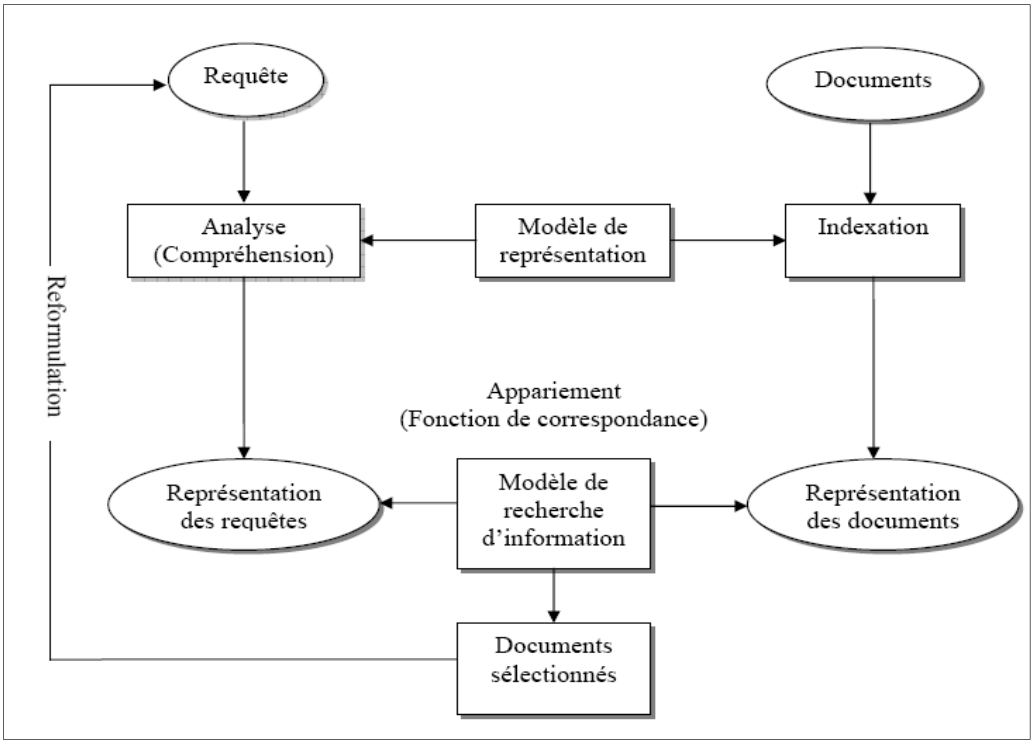

Figure I.1 Processus en U de recherche d'information

Le schéma ci-dessus montre clairement que le processus de recherche d'information se décompose en deux processus comme suit :

- ***Modèle de représentation.***
- ***Modèle de recherche d'information.*** [Boucham, 09]

4.1. L'indexation

Un traitement préliminaire de la base documentaire est nécessaire dans le processus de RI. Il consiste à créer un ensemble de mots clés reflétant aux mieux le contenue sémantique du document, cette liste de mots clés sera plus facilement exploitable lors du processus de la RI. Cette étape est communément appelée l'indexation.[Jian-Yun, 01]

On distingue deux types d'indexation :

–***L'indexation libre (réalisée automatiquement) :*** les termes sont extraits du texte à indexer.

–***L'indexation contrôlée (réalisée manuellement ou semi-manuellement) :*** qui s'appuie sur un ensemble prédéfini de termes : le *thésaurus*. Elle consiste à sélectionner les termes de ce thésaurus qui indexent ce texte.

4.1.1. Les approches d'indexation

✓ ***Indexation manuelle***

C'est le documentaliste ou un spécialiste du domaine qui effectue l'analyse du document, pour identifier son contenu et construire une représentation de ce contenu (choix des mots effectué par des indexeurs).Elle est basée sur un vocabulaire contrôlé (lexique, liste hiérarchiques, thésaurus, ontologie). [Boucham, 09]

✓ ***Indexation semi-manuelle***

Le choix finale revient au spécialiste, qui intervient souvent pour choisir d'autres termes significatifs, l'indexation semi manuelle se divise en deux parties, une partie automatique permettant d'extraire une liste de descripteur, et une deuxième partie qui est manuelle réalisée par un spécialiste du domaine dont la tache est de sélectionner des termes significatifs parmi les descripteurs retournés auparavant. [Merad & Asnoun, 09]

✓ ***Indexation automatique***

C'est le SRI qui génère les indexes des documents. L'indexation automatique a été crée afin de remédier aux problèmes liés aux approches précédentes, elle présente l'avantage d'une régularité du processus, car l'indexation automatique fournit toujours le même index pour le même document, ce qui constitue une qualité du système. En effet, l'indexation automatique pêche par son incapacité à interpréter un texte et son manque d'adaptation à de nouveaux vocabulaires. Il est impossible de trouver dans les documents autre chose que ce que le système peut détecter. [Merad &Asnoun, 09] [Boucham, 09]

4.1.2. Les étapes de l'indexation automatique

Le processus d'indexation automatique (figure I.2) passe obligatoirement par 3 phases, chaque phase pouvant contenir une ou plusieurs étapes selon les usages des utilisateurs, c'est au programmeur de sélectionner les étapes qu'il souhaite intégrer au processus d'indexation automatique du corpus.

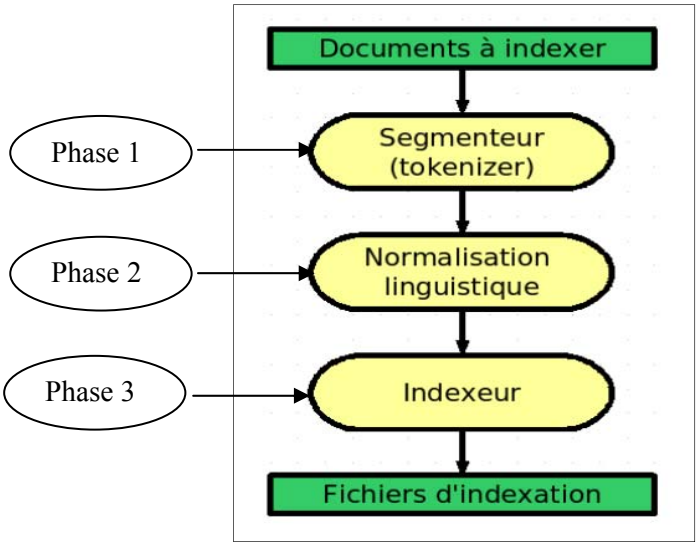

Figure I.2 Les étapes de l'indexation automatique

–***La phase 1:*** cette phase représente la segmentation des documents en unités, cette segmentation est basée en générale sur la ponctuation et sur une liste de séparateur, le résultat de cette étape est un ensemble de mots.

–***La phase 2 :*** cette phase peut contenir plusieurs étapes, dans ce qui suit nous allons expliquer les étapes les plus importantes et les plus utilisées :

- *Elimination des mots vides :* Les mots vides sont des mots qui permettent de lier entre eux les mots d'une phrase pour la structurer (les articles, les conjonctions de coordination, les verbes auxiliaires, etc.). Ces mots ne portent pas de sens, ils ne peuvent pas constituer des index il faut donc les éliminer, cependant il ne faut pas oublier de tenir compte de certain mots vides qui auraient pour homographes des mots significatifs (ex : or).
- *La racinisation* (« stemming » en anglais) : consiste à rechercher la forme tronquée d'un mot à partir de laquelle peuvent être reconstruites ses différentes variantes morphologiques. Cette opération peut être réalisée assez simplement, en utilisant un algorithme comme l'algorithme de Porter, pour l'anglais. Par contre, elle peut entraîner une perte de sens, car la racine extraite peut être commune à des mots se rapportant à des concepts différents : Les mots ***port***, ***port***es (ouverture) et ***port***era ont la même racine ***port*** mais se rapportent à trois concepts différents, la racinisation permet d'augmenter le rappel mais peut diminuer la précision.

- *La lemmatisation :* les mots d'une langue peuvent être classés en deux catégories : les lemmes: formes canoniques (infinitif pour les verbes, singulier pour les noms, etc.) qui constituent en général les entrées dans un dictionnaire de cette langue, les mots obtenus par flexion de ces lemmes: conjugaison d'un verbe, changement de genre ou de nombre, etc. Par exemple, le mot « ***devrait*** » est obtenu par flexion (conditionnel présent, 3e personne du singulier) du verbe « ***devoir*** ».La lemmatisation consiste à remplacer un mot par son lemme, les mots ***port***, ***portes*** et ***portera*** seront remplacés par leurs lemmes : ***port***, ***porter*** ou ***porte*** selon le contexte et *porter***.** C'est une opération plus couteuse que la racinisation car elle nécessite une analyse morphologique et syntaxique des phrases.
- *Extraction des mots composés* : Il est important de reconnaître les mots composés car ce sont des unités de sens. Par exemple : « ***arbre à cames*** » ou « ***pomme de terre*** ».
- *Extraction des entités nommées :* les entités nommées sont des mots ou des groupes de mots qui désignent des personnes, des organisations, des dates, des lieux, etc. Par exemple, si un texte contient l'expression : « ***5 juillet 1962*** » il est plus intéressant de l'indexer globalement par cette date plutôt que les trois termes: « ***5*** », « ***juillet*** » et « ***1962*** ».Si de plus l'indexeur est capable de reconnaître que c'est la date de l'indépendance de l'Algérie, l'indexation sera encore plus précise.

–***L'étape3 :*** Dans cette phase on utilise une approche permettant de sélectionner les index et de leur associer une pondération, cette dernière permet d'assigner aux termes leur degrés d'importance dans les documents, il existe 3 approches pour le choix des index :

- *Approche basée sur la fréquence d'occurrences :* Cette approche consiste à choisir les mots représentants selon leur fréquence d'occurrence. La façon la plus simple consiste à définir un seuil sur la fréquence: si la fréquence d'occurrence d'un mot dépasse ce seuil, alors il est considéré important pour le document. Cependant, en calculant ces fréquences, on s'aperçoit que les mots les plus fréquents sont des mots fonctionnels (mots vides), comme l'explique la loi de Zipf. [Jian-Yun, 01]

- *Approche basée sur la valeur de discrimination :* Par "discrimination", on réfère au fait qu'un terme distingue bien un document des autres documents. C'est-à-dire, un terme qui a une valeur de discrimination élevée doit apparaître seulement pour un petit nombre de documents. L'idée est de garder seulement les termes discriminants, et éliminer ceux qui ne le sont pas. Le calcul de la valeur de discrimination a été développé dans le modèle vectoriel.

Pour calculer la valeur de discrimination d'un terme, on doit comparer une sorte d'uniformité au sein du corpus avec celle du corpus transformé dans lequel le terme en question a été uniformisé (mis au même poids). L'idée est que, si on uniformisant le poids d'un terme dans tous les documents, on obtient une grande amélioration dans l'uniformité du corpus, ce terme était donc très différent (non uniformément distribué) dans différents documents. Il a donc une grande valeur de discrimination. En revanche, si en uniformisant le poids du terme, on n'obtient pas beaucoup d'amélioration sur l'uniformité, ce terme était donc déjà distribué de façon uniforme, donc peu discriminant. [Jian-Yun, 01]

- *Approche basée sur tf*idf :* tf*idf désigne un ensemble de schémas de pondération (et de sélection) de termes. tf signifie "term frequency" et idf "inverted document frequency". Par tf, on désigne une mesure qui a rapport à l'importance d'un terme pour un document. En général, cette valeur est déterminée par la fréquence du terme dans le document. Par idf, on mesure si le terme est discriminant (ou non-uniformément distribué). Ici, on donne quelques formules de tf et d'idf souvent utilisées.

- tf = fréquence d'occurrence du terme dans un document f(t, d);
  - tf = f(t, d) / Max [f(t, d)] où Max [f(t, d)]
  - tf = log (f(t, d))
  - tf = log (f(t, d) + 1)
- idf = log (N/n) où N est le nombre de documents dans le corpus, et n ceux qui contient le terme

Une formule tf*idf combine les deux critères qu'on a vus: l'importance du terme pour un document (par tf), et le pouvoir de discrimination de ce terme (par idf). Ainsi, un terme qui a une valeur de tf*idf élevée doit être à la fois important dans ce document, et aussi il doit apparaître peu dans les autres documents.[Jian-Yun, 2001]

4.1.3. Le résultat de l'indexation

Le résultat d'une indexation est donc un ensemble de termes qui peuvent être soit un mot, soit une racine de mot (soit un terme composé si on possède un mécanisme pour reconnaître des termes composés).

4.2. L'appariement Document/ Requête

La correspondance entre les termes de la requête d'un utilisateur et ceux des documents s'effectue au niveau de l'appariement document-requête, cette étape sert à renvoyer une liste de documents ordonnées selon un degré de pertinence, ce dernier est calculé à partir d'une fonction notée RSV (Q, D) (Retrieval Status Value), ou Q est une requête et D un document. L'expression de la fonction d'appariement est tributaire du modèle de RI choisi. [Merad & Asnoun, 09]

4.3. La reformulation de la requête

Le but essentiel d'un système de recherche d'information est de permettre à l'utilisateur d'avoir un résultat satisfaisant (des documents pertinents) par apport à son besoin exprimé par une requête. La possibilité de reformuler la requête initiale s'avère intéressante dans le processus de la RI. Cela fera en sorte que le résultat retourné soit plus pertinent. Il existe trois méthodes de reformulation de requêtes :

4.3.1. La reformulation manuelle

Elle consiste à présenter à l'utilisateur une liste de documents jugés pertinents en réponse à la requête initiale. C'est à l'utilisateur de sélectionner à partir des documents pertinents ceux dont lesquels le système va extraire les termes à rajouter à la requête initiale dans le but d'effectuer une nouvelle recherche. [Merad & Asnoun, 09]

4.3.2. La reformulation semi-automatique

Cette technique nécessite l'intervention de t'utilisateur qui doit identifier et sélectionner les documents pertinents et les documents non pertinents. [Caro, 97]

4.3.3. La reformulation automatique

L'extension de la requête est faite sans intervention de l'utilisateur grâce à l'utilisation d'un thésaurus contenant des informations de type linguistique (équivalence, association, hiérarchie) et statistique (pondération des termes). [Merad & Asnoun, 09]

5. Modèles de RI

Si c'est l'indexation qui choisit les termes pour représenter le contenu d'un document ou d'une requête, c'est au modèle de leur donner une interprétation. Étant donné un ensemble de termes pondérés issus de l'indexation, le modèle remplit les deux rôles suivants:

- Créer une représentation interne pour un document ou pour une requête basée sur ces termes.

- Définir une méthode de comparaison entre une représentation de document et une représentation de requête afin de déterminer leur degré de correspondance.

Le modèle joue un rôle central dans la RI. C'est le modèle qui détermine le comportement clé d'un système de RI.

5.1. Les modèles ensemblistes

5.1.1. Le modèle booléen

Dans ce modèle, un document est représenté comme une conjonction logique de termes (non pondérés), par exemple, $d = t_1 \wedge t_2 \wedge \ldots \wedge t_n$

Une requête est une expression logique quelconque de termes. On peut utiliser les opérateurs et ($\wedge$), ou ($\vee$) et non ($\neg$). Par exemple: $q = (t_1 \wedge t_2) \vee (t_3 \wedge \neg t_4)$.

Pour qu'un document corresponde à une requête, il faut que l'implication $d \Rightarrow q$. soit valide. Cette évaluation peut être aussi définie de la façon suivante:

Un document est représenté comme un ensemble de termes, et une requête comme une expression logique de termes. La correspondance $R(d, q)$ entre une requête et un document est déterminée de la façon suivante:

Rapport-Gratuit.com

$R(d, t_i) = 1$ si $t_i \in d$; 0 sinon.

$R(d, q_1 \wedge q_2) = 1$ si $R(d, q_1) = 1$ et $R(d, q_2) = 1$; 0 sinon.

$R(d, q_1 \vee q_2) = 1$ si $R(d, q_1) = 1$ ou $R(d, q_2) = 1$; 0 sinon.

$R(d, \neg q_1) = 1$ si $R(d, q_1) = 0$; 0 sinon.

Les principaux problèmes liés à ce modèle sont:

– Les documents retournés comme réponse à une requête par le système seront non-ordonnés. Il n'est pas possible de dire quel document est mieux qu'un autre.

– Il est difficile d'exprimer qu'un terme est plus important qu'un autre dans leur représentation.

– Il est difficile pour les usagers d'exprimer les requêtes car ils manipulent très mal les opérateurs logiques. [Jian-Yun, 01]

5.1.2. Modèle Booléen basé sur des ensembles flous

Ce modèle est une extension du modèle booléen standard, il vise à tenir compte de la pondération des termes dans les documents. Du côté requête, elle reste toujours une expression booléenne classique. Avec cette extension, un document est représenté comme un ensemble de termes pondérés comme suit: $d = \{\ldots, (t_i, a_i), \ldots\}$

L'évaluation d'une requête peut prendre plusieurs formes, dans la première évaluation, les opérateurs logiques ∧ et ∨ sont évalués par min et max respectivement et donc les parties d'une conjonction ou d'une disjonction ne contribuent pas en même temps dans l'évaluation, pour remédier à ce problème une autre évaluation à été proposée, c'est celle de Lukaswicz, dans cette dernière les opérateurs logiques ∧ et ∨ sont évalués par (*) et (+,-,*) respectivement.

Dans ce modèle, le plus important est qu'on peut mesurer le degré de correspondance entre un document et une requête et ainsi, ordonner les documents dans l'ordre décroissant de leur correspondance avec la requête. Au niveau de la représentation, nous pouvons exprimer dans quelle mesure un terme est important dans un document. [Jian-Yun, 01]

5.2. Les modèles algébriques

Nous adoptons ce qui suit, les principales notations suivantes :

Q_k : k ^{ième} requête

D_j : j^{ème} document de la collection

$RSV(Q_k,D_j)$: Valeur de pertinence associée au document *Dj* relativement à la requête Q_k

q_{ki} : Poids d'indexation du terme *ti* dans la requête Q_k

d_{ji} : Poids d'indexation du terme *ti* dans le document D_j

T : Nombre total de termes d'indexation dans la collection

N : Nombre total de documents dans la collection

n_i : Nombre de documents de la collection contenant le terme *ti*

5.2.1. Le modèle vectoriel

Ce modèle préconise la représentation des requêtes utilisateurs et documents sous forme de vecteurs dans l'espace engendré par les N termes d'indexation. De manière formelle, les documents et requêtes sont des vecteurs dans un espace vectoriel de dimension N et représentés comme suit :

$$D_J = \begin{bmatrix} d_{j1} \\ d_{j2} \\ \\ d_{jT} \end{bmatrix} \qquad \boldsymbol{Q}_k = \begin{bmatrix} d_{j1} \\ d_{j2} \\ \\ d_{jT} \end{bmatrix}$$

Sous l'angle de ce modèle, le degré de pertinence d'un document relativement à une requête se traduit par la fonction de pondération.

✓ ***Fonction de pondération***

La fonction de pondération la plus répandue est celle de Sparck Jones & Needham:

$$d_{ji} = tf_{ji} * idf_i$$

Où :tf_{ji} : Décrit le pouvoir descriptif du terme t_i dans le document D_j

idf_i : Décrit le degré de généralité du terme t_i dans la collection

Rapport-gratuit.com
LE NUMERO 1 MONDIAL DU MÉMOIRES

De nombreuses autres fonctions d'indexation sont basées sur une variante du schéma balancé *tf.Idf*, on cite notamment :

– Formule de Salton & Buckley [Salton & Buckley, 88] :

$$d_{ji}=\left(0.5+\frac{0.5*\,\mathrm{freq}_{ij}}{\mathrm{Max}_l\ \mathrm{freq}_{jl}}\right)*\log\frac{N}{n_i}$$

– Formule de Salton & Allan [Salton & Allan, 94] :

$$d_{ji}=\frac{\mathrm{freq}_{ij}}{\sqrt{\sum_{j=1}^{N}\mathrm{freq}_{ji}*\ \log_2{}^2\left(\frac{N}{n_i}\right)}}*\log\frac{N}{n_i}$$

Où : $freq_{ij}$: Fréquence d'apparition du terme *ti* dans le document D_j

Ces mesures supposent que la longueur d'un document n'a pas d'impact sur la mesure de pertinence; or des expérimentations réalisées par Singhal dans [Singhal & al, 97] ont montré que les documents longs ont une plus grande probabilité de pertinence parce que contenant plus de termes d'appariement avec la requête. Les auteurs proposent la fonction suivante :

$$d_{ji}=\frac{tf_{ji}*\log\left(\frac{N\text{-}n_{i+0.5}}{n_{i+0.5}}\right)}{2*\left(0.25+0.75*\frac{|Dj|}{|Dj|}\right)+tfji}$$

Où :

$|D_j|$: Longueur du document D_j

$|D_j|$: Longueur moyenne des documents dans la collection

✓ ***Fonction de similarité***

La fonction de similarité permet de mesurer la ressemblance des documents et de la requête. Les types de mesures les plus répandus sont :

– La mesure du cosinus [Salton, 71] :

$$RSV\ (Q_k, D_j) = \frac{\sum_{i=1}^{T} q_{ki}\ d_{ji}}{\left(\sum_{i=1}^{T} q_{ki}^{\ 2}\right)^{1/2} \left(\sum_{i=1}^{T} d_{ji}^{\ 2}\right)^{1/2}}$$

– La mesure de Jaccard :

$$RSV\ (Q_k, D_j) = \frac{\sum_{i=1}^{T} q_{ki}\ d_{ji}}{\sum_{i=1}^{T} (d_{ji})^2 + \sum_{i=1}^{T} (q_{ki})^2 + \sum_{i=1}^{T} q_{ki}\ d_{ji}}$$

– La mesure de Singhal & al [Singhal & al, 95]:

$$RSV\ (Q_k, D_j) = \frac{\sum_{i=1}^{T} q_{ki}\ d_{ji}}{(1-s) + s * \frac{\sqrt{\sum_{k=1}^{N} d_{kj}^2}}{|D_j|}}$$

Où : $|D_j|$: Longueur du document D_j

s : Constante

Ce modèle présente un inconvénient majeur lié au traitement des termes de documents de manière indépendante. Ceci ne permet pas en effet de reconstituer à travers le processus de recherche, la sémantique associative de termes et ainsi, de la comparer à celle véhiculée par la requête.

5.2.2. Le modèle vectoriel généralisé

Dans le but de pallier au problème d'indépendance des termes, posé par le modèle vectoriel classique, dans [Wong & al, 1985], l'auteur a proposé une nouvelle base de référence pour la représentation des documents et requêtes. A cet effet, il définit sur une collection de termes d'indexation $\{t_1, \ldots, t_t\}$:

–Une base de vecteur binaires, non orthogonaux $\{m_i\}\ i=1..2^T$.

–Un ensemble de min-termes associé à la base; chaque min-terme correspond à l'ensemble de documents comprenant les termes d'indexation positionnés à *1* dans le vecteur de base correspondant.

– Une fonction de pondération $g_i(m_j)$ qui donne le poids du terme t_i dans le min-terme m_j, soit w_{ij}

La base ainsi décrite supporte la représentation de la cooccurrence entre termes. Chaque document et requête est décrit dans la nouvelle base comme suit :

$$D_j = \sum_{i=1}^{T} d_{ji} K_i \qquad Q_k = \sum_{i=1}^{T} q_{ki} K_i$$

$$K_i = \frac{\sum_{\forall r, gi(mr)=1} C_{ir} m_r}{\sqrt{\sum_{\forall r, gi(mr)=1} C_{ir}^{2}}} \qquad C_{ir} = \sum_{d_j / gl(d_j) = gl(mr) \forall l} w_{ij}$$

Le calcul de pertinence *RSV(Q,D)* combine alors le poids des documents w_{ij} et le facteur de corrélation entre termes C_{ir} . Malgré un accroissement du coût de calcul pour la mesure de similarité, relativement au modèle vectoriel classique, le modèle vectoriel généralisé a l'intérêt d'introduire l'idée de considérer la relation entre termes de manière inhérente au modèle de la fonction de pertinence.

Remarque : Il existe cependant d'autres modèles algébriques comme le modèle connexionniste et le modèle LSI.

5.3. Les modèles probabilistes

5.3.1. Le modèle de base

Le modèle de recherche probabiliste utilise un modèle mathématique fondé sur la théorie de la probabilité. Le processus de recherche se traduit par calcul de proche en proche, du degré ou probabilité de pertinence d'un document relativement à une requête. Pour ce faire, le processus de décision complète le procédé d'indexation probabiliste en utilisant deux probabilités conditionnelles :

–***P*** $(w_{ji}$ ***/ Pert) :*** Probabilité que le terme *ti* occure dans le document D_j sachant que ce dernier est pertinent pour la requête

–***P***$(w_{ji}$***/ NonPert)*** **:** Probabilité que le terme t_i de poids d_{ji} occure dans le document D_j sachant que ce dernier n'est pas pertinent pour la requête.

Rapport-gratuit.com
LE NUMERO 1 MONDIAL DU MÉMOIRES

Le calcul d'occurrence des termes d'indexation dans les documents est basé sur l'application d'une loi de distribution (type loi de poisson) sur un échantillon représentatif de documents d'apprentissage.

En posant les hypothèses que :

–La distribution des termes dans les documents pertinents est la même que leur distribution par rapport à la totalité des documents

–Les variables « documents pertinents », « document non pertinent » sont indépendantes, la fonction de recherche est obtenue en calculant la probabilité de pertinence d'un document *D*, notée *P (Pert/D)* [Risjbergen, 79] :

$$P\ (Pert/\ D_j)=\sum_{i=1}^{T} \log \frac{P(w_{ji}\ /Pert)}{P(w_{ji}\ /NonPert)}$$

L'ordre des documents est basé sur l'une des deux méthodes :

–Considérer seulement les termes présents dans les documents et requêtes.

–Considérer les termes présents et termes absents dans les documents et requêtes.

La similitude entre requête et document est calculée comme suit :

$$RSV\ (Q_k\ ,D_j)=c\sum_{i=1}^{T} q_{ki}\ d_{ji}+\sum_{i=1}^{T}\ f_{ij}q_{ki}\ d_{ji}\ \log \frac{N.\ n_i}{n_i}$$

Ou :

$$f_{ji}=\frac{tf_{ji}}{\max tfj}$$

C : Constante

De manière générale, le modèle probabiliste présente l'intérêt d'unifier les représentations des documents et concepts.

5.3.2. Le modèle de réseau inférentiel bayésien

Un réseau bayésien est un graphe direct acyclique où les nœuds représentent des variables aléatoires et les arcs des relations causales entre nœuds.

Ces derniers sont pondérés par des valeurs de probabilités conditionnelles. Le travail original en recherche d'information et basé sur le modèle des réseaux bayésiens, est développé par Turtle [Turtle & Croft, 91] :

Dans l'espace défini par les termes d'indexation, on définit :

T : Variables aléatoires binaires $t_1, \ldots, t_T$ associés aux termes d'indexation

D_j : Variable aléatoire associée à un document

Q_k : Variable aléatoire associée à une requête

On calcule alors la mesure de pertinence de Q_k relativement à D_j en traitant les probabilités conditionnelles de Bayes selon la formule :

$$RSV(Q_k, D_j) = 1 - P(Q_k \wedge D_j)$$

Ou :

$$P(Q_k \wedge D_j) = \sum_{i=1}^{T} P(Q_k/t_i) * \left(\prod_{t_i \in D_j} P(t_i/D_j) * \prod_{t_i \notin D_j} P(\bar{t}_i/D_j)\right) * P(D_j)$$

Avec :

$P(Q_k/t_i)$: Probabilité que le terme t_i appartienne à un document pertinent de Q_k

$P(t_i/D_j)$: Probabilité que le terme t_i appartienne au document D_j sachant qu'il est pertinent.

$$P(t_i/D_j) = 1 - P(t_i/D_j)$$

$P(D_j)$: Probabilité d'observer D_j

Les probabilités conditionnelles de chaque nœud sont calculées par propagation des liens de corrélation entre eux. Ce modèle présente l'intérêt de considérer la dépendance entre termes mais engendre une complexité de calcul importante.

6. Evaluation d'un système

La qualité d'un système doit être mesurée en comparant les réponses du système avec les réponses idéales que l'utilisateur espère recevoir. Plus les réponses du système correspondent à celles que l'utilisateur espère, mieux est le système.

L'évaluation constitue donc une étape importante lors de la mise en œuvre d'un modèle de recherche d'information puisqu'elle permet de paramétrer le modèle,

d'estimer l'impact de chacune de ses caractéristiques et enfin de fournir des éléments de comparaison entre modèles.

6.1. Corpus de test

Pour arriver à une telle évaluation, on doit connaître d'abord les réponses idéales de l'utilisateur. Ainsi, l'évaluation d'un système se fait à l'aide d'un corpus de test.

Dans un corpus de test, il y a:

–un ensemble de documents.

–un ensemble de requêtes.

–la liste de documents pertinents pour chaque requête.

Pour qu'un corpus de test soit significatif, il faut qu'il possède un nombre de documents assez élevé. Pour la construction d'un corpus de test, les jugements de pertinence constituent la tâche la plus difficile. [Jian-Yun, 01]

Différentes collections de test sont utilisées en recherche d'information. Parmi elles nous citons :

6.1.1. Les collections TREC

Le projet TREC est un programme international initié au début des années 90 par le NIST (National Institute of Standards and Technology) et du DARPA (Defense Advanced Reserach Projet Agency). Ce programme offre des moyens homogènes d'évaluation des systèmes de recherche d'information. Il est devenu la référence en recherche d'information pour diverses raisons. En effet, il a permis de définir les tâches en recherche d'information et de construire de larges collections de test.

Dans ce qui suit, nous allons définir les différents éléments qui constituent le projet TREC :

–***Tâches***: L'objectif est de permettre l'évaluation d'approches spécifiques en recherche d'information concernant le filtrage, le croisement de langues, la recherche dans de très large corpus (100 giga octet et plus), les modèles d'interactions.

–***Les participants:*** 25 groupes ont participé à la première édition de TREC en 1992 et 66 groupes de 16 pays différents ont également participé à TREC8.

–***Source d'information:*** Les documents de la collection sont issus de la presse écrite en 1999 (Financial Time, Résumés de publication USDOE, SAN jose Mercury news, etc.).

–***Structure et principe de construction de la collection***: un document TREC est généralement présenté sous le format SGML. Il est identifié par un numéro et décrit par un auteur, une date de production et un contenu textuel.Une requête TREC est également identifiée par un numéro. Elle est décrite par un sujet générique, une description brève et une description étendue sur les caractéristiques des documents pertinents associés à la requête.

6.2. Mesures d'évaluation

La comparaison des réponses d'un système pour une requête avec les réponses idéales nous permet d'évaluer les deux métriques suivantes:

$$\boldsymbol{Précision = \frac{\#documents\ pertinents\ retrouvés}{\#documents\ \ retrouvés}}$$

$$\boldsymbol{Rappel = \frac{\#documents\ pertinents\ retrouvés}{\#documents\ \ pertinents\ dans\ la\ base}}$$

En générale, on peut obtenir un taux de précision et de rappel aux alentours de 30%. Les deux métriques ne sont pas indépendantes. Il y a une forte relation entre elles: quand l'une augmente, l'autre diminue. Il ne signifie rien de parler de la qualité d'un système en utilisant seulement une des métriques. Ainsi, pour un système, on a une courbe de précision-rappel qui a en général la forme suivante (Figure I. 3): [Jian-Yun, 01]

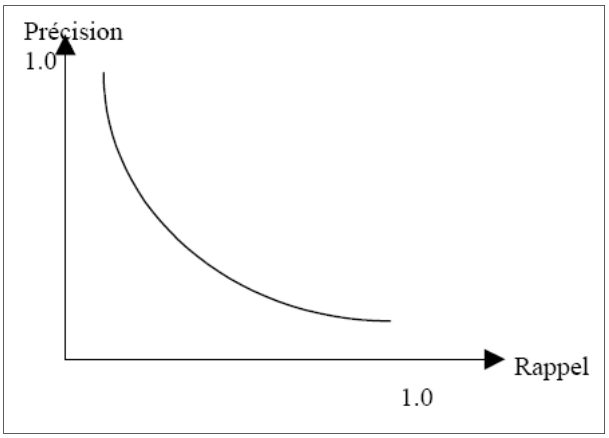

Figure I.3 Courbe générale de précision/rappel

6.3. Comparaison de systèmes et Précision moyenne

Si on veut comparer deux systèmes de RI, il faut les tester avec le même corpus de test (ou plusieurs corpus de test). Un système dont la courbe dépasse (c'est-à-dire qu'elle se situe en haut à droite de) celle d'un autre est considéré comme un meilleur système.

Il arrive parfois que les deux courbes se croisent. Dans ce cas, il est difficile de dire quel système est meilleur. Pour résoudre ce problème, on utilise aussi la précision moyenne comme une mesure de performance. La précision moyenne est une moyenne de précision sur un ensemble de points de rappel. Cette dernière est possible seulement avec la polarisation.

Pour comparer deux systèmes ou deux méthodes, on utilise souvent l'amélioration relative qui est calculée comme suit :

Amélioration de méthode 2 sur méthode 1 = (performance de méthode 2 – performance de méthode 1) / performance de méthode 1. [Jian-Yun, 01]

7. Relations avec d'autres domaines

7.1. La RI et les Bases de Données

On peut imaginer un système de RI comme un système de BDD textuelles. Dans la RI, une partie des spécifications de documents est structurée, notamment les attributs externes. Cette partie peut être organisée assez facilement comme une relation en BDD, et ainsi utiliser des SGBD existants pour rechercher des documents selon des critères sur les attributs externes. Cette partie ne représente pas le cœur de la RI, ce dernier se situe dans la recherche selon le contenu.

Après l'indexation de document, la connexion entre la recherche d'information et les bases de données devient plus étroite. Le résultat de l'indexation est d'associer à chaque document un ensemble d'index. Ce dernier peut être vu comme une relation en BDD. L'inconvénient est que les sélections ne retournent qu'un ensemble de documents sans ordonnancement. [Jian-Yun, 01]

Rapport-Gratuit.com

7.2. La RI et les systèmes question\réponse

Un système de Q\R permet de répondre aux questions relatives à un petit domaine. Pour cela, il faut qu'on crée une modélisation du domaine d'application dans lequel les concepts ou objets sont reliés par des relations sémantiques. Ce modèle permettra de retrouver le concept ou l'objet et ainsi donner une réponse directe à la question.

Cependant dans la RI, c'est une réponse indirecte à une question: on identifie les documents dans lesquels l'utilisateur peut trouver des réponses directes à sa question.

Il y a des tentatives pour rapprocher la RI des systèmes Q/R. Dans certains contextes très spécialisés, la recherche d'information incorpore une base de connaissances. Elle utilise aussi des raisonnements pour déduire si un document peut être pertinent ou pas.

Une tentative plus restreinte consiste à raffiner la notion de document dans la réponse: au lieu de fournir un document complet comme une réponse, on essaie d'identifier un passage dans le document (passage retrieval). C'est une étape qui diminue un peu la distance entre la RI et la Q/R. [Jian-Yun, 01]

8. Difficultés de la RI

–***Difficultés d'accès, couverture, temps de traitement***.

–***Difficultés de définition de la pertinence***.

–***Difficulté de définition du besoin de l'utilisateur :***

- Le besoin d'information formulé par une requête est généralement tellement vague et imprécis que l'objet de la recherche d'information est à priori inconnu.
- La perte d'information entre la réalité du besoin d'information et son expression.

–***Difficulté du langage naturel (Implicite, redondant, ambigu).***

–***Difficulté d'extraction de l'information***.

–***Difficultés d'exploitation de l'information*** : Les documents pertinents ne sont pas nécessairement dans la langue de la requête :

- Les corpus contiennent de plus en plus de document écrit dans différentes langues.
- Un utilisateur qui soumet une requête dans une langue pourrait aussi être intéressé par des documents dans d'autres langues. D'où l'apparition de la Recherche d'Information Multilingue (RIM). [Boucham, 09] [Tamine, 00]

9. Conclusion

Ce premier chapitre a porté essentiellement sur l'étude des SRI de manière générale, nous avons présenté l'architecture commune à tous les systèmes de recherche d'information notamment l'appariement document/requête et la reformulation des requête puis nous avons présentés les étapes essentielles d'une bonne indexation et enfin les différents modèles et stratégies utilisés lors de la mise en œuvre d'un SRI.

Il en ressort que chacun de ces modèles ou stratégies contribue en partie à la résolution des problèmes inhérents à la recherche d'information : perception du besoin en information, représentation du sens véhiculé par les documents, formalisation de la pertinence etc.…

A l'issue de cette étude et en vue des difficultés rencontrées actuellement dans le domaine de RI comme la disponibilité de documents pertinents mais dans plusieurs langues nous allons nous intéresser dans le chapitre qui suit à la conception de système de recherche d'information multilingue.

Rapport-gratuit.com
LE NUMERO 1 MONDIAL DU MÉMOIRES

Chapitre II

La recherche d'information par croisement de langues

1. Introduction

L'utilisation d'une langue universelle, vieux rêve philosophique, semble être encore pour longtemps une utopie. La multitude de langues actuellement présentes sur notre planète restera encore une source de problèmes pour tous ceux désirant trouver des informations. Ces problèmes apparaitront qu'elle que soit la langue dans laquelle celles-ci s'expriment.

Avec le développement d'Internet au niveau mondial, les échanges de documents s'intensifient entre les pays, les cultures et par conséquent, les corpus contiennent de plus en plus de document écrit dans différentes langues, la recherche devient alors multilingue et doit retrouver tous les documents concernés par un besoin d'information.

La sélection d'informations pertinentes est donc confrontée à un double problème : le premier, spécifique a la RI, réside dans la capacité du SRI à séparer les informations pertinentes de celles qui ne le sont pas. Le second, lié au multilinguisme, correspond à la capacité du système d'aller au delà de la langue de la requête en sélectionnant des informations pertinentes écrites dans des langues autres que celle de la requête. Ce second point est communément appelé recherche d'information multilingue.

2. Quesque la recherche d'information multilingue

La recherche d'information multilingue (RIM) est un type de recherche qui permet de repérer l'information lorsque la langue des requêtes est différente de la langue des documents repérés. Un utilisateur peut présenter une requête dans sa propre langue et le système retrouve des documents dans une autre langue.

L'objectif principal de la RIM est de fournir des outils à l'utilisateur qui serait intéressé par l'obtention de documents dans d'autres langues que sa langue maternelle. L'utilisation d'un système de recherche monolingue peut s'avérer fort problématique pour l'usager lorsqu'il effectue une recherche dans une langue qui ne lui est pas familière. La recherche d'information multilingue tente donc d'apporter une solution à ce problème qui devient de plus en plus préoccupant, depuis l'avènement d'Internet et de son contenu multilingue [Nassr, 02] [Boucham, 09] [Harrathi, 09].

La notion de multilinguisme est utilisée selon des sens différents :

- ✓ ***Requête multilingue : Interrogation monolingue de plusieurs bases de documents monolingues***

Ce genre de système, s'apparente à une recherche monolingue, cependant le processus de recherche est capable de traiter des requêtes dans différentes langues. Le corpus est découpé en bases documentaires monolingues, indépendantes les unes des autres. Les documents de chacune des bases ne peuvent être retrouvés que par une requête dans leur langue. [Boucham, 09] [Nassr, 02]

- ✓ ***Document multilingue : Interrogation de documents multilingues***

La recherche s'effectue sur des documents multilingues où des parties du document sont écrites dans des langues différentes. Par exemple, à partir d'une requête dans une langue donnée on peut retrouver des documents multilingues dont le résumé est écrit en français et le corpus de texte en arabe. [Boucham, 09] [Nassr, 02]

- ✓ ***Base multilingue de documents: Interrogation multilingue de plusieurs bases de documents monolingues***

A partir d'une requête dans une langue donnée, on peut retrouver des documents écrits dans chacune des langues du corpus.

La spécificité de ce système réside dans sa capacité à sélectionner des documents écrits dans une langue différente de celle de la requête. Ce genre de systèmes Porte le nom « Cross-Language Information Retrieval » (CLIR) appelé aussi système de recherche d'information par croisement de langues. [Boucham, 09] [Nassr, 02]

3. Les différentes approches de l'indexation multilingue

Dans un contexte multilingue, la requête n'est pas écrite dans la langue des documents. La représentation de la requête est alors dans un espace d'indexation différent de l'espace d'indexation des documents. Dans ce contexte et afin de rendre une recherche documentaire possible il est nécessaire d'utiliser les mêmes descripteurs pour décrire la requête et les documents. Ceci est possible en procédant soit par indexation en langage contrôlé soit par une indexation en texte libre, comme le montre la figure suivante :

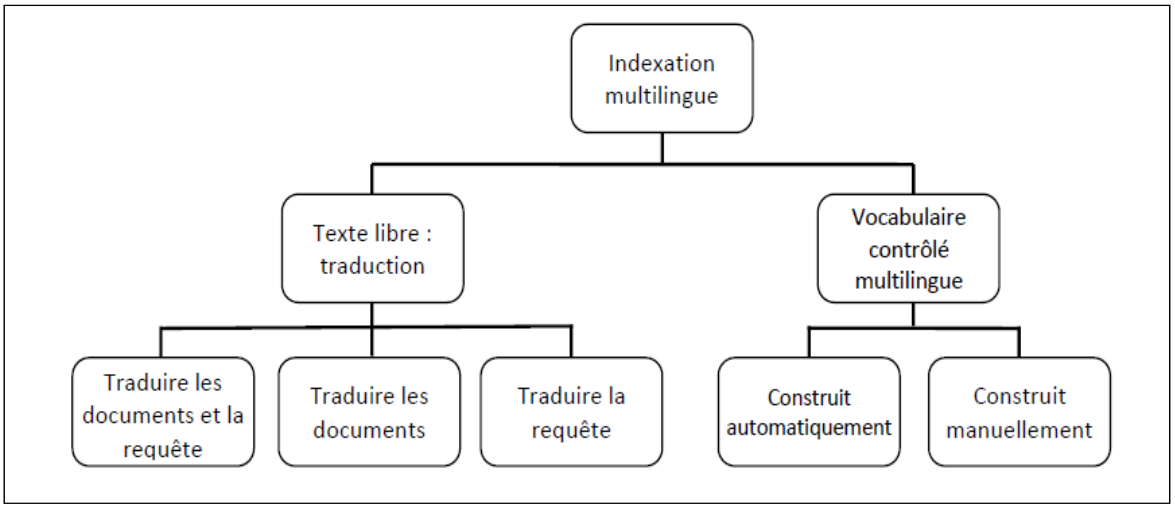

Figure II.1 Les différentes approches de l'indexation multilingue

3.1. Approche basée sur un vocabulaire contrôlé

Dans cette approche la liste des descripteurs est établie préalablement avant l'indexation et elle est utilisée pour indexer les documents et la requête. L'élaboration de cette liste peut être automatique ou manuelle.

L'avantage de cette approche et qu'elle permet d'assister l'utilisateur lors de sa recherche. Cependant, sa difficulté réside dans l'automatisation de l'information de ce vocabulaire. C'est pourquoi elle nécessite souvent une exploitation et une intervention humaine.

3.2. Les différentes approches de la traduction (texte libre)

La plupart des solutions proposées aujourd'hui ont adopté la traduction des documents et/ou des requêtes comme moyen pour mettre ces documents et ces requêtes dans un même référentiel. Nous présentons dans ce qui suit, les différentes approches de la traduction.

3.2.1. Approche basée sur la traduction de la requête

Les entités d'indexation appartiennent à la langue du corpus c'est-à-dire que les requêtes doivent être traduites dans la langue du corpus avant d'effectuer l'indexation.

Cette approche est souvent préférée par les chercheurs en RIM puisqu'il s'agit d'un moyen efficace et peu coûteux, étant donné qu'en général, les requêtes sont composées

Rapport-gratuit.com
LE NUMERO 1 MONDIAL DU MÉMOIRES

de mots simples et sont souvent plus courtes et moins complexes à traduire que les documents. Un autre avantage de la traduction des requêtes est qu'il est possible d'intégrer un traducteur de requêtes sans modifier un système de recherche déjà existant.

Mais la difficulté que l'on rencontre pour traduire une requête est le manque de contexte, ce qui conduit à des interprétations erronées. Cette difficulté réside dans le fait que les requêtes sont généralement composées de quelques mots. Ce manque de contexte est souvent créateur d'ambiguïté ce qui diminue les chances de trouver les bonnes traductions des termes de la requête et par conséquent augmente le bruit généré par le SRI. [Boucham, 09] [Radwan, 94] [Harrathi, 09].

3.2.2. Approche basée sur la traduction des documents

La traduction du texte de la base dans les différentes langues des interrogateurs est une des possibilités pour satisfaire le besoin des interrogateurs.

Le principal avantage de cette approche, est d'offrir une plus grande précision de recherche apparente puisqu'un texte plus long pose moins de problème de polysémie lors de la traduction. Donc, il permet de compenser le manque de contexte de l'approche précédente.

Cette technique peut s'avérer longue et forte coûteuse, sachant que le volume de la documentation produite ne cesse de s'accroître, il s'ajoute à cela le coût du stockage des textes dans les différentes langues d'interrogation [Radwan, 94][Boucham, 09].

3.2.3. Approche basée sur le langage pivot

Le langage pivot se veut une représentation universelle dans un langage formel indépendant de la représentation de la langue du corpus et de la requête. L'analyseur transforme le texte en langue pivot, le texte sera ensuite généré à partir du langage formel dans la langue cible.

Cette approche implique donc une traduction des documents et de la requête dans ce langage pivot. La première remarque est évidente : l'analyse linguistique est donc deux fois plus lourde car il faut traduire le corpus et les requêtes. Par contre, le passage à une recherche d'information vraiment multilingue (collection de document incluant au moins deux langues) est facilité. [Radwan, 94] [Boucham, 09]

4. Les ressources multilingues

Une plus grande partie des approches proposées par le croisement de langues en recherche d'information se basent sur des ressources comme par exemple (Les corpus, dictionnaires, thesaurus).Il est clair que le développement et la disponibilité de ces ressources sont des points cruciaux en recherche d'information par croisement de langues. En effet, la qualité des résultats des systèmes à croisement de langues est fortement liée à la qualité des ressources utilisées.

Une référence globale pour retrouver toutes les ressources utilisées par un très grand nombre d'utilisateurs est le site Web de Douglas Oard[1]. [Nassr, 02]

4.1. Les traducteurs automatiques

L'utilisation d'un logiciel de traduction automatique est l'approche la plus directe. Ces systèmes "Machine Translation " sont utilisés pour obtenir différentes versions d'un même texte dans plusieurs langues or le but d'un système de traduction automatique est de produire une version lisible et fiable dans la langue cible du texte source.

4.1.1. Les problèmes

–***Choix incorrect de la traduction du mot ou terme :*** Un traducteur automatique doit forcément lever toutes les ambiguïtés pour ne fournir qu'une unique version de la traduction, ce qui génère parfois des choix incorrects. Malheureusement, les SRI sont plus pénalisés par ce dernier que par la persistance d'une ambiguïté

–***Syntaxe incorrecte :*** Un traducteur automatique doit fournir en sortie un texte grammaticalement correct. Or les SRI sont plus sensibles à des traductions sémantiquement correctes qu'à des constructions syntaxiquement valides. Un traducteur remplit donc des tâches pas nécessairement utiles pour le SRI.

–***Traduction des mots inconnus :*** La ressource linguistique doit être adaptée aux vocabulaires du corpus et des requêtes ou des documents sinon leurs termes ne seront

[1]Site : (http://www.ee.umd.edu/medlab/mlir/)

pas reconnus dans l'étape de traduction et ils ne seront pas pris en compte, ni par le processus d'indexation pour représenter un document, ni par la fonction de comparaison pour comparer les représentations.

4.1.2. Exemples de systèmes de traductions

Quelques systèmes de traductions sont offerts par Lernout et Hauspi[2], Alis Technolgies[3] et aussi par Schauble et McNamee.

–***Systran***[4] est un fournisseur de service de traduction automatique facilitant la communication pour 36 paires de langue et dans 20 domaines spécialisés.

–***CAT2*** est un système de traduction automatique multilingue qui a pris ses racines dans le projet EUROTRA.

–***Reverso*** permet de traduire des documents peu importe leur taille ou leur format (Excel, Word, etc.), tout en respectant la mise en page.

–***Logomedia*** offre un système de traduction automatique des documents, sites Web, courriers électroniques, etc. [Nassr, 02] [Boucham, 09]

4.2. Les dictionnaires

Recueil des mots d'une langue, des termes d'une science, d'un art, rangé par ordre alphabétique, avec leurs significations. Un dictionnaire de la langue indique la définition, l'orthographe, les sens et les emplois des mots d'une langue. Parallèlement au dictionnaire monolingue, il existe des dictionnaires bilingues voir multilingue qui donnent des traductions des mots d'une langue vers une ou plusieurs langues étrangères.

4.2.1. Les dictionnaires bilingues

A titre d'exemple, nous citons la série de Collins des dictionnaires bilingues (Anglais-Espagnol, Anglais-Italien, Anglais-Français, etc.). Internet propose aussi

---

[2] Site : (http://www.lhs.com/itranslator)

[3]Site :(http://www.alis.com/LangSol/whatGIT.html)

[4]Site: (http://www.systran.fr/)

différents dictionnaires bilingues, nous citons BABYLONE[5] et FreeDict[6]. D'autres dictionnaires bilingues sont également disponibles : SunRecom[7], Seasite[8], et enfin Freelang. Les dictionnaires bilingues posent souvent des problèmes, en effet :

– le dictionnaire ne contient pas tous les mots possibles retrouvés dans un texte, certains termes sont explicites mais ils ne sont pas nécessairement dans un dictionnaire, car l'utilisateur humain est capable de dériver automatiquement ses formes

– le dictionnaire contient des définitions longues, avec beaucoup de bruit. En tenant compte de chaque traduction possible, on augmente ainsi le bruit de recherche documentaire;

–le problème de couverture;

– le dictionnaire ne contient pas les noms propres (noms des pays …).

4.2.2. Les dictionnaires multilingues

Il est beaucoup plus difficile de trouver des exemples de dictionnaires multilingues[9] plutôt que les dictionnaires bilingues. Memodata est un de ces dictionnaires. C'est un lexique général multilingue pour le Français, l'Italien, l'Espagnol, l'Anglais et l'Allemand où les termes sont reliés par sens. Le manque de dictionnaires multilingues entrave énormément le développement de nouveaux systèmes de croisement de langues. [Nassr, 02] [Boucham, 09]

---

[5] Disponible en plusieurs langues à l'adresse :(http://www.babylon.com)

[6] Site: (http://www.freedict.com/)

[7] Site : (http://sunrecomgen. Univ rennes1.fr/FR-Eng.html)

[8] Site : (http://www.seasite.niu.edu/Thai/home-page/onlinethai-dictionaries.html)

[9]On peut trouver quelques dictionnaires multilingues aux adresses suivantes : (http://www.emich.edu/linguist/dictionaries.html), (http://www.june29.com/).

Rapport-Gratuit.com

4.3. Les Corpus Alignés

Un corpus est un ensemble de documents exprimés dans une langue source $L1$ alignés avec d'autres documents exprimés dans une langue cible $L2$. Il peut être utilisé pour :

– l'extraction automatique de l'information,
– la désambiguïsation des différentes alternatives du dictionnaire,
– la construction du thesaurus de similarité,
– l'évaluation.

Les corpus parallèles sont des ensembles de textes traduits équivalents, constitués généralement du texte source et d'une ou plusieurs traductions, les corpus comparables sont plutôt des textes pouvant être associés, non pas par leur traduction mais plutôt par leurs caractéristiques communes (même sujet, par exemple).

4.3.1. Exemples de Corpus Alignés

–***Multext*** est développé sur une base volontaire et est disponible au grand public à des fins non commerciales ni militaires.

–***MultiTrans*** est un corpus multilingue plein texte intégré à une infrastructure évoluée de gestion terminologique. MultiTrans se veut un outil de soutien à la traduction.

–***European Corpus Initiative Multilingual (ECI)*** est un corpus contenant plus de 98 millions de mots assurant la couverture de la majorité des langues européennes, de même que la langue turque, le japonais, le chinois, le malais et plusieurs autres. Cependant le corpus le plus utilisé pour les recherches en croisement de langues est HANSARD[10]. Il existe également un autre corpus aligné appelé WAC (Word-wide-web Aligned Coprus).

[10] Il s'agit d'un corpus parallèle (Anglais- Français) pour les textes parlementaires site :(http://morph.ldc.upenn.edu/).

4.3.2. Les techniques d'alignement

–***L'alignement par phrases* :** il s'agit de mettre en correspondance des phrases d'une langue $L1$ avec d'autres phrases d'une langue $L2$

–***L'alignement de mots et expressions* :** il s'agit de repérer les mots et expressions du texte source et du texte cible, puis de les mettre en correspondance.

–***Autres types d'alignement* :** il s'agit de l'alignement de segments linguistique supérieurs aux mots ou termes, et inférieurs à la phrase. [Nassr, 02] [Boucham, 09]

4.4. Les thesaurus

Un thesaurus est une structure qui contrôle les complexités de la terminologie dans un langage. Il fournit aussi des relations conceptuelles idéales à travers la classification. Plus particulièrement, dans un thesaurus multilingue, la relation d'équivalence inclue dans l'ensemble des termes choisis comme représentant du concept, la traduction et les synonymes de ces termes. La relation d'association permet d'améliorer la traduction des expressions. En effet, au lieu d'effectuer une traduction mot à mot en considérant les termes comme indépendants, il faut d'abord chercher à les relier par des relations d'associations pour obtenir la traduction exacte d'un concept multi terme.

4.4.1. Exemples de Thésaurus

–***Le thesaurus INIS*** est spécialisé dans le domaine de l'énergie nucléaire. Il existe en quatre langues : Français, Anglais,Allemand et Russe.

–***EuroWordNet***[11] est une base de données multilingue extraite à partir de WordNet et étendue pour inclure d'autres langues. Ce thesaurus est constituée de relations sémantiques établies entre les différents termes des diverses langues.

[11] Il peut être obtenu à partir d'ELRA. Site : (http://www.icp.ox.ac.uk/)

Rapport-gratuit.com
LE NUMERO 1 MONDIAL DU MÉMOIRES

5. Etat des recherches dans le domaine de la RI multilingue

En recherche d'information la notion de multilinguisme peut se présenter sous différentes facettes [Oard, 96**].** La facette à laquelle nous nous intéressons est la recherche d'information par croisement de langues ou cross-language Information Retrieval (CLIR) [Greffenstette, 98]. Dans ce contexte la, différentes approches ont été proposées :

5.1. Approches basées sur le langage pivot

Les approches basées sur le langage pivot impliquent une traduction des documents et de la requête dans un langage pivot, cela peut s'avérer long et couteux car l'analyse linguistique est deux fois plus lourde puisqu'il faut traduire le corpus et les requêtes.

Le langage pivot se veut une représentation universelle dans un langage formel indépendant de la représentation de la langue du corpus et de la requête, et cela peut s'avérer comme seule alternative au problème de ressources non disponibles dans les paires de langues.

Dans [Nassr, 02], l'auteur a utilisé la traduction basée sur le langage pivot, cette dernière a remarqué que les résultats obtenus ne sont pas bons par rapport aux tests de la base de comparaison (test monolingue et test du dictionnaire). Par contre les résultats sont améliorés, dans le cas où cette technique est combinée à une des stratégies de désambiguïsation telle que la traduction bidirectionnelle.

5.2. Approches basées sur les traducteurs automatiques

Ces approches nécessitent l'intégration d'un logiciel de traduction automatique dans le SRI [Radwan, 94]. Les systèmes basés sur les traducteurs automatiques (Machine Translation (MT)) sont utilisés pour obtenir un même texte dans plusieurs langues avec ou sans l'aide d'un expert. Ces systèmes sont généralement plus complexes et loin d'être parfaits, car ils s'appuient sur des grammaires et autres méthodes linguistiques, même s'ils donnent des résultats satisfaisants pour la traduction des documents, leur utilisation pour la traduction de requêtes n'a pas connu le même

sucés, du fait que ces dernières, sont souvent courtes et exprimées par des mots indépendants. Certaines de leurs fonctionnalités semblent combler les attentes d'un système de recherche d'information par croisement de langues, cependant, certaines d'entre elles sont pénalisantes pour la RI.

Dans [Yamabana & al, 98 ; Oard & al, 96 ; Gey & al, 97], les travaux basés sur la traduction automatique de requêtes ont montré des performances plus faibles que d'autres techniques. Ceci est dû au fait que la requête est souvent une liste de mots dépourvue de sémantique. Dans ce cas précis, les traducteurs automatiques ne produisent pas de bonnes traductions [Pirkola, 98].

Dans [Gey & al, 97], les auteurs ont utilisé le traducteur automatique Globalink pour traduire les requêtes dans le cadre de la tâche de croisement de langues de TREC 6. L'absence de certaines paires de langues dans les lexiques de Globalink, les a obligés à utiliser l'Anglais comme langage pivot intermédiaire entre les différentes langues. Ils ne font aucun commentaire sur l'impact de ce processus sur leurs résultats.

Dans [Oard, 98], l'auteur a également comparé dans le cadre de TREC 7, la traduction des requêtes par le traducteur automatique Logos à celle basée sur les dictionnaires. Il a montré que la technique basée sur les traducteurs automatiques est clairement moins efficace que celle basée sur les dictionnaires pour la traduction des requêtes courtes.

Dans [Yamabana & al, 98], les auteurs ont également utilisé le traducteur automatique Kana-Kanji pour traduire les requêtes. Ils concluent que les traducteurs automatiques sont peu adaptés pour la traduction des requêtes, puisque les requêtes sont rarement exprimées par des phrases et plus souvent par des termes indépendants.

Du fait que les traducteurs automatiques sont loin de produire des traductions de requêtes de bonne qualité [Kay, 95], les travaux élaborés dans [Savoy, 02] proposent une méthode permettant d'améliorer la traduction d'une requête (anglais vers d'autres langues) en utilisant le système de traduction automatique SYSTRAN d'une part, et d'autre part, le dictionnaire bilingue BABYLON, concernant le dictionnaire, la proposition de Savoy émet l'hypothèse que la meilleure traduction est toujours présentée comme premier choix dans le dictionnaire. L'utilisation combinée de ressources pour la traduction de la requête présente une performance intéressante comparée à celle obtenue en recourant à un seul outil.

5.3. Approches basées sur les dictionnaires

Les dictionnaires bilingues tels que ceux développés par les humains sont actuellement la forme la plus répandue des structures ayant une couverture suffisante pour réaliser les applications de croisement de langues, c'est pour cela que les méthodes basées sur les dictionnaires sont les plus utilisées dans la recherche d'information par croisement de langues. Contrairement aux systèmes de traduction automatique qui sur la base d'une phrase, restituent une phrase traduite, les approches basées sur les dictionnaires proposent une traduction mot à mot sans se préoccuper de la syntaxe, ainsi, les termes ***mad cow*** seront traduits ***fou vache*** et non ***vache folle*** [Savoy, 01].

Les dictionnaires, utilisés dans ce domaine, sont généralement des listes de termes donnés dans la langue source alignés avec d'autres termes de la langue cible. La traduction basée sur ces dictionnaires fournit en sortie les traductions d'un terme donné en entrée, c'est pour cela que l'idée principale des techniques proposées dans [Davis, 96 ; Ballestros, 96 ; Hull, 96 ; Sanderson, 00 ; Baziz, Boughanem & Nassr, 04] a été de remplacer chaque terme de la requête par le(s) terme(s) approprié(s) dans la langue cible, ces techniques n'ont pas été totalement satisfaisantes à cause de la difficulté de la traduction automatique et des imperfections des dictionnaires bilingues, qui posent souvent des problèmes.

Dans [Ballestros & al 96, Ballestros & al 97], furent élaborés les premiers travaux basés sur les dictionnaires. Ces derniers ont montré que l'utilisation du dictionnaire Collins (Espagnol-Anglais) pour la traduction de requêtes peut mener à une baisse de 40-60% au niveau des performances des résultats par rapport aux résultats du monolingue (requête en Anglais contre documents en Anglais). Ils attribuent ceci à trois problèmes principaux: Le manque d'un vocabulaire spécialisé dans le dictionnaire, l'ambiguïté des termes lors de la traduction et la non traduction des concepts multi termes tels que l'expression. Ils ont également montré dans le cadre de ces travaux que des améliorations au niveau des résultats de la traduction de requêtes par le dictionnaire peuvent être obtenues en utilisant la pseudo réinjection de la pertinence (pseudo-relevance feedback) avant et après la traduction. Les performances, en termes de précision, ont augmenté de 16% à 34% quand la réinjection est appliquée avant la traduction et entre 14.3% et 47.5% quand elle est appliquée après la traduction. La combinaison des deux niveaux (avant et après la traduction) donne une amélioration

entre 40% et 51%. La collection de test et les requêtes utilisées pour ces différentes évaluations sont issues de la collection TREC.

Dans [Davis, 96 ; Davis, 98], l'auteur a exploité une version électronique du Collins pour réaliser la traduction des termes de la requête de l'Anglais vers l'Espagnol, la collection de test et les requêtes utilisées sont issues de la collection de TREC 6. Davis a montré que la traduction terme par terme de la requête utilisant le dictionnaire bilingue COLLINS (Anglais-Espagnol), produit un déficit de 58.2% par rapport aux performances de la recherche monolingue (requêtes Espagnol contre documents Espagnols).La recherche monolingue a été effectuée en considérant des requêtes en Espagnols construites manuellement et fournies dans le cadre de TREC 6. Ces requêtes sont évaluées sur des documents en Espagnols. Tous les travaux utilisant les dictionnaires pour la traduction de requêtes, ont démontré que pour améliorer les résultats de la recherche d'information par croisement de langues il est nécessaire de combiner le dictionnaire avec une méthode de désambiguïsation stricte qui permet de réduire l'ambiguïté des termes fournit par le dictionnaire.

5.4. Approches pour la désambigüisation des requêtes

Une part importante des travaux effectués actuellement explorent cette direction et tentent de chercher des stratégies de désambiguïsation efficaces.

Dans [Grefenstette, 98 ; Oard, 98], une variété de stratégies pour la désambiguïsation des termes de la requête a été proposée. Les travaux recensés sont principalement basés sur les corpus alignés parallèles et comparables. La plupart des approches de désambiguïsation basées sur les corpus alignés utilisent des cooccurrences entre termes calculées à partir de ce corpus pour choisir la(es) meilleur(es) substitution(s) possibles pour un terme donné.

Ainsi dans [Ballestros, 97], les valeurs de cooccurrences sont calculées entre les termes anglais et espagnols en se basant sur un corpus parallèle (espagnole-anglais).La désambiguïsation consiste à retenir pour chaque terme anglais le terme espagnol le plus co-occurrent parmi les substitutions possibles obtenues par le dictionnaire COLLINS (anglais-espagnol) pour ce terme anglais. Elle a montré que la précision moyenne est améliorée de 31% par rapport aux résultats obtenus par le dictionnaire.

Dans [Davis, 97], l'approche proposée par Davis et Odgen n'utilise pas de valeurs de cooccurrence entre termes, mais effectue plusieurs recherches monolingues sur

chacune des parties du corpus parallèle (anglais-espagnol) à l'aide d'un SRI basé sur le modèle vectoriel QUILT. Tout d'abord, une recherche monolingue est effectuée avec la requête sur une partie du corpus parallèle pour trouver la liste ordonnée de documents résultats. Ensuite, une recherche monolingue sur l'autre partie du corpus parallèle est effectuée pour chacune des traductions possibles d'un terme de la requête. Le produit scalaire entre les différents vecteurs est ensuite calculé, entre les vecteurs de documents de chaque traduction et le vecteur de document du terme source. La traduction choisie est celle qui obtient la liste de documents la plus proche de la liste de la requête. Dans cette approche, il s'agit encore de faire une traduction mot à mot des termes de la requête. Ils ont montré que la désambiguïsation améliore de 37% les résultats obtenus par la traduction simple par le dictionnaire. Ils ont remarqué également que la traduction choisie par le système ne favorise pas forcément les traductions les plus fréquentes dans le corpus.

Dans [Yamabana, 98], l'auteur a développé une méthode de désambiguïsation utilisant un corpus comparable. L'approche proposée consiste à calculer automatiquement à partir de ce corpus l'ensemble des valeurs de cooccurrence entre les termes de la langue source et les termes de la langue cible. Ce thesaurus est utilisé pour sélectionner la meilleure traduction en langue cible.

Dans [Nassr, 02], l'auteur a utilisé deux méthodes de désambiguïsation: La première concerne l'utilisation du corpus aligné, les meilleurs résultats sont obtenus par l'utilisation du contexte de la requête. L'utilisation des phrases alignées pour la désambiguïsation a montré également de bonnes performances. La deuxième utilise les dictionnaires bilingues comme un moyen de désambiguïsation. Cependant, il n'y a pas une grande différence entre les résultats obtenus par les deux mesures de similarité et la traduction bidirectionnelle.

Dans [Baziz, Boughanem & Nassr, 04], les travaux se sont dirigé vers une approche de désambiguïsation qui s'appuie sur des concepts issus d'une base de données lexicographique externe (WordNet 1.7) comme des liens sémantiques dénotant des relations telles que spécifique/générique ou partie/tout sont ensuite utilisés pour l'expansion de ces requêtes. Toutes ces expérimentations ont été effectuées sur la base CLEF et utilisent le moteur de recherche d'information "Mercure".

Deux groupes de tests ont été réalisés, le premier groupe concerne l'utilisation des dictionnaires bilingues sans la désambiguïsation, le deuxième groupe concerne

l'utilisation des dictionnaires bilingues pour la traduction combinée avec la désambiguïsation et l'expansion, basées sur les concepts de WordNet. Les résultats obtenus par cette combinaison sur toutes les précisions sont meilleurs que ceux obtenus par le dictionnaire.

5.5. Approches basées sur les Corpus alignés

Rappelons que l'utilisation des dictionnaires pour la traduction de requêtes pose des problèmes liés à l'absence des termes spécifiques à un domaine ou l'absence de certaines formes d'un terme, et dans la plupart des cas, ces dictionnaires proposent pour un terme donné différentes traductions. C'est la raison pour laquelle la communauté de la recherche d'information par croisement de langues s'est orientée vers les méthodes basées sur les corpus alignés. Ces derniers tentent d'y répondre par extraction automatique de l'information manquante.

Les méthodes proposées dans [Davis, 96 ; Ballestros, 98 ; Sheridan, 96 ; Schauble, 00 ; Boughanem, 00 ; Nassr, 02] utilisent directement le contenu d'un ensemble de documents, regroupés dans un corpus alignés soit pour la traduction ou pour la désambiguïsation des requêtes. L'approche basée sur les corpus alignés est capable de fournir des termes reliés, a titre d'exemple, face à la requête sur « ***IRA attacks in airports »***, elle redonne le terme « ***bomb »*** relié au thème de la requête, mais ne correspondant pas à une traduction mot à mot. [Savoy, 01]

Un corpus aligné est constitué d'un ensemble de documents exprimés dans une langue, alignés avec des documents dans une autre langue. L'alignement entre ces documents consiste à mettre en correspondance les documents de langues différentes selon un critère donné. Il peut être *parallèle* ou *comparable* ***:***

–***L'alignement parallèle*** consiste à mettre en correspondance chaque document d'une Langue source L1 avec le document représentant sa traduction dans la langue cible L2. Dans ce cas, l'alignement peut être fait sur: le document, les paragraphes, les phrases ou les termes. Les corpus basés sur ce type d'alignement sont appelés les *corpus parallèles***.**

Les premiers travaux ont été élaborés dans [Sheridan & al, 96 ; Littman & al, 98], les auteurs ont utilisé des corpus parallèles basés sur l'utilisation des méthodes statistiques pour l'extraction des termes équivalents multilingues à partir de ces corpus,

le but de ces approches est de construire un thesaurus de termes exprimés dans une langue reliés à d'autres termes dans d'autres langues. Ce thesaurus est utilisé par la suite pour la traduction des requêtes.

Dans [Nassr, 02], l'auteur a effectué une série de tests pour évaluer l'impact des corpus alignés indépendants des collections de tests sur la traduction de requêtes. Dans ce cas, les tests ont été effectués sur deux types d'alignement : *l'alignement par document* et celui par *phrases*. Elle a observé que les résultats obtenus par les phrases alignées sont meilleurs que ceux obtenus par les documents alignés. Elle explique cette différence de résultats par le fait que l'alignement par phrases produit de bonnes correspondances entre les termes. Contrairement à l'alignement par documents qui est général et où la correspondance des termes est moins évidente. Pour compléter ses tests sur les phrases alignés, ces derniers ont également été comparés au dictionnaire. Une fois encore, les résultats obtenus par les phrases alignées sont meilleurs que ceux du dictionnaire. Il faut noter que ces résultats sont obtenus en utilisant le corpus WAC qui est totalement indépendant des documents de la collection de test.

–***L'alignement comparable*** est plus délicat à réaliser, en effet cela revient à mettre en correspondance des documents en se basant sur des critères comme par exemple la présence de même dates, de même noms de personnes dans des documents de langues différentes. Les corpus basés sur ce type d'alignement sont appelés les corpus comparables.

Une des stratégies intéressantes dans cette catégorie d'approches est celle proposée dans [Sheridan & al, 96], les auteurs ont construit un thesaurus de similarité entre termes de différentes langues à partir d'un corpus comparable. Cet alignement est réalisé par sujet et date et à partir des articles en Allemand et en Italien de journaux issus de l'agence suisse (Schweizerische Depeschen Agentur (SDA)).Les requêtes exprimées en Allemand sont posées sur les collections de documents en Italien. Les auteurs ont constaté que les résultats obtenus sont meilleurs que la recherche monolingue.

Dans [Nassr, 02], Les travaux de l'auteur ont également été portés sur l'utilisation d'un corpus aligné comme moyen de traduction, l'auteur a réalisé un premier test sur les associations entre termes. Cette technique a déjà été utilisée par d'autres travaux sur des collections non homogènes, mais son but est de mesurer l'impact des collections de documents orientées domaine comme Amaryllis dans l'établissement de relations. Les

résultats obtenus par cette technique dépassent même les résultats monolingues. Ces résultats sont attendus pour la simple raison que le corpus aligné utilisé pour les associations entre termes traite le même domaine que les documents de la collection de test.

5.6. Approches basées sur le vocabulaire prédéfini

Cette approche consiste d'une façon générale à utiliser un vocabulaire contrôlé, représenté sous forme d'un thésaurus multilingue. Les correspondances entre termes de différentes langues sont prédéfinies par le vocabulaire et regroupées dans des classes.

Une classe représente une entrée du vocabulaire. Ces approches sont utilisées pour la représentation des documents et des requêtes. L'indexation des documents est réalisée de façon guidée par le vocabulaire. Ainsi, chaque document est représenté par une liste de classes de termes. La recherche d'information revient donc à représenter la requête dans ce référentiel (liste de classes) et à récupérer les documents exprimés dans les différentes langues et indexés par cette liste.

Les premiers travaux dans cette direction de recherche ont été faits par G.Salton en 1970.Dans son expérimentation, l'auteur a utilisé une liste de concepts multilingues anglais – allemand construite manuellement à partir d'une traduction de l'anglais vers l'allemand. L'expérimentation a été réalisée sur un corpus contenant 468 résumés en allemand et 1095 résumés en anglais. La précision moyenne a été d'environ 95% par rapport à celle d'un système monolingue, cette expérimentation représente la première expérience d'utilisation d'un thésaurus multilingue prédéfini.

Dans [Salton, 71 ; Pevzner, 72], d'autres résultats expérimentaux démontrant l'efficacité d'une telle approche ont été rapportés, ils ont montré qu'avec un thesaurus (Français-Anglais) soigneusement construit, les résultats de la recherche étaient presque aussi efficaces que la recherche monolingue. Plus particulièrement l'approche de Salton consistait à construire manuellement des classes de termes à partir d'une petite collection de documents en français et de leurs traductions en anglais. Les groupes de termes liés de chaque langue sont placés dans une classe individuelle de telle manière que les groupes correspondants dans deux langues différentes aient le même identifiant de classe. Bien que les résultats étaient aussi bon que ceux obtenus par le monolingue, les résultats obtenus par la recherche pour des requêtes en français et des documents en

Rapport-Gratuit.com

anglais étaient moins efficaces. Le problème principal réside dans le fait qu'un terme de la requête exprimé en français avait plusieurs traductions en anglais est donc plusieurs classes.

Dans [Diekema & al, 98], les auteurs étant d'ailleurs les concepteurs du système CINDOR, utilisent la hiérarchie du WordNet comme modèle pour construire un thésaurus multilingues. Les termes synonymes dans différentes langues sont regroupés en synsets. Les auteurs ont montré que les performances du croisement Français-Anglais et Anglais-Français arrivent à 75% des performances de la recherche monolingue. Dans cette expérimentation ils ont remarqué également, certains problèmes liés au vocabulaire, comme le problème d'ambiguïté et de couverture qui influent sur les résultats de la recherche.

Dans [Boucham, 09], l'auteur a proposé une approche pour une indexation contrôlée de la requête et d'un corpus (corpus trilingue : arabe, français et anglais), cette approche est basée sur la construction d'une ontologie de domaine et ceux afin d'explorer l'apport des approches Web Sémantique (en particulier l'utilisation des ontologies pour améliorer la description sémantique des documents et des requêtes).

Le système proposé est fondé sur un formalisme de représentation de connaissances, plus précisément les graphes sémantiques qui supportent une ontologie de domaine. Les documents et les requêtes sont aussi représentés dans ce formalisme. L'ontologie du domaine constitue le noyau du système, elle est utilisée aussi bien pour l'indexation que pour la recherche. L'approche utilisée pour l'indexation utilise une méthode d'extraction qui est basée sur le calcul de segments répétés en utilisant des filtres linguistiques. L'approche utilisée pour la recherche consiste en une comparaison de graphes pour trouver les documents qui répondent à la requête étendue de l'utilisateur.

Dans [Harrathi, 09], l'auteur a proposé une méthode fondée sur la distance intertextuelle inter-domaine, des mesures statistiques et une ressource sémantique externe (l'ontologie multilingue du domaine) permettant l'extraction des concepts et relations entre concepts à partir des documents multilingues écrits en anglais et en langue latine qui serviront de descripteurs sémantiques pour la représentation d'un document. L'evaluation a été faite en utilisant la collection CLEF médicale 2007 et en comparant les resulats obtenu en appliquant sa méthode aux resultats obtenu par

l'utilisation des analyseurs linguistiques :MetaMap ,MiniPar ,TreeTagger et au final les résultat sont comparables.

6. Les problèmes de la recherche d'information multilingue

La Recherche d'information multilingue et la recherche d'information monolingue sont confrontées aux mêmes problèmes qui résident dans les pièges et difficultés du langage naturel (*Polysémie, Sens large, Homographie).*

Un autre problème de l'indexation multilingue lors de la traduction : la car les concepts n'existent pas dans toutes les langues par exemple: *traduction concept – concept* .

–Concept de L1 sans équivalent dans L2.

–Concept de L1 = agrégat de concepts de L2.

–Concept de L1 ≈ concept de L2. [Nassr, 02] [Boucham, 09]

7. Conclusion

Ce deuxième chapitre a porté essentiellement sur l'étude des systèmes de recherche d'information par croisement de langues. L'objectif principal d'un tel système est de permettre à un utilisateur donné de formuler sa requête dans une langue et de sélectionner des documents pertinents répondant à sa requête dans une langue différente.

Dans ce chapitre, nous avons introduit en premier lieu un cadre général concernant les systèmes de recherche d'information par croisement de langue, puis nous avons élaboré un récapitulatif concernant les travaux accomplit dans ce domaine c'est-à-dire un état d'art de la recherche d'information multilingue, au final, nous avons cité les différents problèmes rencontrés dans ce domaine.

Rapport-gratuit.com
LE NUMERO 1 MONDIAL DU MÉMOIRES

Chapitre III

Expansion de requête pour un Système de Recherche d'Information par croisement de langues

1. Introduction

Dans le chapitre précédent, nous avons étudié les différentes approches proposées pour mettre les requêtes et les documents dans un même référentiel en croisement de langues. Nous rappelons que le but d'un système de recherche d'information par croisement de langues est de retrouver des documents pertinents exprimés dans une langue différente de celle de la requête. Actuellement la plupart des travaux dans ce domaine se focalisent sur la traduction de la requête. Cette traduction est moins coûteuse que celle de tous les documents de la collection. Pour notre part, étant donné que les requêtes sont plus courtes que les documents, nous pensons aussi qu'il est plus réaliste de traduire la requête seulement.

La requête est souvent une suite de termes indépendants, situation que l'on rencontre couramment dans les moteurs de recherche. Cependant, la traduction de ces requêtes n'est pas sans engendrer des problèmes. Le problème d'expressivité de la requête traduite est posé quand les termes issus de la traduction ne sont pas suffisants pour représenter la requête initiale. D'où la nécessité d'expansion pour enrichir la requête avec des termes plus courants. Mais le problème le plus crucial à résoudre est comment désambigüiser et étendre les requêtes et à quel moment ?

Etant donnée l'utilisation d'un traducteur automatique pour la traduction de notre requête (traduction faite mot par mot), le traducteur automatique va sélectionner une seule traduction pour chaque mot de la requête et non pas plusieurs.

Dans le cas ou le mot possède plusieurs traduction et que celle retourné par le traducteur ne correspond pas au sens dénoté, cela engendre une perte de sens.

Sachant qu'au départ nous avons déjà une perte d'information importante entre la réalité du besoin d'information de l'utilisateur et son expression par ce dernier sous forme de requête, cela pourrait empirer après la traduction de cette dernière c'est pour cela que nous choisissions de faire l'expansion de la requête avant la traduction en offrant à l'utilisateur la possibilité de spécifier le sens dénoté par chaque terme de la requête.

Nous nous intéresserons dans ce chapitre à l'étude de techniques d'expansion des termes de la requête, ces dernières s'appuient sur des concepts issus d'une base de données lexicographique externe (WordNet 2.0) et sur l'intervention de l'utilisateur.

2. Problématique

Dans le domaine de la recherche d'information, la source principale de l'ambiguïté réside dans les variations linguistiques présentes dans le texte des requêtes et des documents. Ces variations linguistiques peuvent être interprétées par le fait que le langage n'est pas simplement une collection de mots, mais un moyen de communiquer au sujet de concepts. Ce qui rend l'hypothèse de récupération de mots clé insuffisante. Ceci est dû au fait que les termes utilisés par l'utilisateur dans sa requête, peuvent présenter par rapport à ceux des documents de la base, des variations morphologiques (comme dans « ***wolf*** » et « ***wolves*** »), des variations lexicales ou des mots différents sont utilisés pour représenter le même sens (« ***film*** » et « ***movie*** ») ou encore des variations sémantiques, où des mots ont plusieurs sens: Un pétrolier cherchant par exemple le mot «***oil*** » sera confronté à «***olive oil »*** et « ***kitchen***».

D'où l'idée d'une désambiguïsation des requêtes initiales avec des concepts issus d'une base lexicographique externe et à l'aide de l'utilisateur, ce dernier va sélectionner le concept qui lui semble approprié. Ces requêtes seront ensuite étendues avec d'autres concepts reliés à ceux sélectionnés par l'utilisateur par des liens sémantiques tels que, spécifique/générique ou partie/tout. Ceci permet, à titre d'exemple pour le mot « ***country*** », de récupérer en premier en utilisant la synonymie, les mots « ***administrative*** », « ***district*** », « ***administrative division*** », « ***territorial division*** », l'hyperonymie, de récupérer les mots « ***political unit*** »et l'hyponymie, de récupérer les mots « ***banana republic*** », « ***fatherland*** ». Ou encore de passer du sigle « ***EU*** » à « ***European Union, EU, European Community, EC, European Economic Community, EEC, Common Market, Europe*** ». Une fois étendues, les requêtes seront ensuite traduites et soumises au système de recherche d'information.

3. Description de l'approche suivie

Comme pour tout système de RI, l'étape d'indexation est nécessaire, les documents exprimés dans une langue L1 (français) y sont analysés. Les requêtes quant à elles sont exprimées dans la langue L2 (anglais), ces dernières seront d'abord désambiguïsées et étendues à l'aide de la base lexicographique WordNet 2.0 pour être traduites ensuite vers la langue L1 (français). La requête traduite va passer ensuite par les mêmes étapes d'indexation que le corpus.

Pour illustrer notre approche, une vue générale et schématique de cette dernière est présentée dans la figure III.1 ci-dessous :

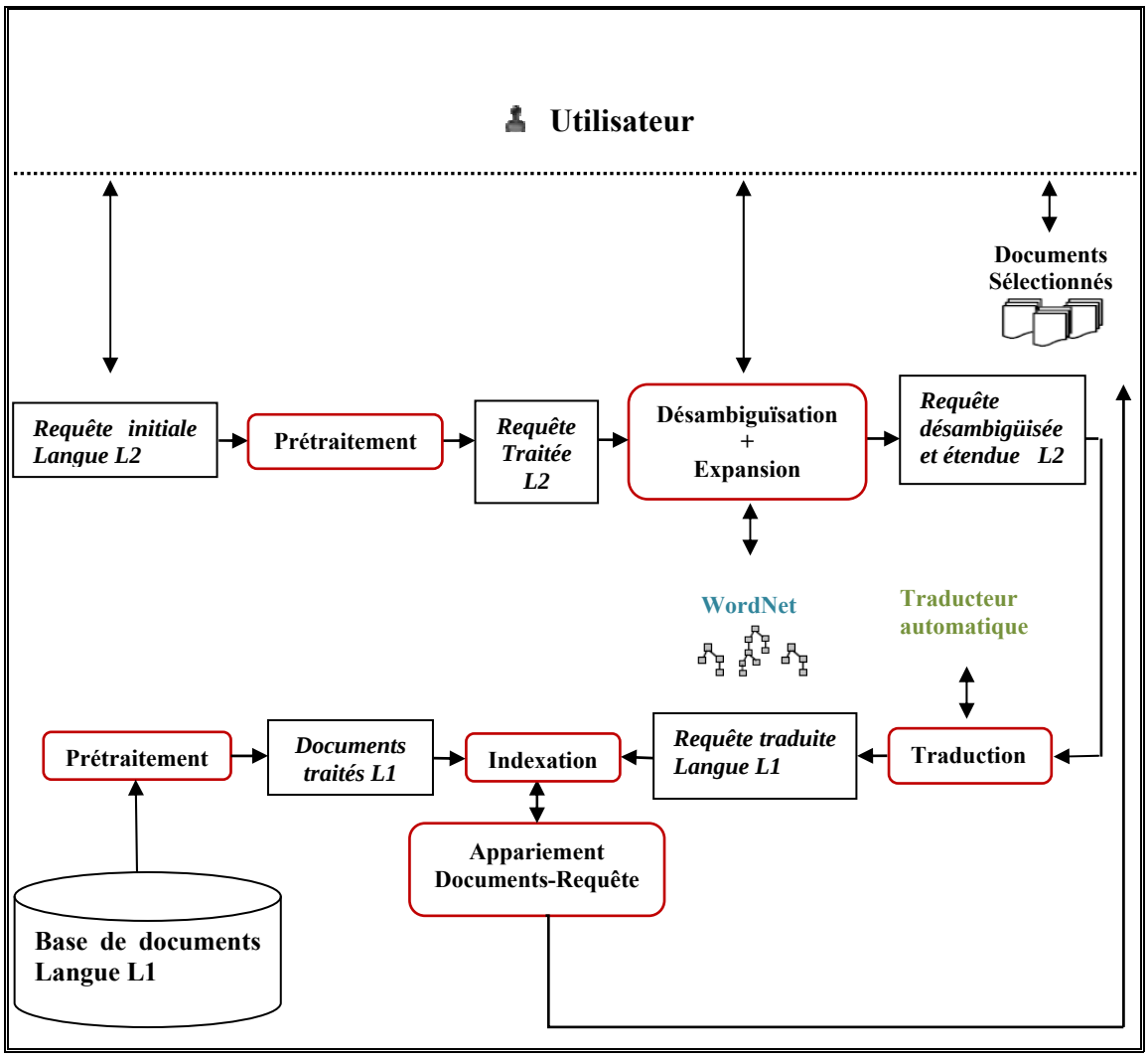

Figure III.1 Schéma synoptique de la stratégie suivie

L'approche illustrée dans le schéma ci-dessous est composée de plusieurs étapes, dans ce qui suit nous allons expliquer chacune d'entre elles :

3.1. Prétraitement

Etant donné que les requêtes et les documents de la base documentaire sont exprimés à l'aide du langage naturel, il est nécessaire d'appliquer un prétraitement sur ces deux derniers, ce prétraitement consiste en 3 étapes essentielles à savoir :

- Conversion des termes en minuscules.
- Elimination des caractères de ponctuation.

- Elimination des mots vides.

Dans le cas ou le prétraitement est appliqué à la requête sachant que cette dernière est exprimée dans la langue L2(anglais) l'élimination des mots vides se fera en fonction de la liste des mots vides de langue anglaise, de même pour le prétraitement sur la base documentaire exprimée dans la langue L1 (français), l'élimination des mots vides se fera en fonction de la liste des mots vides de la langue française [12].

3.2. Désambigüisation et expansion de la requête

La désambiguïsation de la requête initiale a pour objectif d'améliorer la pertinence des documents sélectionnés par le SRI. Elle consiste à se focaliser sur le sens dominant de la requête et à se détacher de ses sens secondaires. Elle se situe entre le moment où l'utilisateur introduit la requête et le moment ou la traduction de cette dernière se fait.

Il est nécessaire de procéder à l'étape de désambigüisation avant l'étape d'expansion et ceux afin de préciser le sens à partir du quel nous allons extraire les mots pour étendre la requête.

3.2.1. La désambiguïsation

Une fois la requête traitée, nous allons appliquer une désambigüisation sur l'ensemble des mots de cette dernière.

La désambigüisation peut être effectuée de façon automatique ou manuelle, nous avons choisit d'effectuer une désambigüisation manuelle étant donné la fiabilité de cette dernière, celle-ci consiste à faire intervenir l'utilisateur en lui offrant la possibilité de spécifier le sens dénoté par les mots de la requête.

Pour chaque mot de la requête faire :

- Extraire les sens possible du mot sélectionné (présentation des glossaires contenus dans les synsets).
- Choisir le sens (synset) le plus approprié (adéquat).

[12] La Stoplist du français proposée par Jacques Savoy est disponible sur le site : (http://www.unine.ch/info/CLEF/frenchST.txt)

Remarque : Les sens (synsets) présentés sont tirés de la base lexicographique WordNet et regroupés par noms, verbes, adjectifs et adverbes.

3.2.2. Expansion de la requête

En se basant sur le fait qu'un sens peut être véhiculé par des mots différents, l'expansion de la requête considère des mots (ou termes) reliés pour étendre la requête.

Cette dernière est aussi vue comme un traitement pour "Cibler" le champ de recherche de cette requête. Dans notre cas, nous utilisons le modèle vectoriel pour la représentation des documents et de la requête, cela implique que les termes reliés à la requête seront rajouté à son vecteur.

Nous avons élaboré plusieurs stratégies destinées à faire l'expansion de la requête, dans ce qui suit nous allons détailler ces dernières :

- ✓ ***Stratégie n°1 :L'hyperonymie***

L'hyperonymie est le terme générique utilisé pour désigner une classe englobant des instances de classes plus spécifiques. Y est un hyperonyme de X si X est un type de (kind of) Y.

Dans ce cas la, l'expansion de la requête initiale se fait à l'aide de l'ensemble des mots contenue dans le synset (sens choisit dans l'étape de désambigüisation) mais aussi des mots contenues dans les sens pères de ce synset.

Sachant que les sens pères (hyperonymes) sont organisés sous forme d'hiérarchie, il faut spécifier le niveau dans cette hiérarchie qui sera pris en compte pour l'expansion de la requête.

- Extraire les mots reliés au synset(sens) choisit.
- Pour chaque niveau dans la hiérarchie faire :
  - -Extraire les sens pères de ce synset.
  - -Extraire les mots contenus dans les sens pères.

Rapport-gratuit.com
LE NUMERO 1 MONDIAL DU MÉMOIRES

✓ ***Stratégie n°2 :L'hyponymie***

L'Hyponymie est le terme spécifique utilisé pour désigner un membre d'une classe (Relation inverse de Hyperonymie). X est un hyponyme de Y si X est un type de (kind of) Y.

Dans ce cas la, l'expansion de la requête initiale se fait à l'aide de l'ensemble des mots contenue dans le synset (sens choisit dans l'étape de désambigüisation) mais aussi des mots contenus dans les sens fils de ce synset.

Sachant que les sens fils (hyponymes) sont organisés sous forme d'hiérarchie, il faut spécifier le niveau dans cette hiérarchie qui sera pris en compte pour l'expansion de la requête.

- Extraire les mots reliés au synset(sens) choisit.
- Pour chaque niveau dans la hiérarchie faire :
  - -Extraire les sens fils de ce synset.
  - -Extraire les mots contenus dans les sens fils.

✓ ***Stratégie n°3 : Synonymie***

La synonymie est le terme spécifique utilisé pour désigner deux mots qui sont interchangeables dans certains contextes linguistiques.

Dans ce cas la, l'expansion de la requête initiale se fait à l'aide de l'ensemble des mots contenue dans le synset (sens choisit dans l'étape de désambigüisation) mais aussi des mots contenues dans les sens synonymes de ce synset.

- Extraire les mots reliés au synset(sens) choisit.
- Extraire les mots contenus dans les sens synonymes de ce synset.

Remarque : Après avoir longuement observé les relations sémantiques existantes dans la base lexicographique WordNet 2.0, nous nous sommes intéressées aux relations sémantiques (synonymie, hyperonymie, hyponymie) et nous avons remarqué (tableau III.1) que ces relations sémantiques varient en fonction du type (part of speech « POS ») du mot (adjectif, adverbes, verbes, noms) :

–Les adjectifs et les adverbes ne possèdent pas d'hyperonymes ni d'hyponymes mais ils possèdent des synonymes.

–Les verbes et les noms possèdent des synonymes, des hyperonymes et des hyponymes sauf que les synonymes des verbes et des noms présentent des analogies avec leurs hyperonymes, plus précisément les synonymes représentent le niveau 1 dans la hiérarchie des pères.

Relation / POS	**Synonymie**	**Hyperonymie**	**Hyponymie**
Adverbe	✓	×	×
Adjectif	✓	×	×
Nom	Niveau1 (hyperonyme)	✓	✓
Verbe	Niveau1 (hyperonyme)	✓	✓

Tableau III.1 Récapitulatif des observations menées

3.3. Traduction de la requête désambiguïsée et étendue

Une fois la requête désambigüisée et étendue, les mots de cette dernière vont subir une traduction. Cette opération offre une seule traduction possible pour chaque terme de la requête source.

Ce processus considère une traduction mot par mot réalisé à l'aide d'un traducteur automatique GoogleTranslate.Le résultat est une requête traduite dans la même langue que la base documentaire à savoir la langue L1 (français).

3.4. Indexation

L'objectif de l'analyse et de l'indexation est de trouver en premier les concepts les plus importants dans les documents et dans les requêtes et ensuite créer une représentation interne en utilisant ces concepts.

Étant donné que les requêtes, une fois traduites sont exprimées dans la même langue que la base documentaire, nous pouvons donc utiliser le même processus d'indexation pour la base documentaire ainsi que pour les requêtes et ceux afin de

Rapport-Gratuit.com

pouvoir les représenter dans le même référentiel. Nous avons choisi le modèle vectoriel comme modèle pour la représentation interne des documents et des requêtes.

L'indexation des documents et des requêtes passent par les étapes suivantes :

–La racinisation [13].

–Concernant la pondération des termes des documents nous avons choisit d'appliquer la formule de tf*idf, quand aux termes de la requêtes, la pondération est soit de 1 soit de 0 (si le terme appartient à l'ensemble des termes issus de l'indexation des documents sa pondération est de 1 ; 0 sinon).

Dans le modèle vectoriel, chaque document est représenté par un vecteur de poids comme suit:

$$\begin{array}{lcl} & & t_1 \quad t_2 \quad t_3 \quad \ldots t_n \\ d_i & \rightarrow & < p_{i1} \quad p_{i2} \quad p_{i3} \quad \ldots \quad p_{in} > \end{array}$$

Ou :

p_{ij} : le poids du terme t_j dans le document **di**.

$\{t_1 \quad t_2 \quad t_3 \quad \ldots t_n\}$: ensemble des termes issus de l'indexation.

La requête est aussi représentée sous forme de vecteur, comme ceci:

$$\begin{array}{lcl} & & t_1 \quad t_2 \quad t_3 \quad \ldots t_n \\ q_k & \rightarrow & < p \quad p \quad p \quad \ldots p > \end{array}$$

Ou: $\mathbf{p} \in \{\mathbf{0}, \mathbf{1}\}$

3.5. Appariement documents-requête

Une foi la requête analysée et représentée dans le même espace vectoriel que celui des documents, nous avons utilisé la ***mesure du cosinus*** (voir chapitre I (5.2.1)) comme méthode de comparaison entre les deux représentations et ceux, afin de déterminer leurs degrés de correspondance et de classer les documents pertinents par ordre décroissant.

[13] L'algorithme de racinisation pour le français est disponible le site : (http://snowball.tartarus.org/algorithms/french/stemmer.html)

Les documents classés constituent le résultat de la recherche, ce dernier sera retourné à l'utilisateur sous forme d'une liste.

4. Expérimentation et évaluation

4.1. Environnement d'expérimentation

4.1.1. Présentation de WordNet

WordNet 2.0[14] est une base de données lexicale développée par des linguistes du laboratoire des sciences cognitives de l'université de Princeton. Son but est de répertorier, classifier et mettre en relation de diverses manières le contenu sémantique et lexical de la langue anglaise.

La composante atomique sur laquelle repose le système entier est le synset (synonym set), un groupe de mots interchangeables, dénotant un sens ou un usage particulier, ses caractéristiques sont présentées dans le (tableau III.2).

Part of speech	Words	Concepts	Total Word-Sense Pairs
Noun	114648	79689	141690
Verb	11306	13508	24632
Adjective	21436	18563	31015
Adverb	4669	3664	5808
Totals	**152059**	**115424**	**203145**

Tableau III.2 Nombre de mots et de concepts de la base lexicographique WordNet 2.0

4.1.2. Présentation de NetBeans

NetBeans[15] est un environnement de développement intégré (EDI), placé en open soure par Sun en juin 2000 sous licence CDDL et GPLv2 (Common Development and Distribution License). En plus de Java, NetBeans permet également de supporter différents autres langages, comme Python, C, C++, JavaScript, XML, Ruby, PHP et

[14] Lien : (http://wordnet.princeton.org)

[15] Site : (www.netbeans.org)

Rapport-gratuit.com
LE NUMERO 1 MONDIAL DU MÉMOIRES

HTML. Il comprend toutes les caractéristiques d'un IDE moderne (éditeur en couleur, projets multi-langage, refactoring, éditeur graphique d'interfaces et de pages Web).Conçu en Java, NetBeans est disponible sous Windows, Linux, Solaris (sur x86 et SPARC), Mac OS X ou sous une version indépendante des systèmes d'exploitation (requérant une machine virtuelle Java).

Un environnement Java Development Kit JDK est requis pour les développements en Java. NetBeans constitue par ailleurs une plate forme qui permet le développement d'applications spécifiques (bibliothèque Swing (Java)). L'IDE NetBeans s'appuie sur cette plate forme, il s'enrichit à l'aide de plugins.

4.1.3. La base de test

L'expérimentation a été effectuée sur un corpus de documents issus du projet TREC 2000[16].

Cette collection contient plus de 487 000 articles en treize langues (néerlandais, français, allemands, chinois, japonais, russe, portugais, espagnol, espagnol latino-américain, italien, danois, norvégien et suédois.). Les articles ne sont pas parallèles, mais ils sont écrits par des journalistes locaux.

Nous avons choisit d'extraire du corpus multilingue, 2486 documents en français (tableau III.3). Ces documents sont classés par thème (désastres et accidents, sport, religion, privatisation des entreprises, violence et guerre civil, questions de travail, mode). Nous avons construit un ensemble de requête par thème et ceux en se basant sur les titres des documents appartenant à chaque thème. Afin que les requêtes traitent des sujets variés nous avons choisit des thèmes différents. Les requêtes en français ont été traduites en anglais à l'aide d'un expert. Le nombre total de ces dernières est de 20.

Quand à la liste des documents pertinents pour chaque requête, elles ont été construites sur la base de notre jugement et à partir des documents traitant le thème de la requête.

[16] Site : (http://trec.nist.gov/data/reuters/reuters.html)

Base de test		
Nombre de documents dans la base	Nombre de termes Dans la base	Taille moyenne d'un document (termes)
2486	**561836**	**226**

Tableau III.3 Description de la base de test utilisée

4.2. Evaluation

Le but de nos expérimentations est de montrer l'amélioration relative à la technique d'expansion, basée sur les concepts de WordNet 2.0. Pour ce faire, nous évaluons les stratégies d'expansion expliquées précédemment (voir 3.2) et nous comparons ensuite les résultats obtenus afin de déterminer la meilleure d'entre elles. Au final, pour mesurer l'efficacité de cette stratégie d'expansion, les résultats obtenus sont comparés aux résultats du test basé sur la traduction sans désambigüisation et expansion.

Pour que l'évaluation soit fiable, il ne faut pas que l'étape de désambigüisation varient d'une stratégie à l'autre en d'autre termes, le choix des sens pour les mots de la requête est fixe à chaque évaluation.

Le test de la traduction sans désambigüisation consiste à traduire les requêtes de l'Anglais vers le Français puis, les comparer aux documents français issus du corpus TREC.

L'évaluation des performances est effectuée sur l'ensemble des documents sélectionnés pour les 20 requêtes. Elle se base sur les mesures de rappels et de précision.

Une précision moyenne et un rappel moyen sont calculés sur l'ensemble des documents sélectionnés.

4.2.1. Evaluation des stratégies

Les stratégies d'expansion sont basées sur les relations sémantiques de la base lexicographique WordNet 2.0 (hyperonymie, hyponymie, synonymie), comme nous l'avons expliqué précédemment.

Pour évaluer l'impact de l'expansion de la requête à l'aide des hyperonymes et des hyponymes, sachant que ces deux dernières sont organisées sous forme d'hiérarchie, le

niveau constitue un critère de comparaison. Nous avons choisit donc de combiner les deux stratégies pour l'expansion (tableau III.4) en fixant le niveau de l'une et en faisant varier celui de la seconde (élevé, moyen et bas) et vice versa.

Hyponymes / Hyperonymes	Niveau 0	Niveau 1	Niveau 4	Niveau 9
Niveau 0	×	P = 0.5955	P = 0.5590	P = 0.5567
		R = 0.9392	R = 0.9441	R = 0.9330
Niveau 1	P = 0.6071	P = 0.4343	P = 0.3971	P = 0.3958
	R = 0.9501	R = 0.9576	R = 0.9624	R = 0.9573
Niveau 4	P = 0.2669	P = 0.2341	P = 0.2286	P = 0.2088
	R = 0.9615	R = 0.9668	R = 0.9684	R = 0.9714
Niveau 9	P = 0.2567	P = 0.2240	P = 0.2013	P = 0.2016
	R = 0.9624	R = 0.9677	R = 0.9722	R = 0.9722

Tableau III.4 Comparaison entre l'expansion faite à l'aide des hyperonymes et l'expansion faite à l'aide des hyponymes en termes de précision et de rappel

Remarques :

–Concernant la combinaison de l'expansion faite à l'aide des deux relations hyperonymie et hyponymie, nous remarquons que plus le niveau est élevé plus la précision diminue. En effet au (niveau 1, niveau 1) on obtient une précision de 0.4343, au (niveau 4, niveau 4) on obtient une précision de 0.2286 et au (niveau 9, niveau 9) on obtient une précision de 0.2016. Nous expliquons cela par le fait que plus le niveau sélectionné dans la hiérarchie est élevé, plus on élargie le champ de recherche de la requête (c'est-à-dire que le nombre de documents retournés est élevé) plus la précision diminue.

–En fixant le niveau des hyperonymes à un niveau bas et en faisant varier le niveau des hyponymes à un niveau moyen puis élevé, nous remarquons que la précision est meilleure que dans le cas inverse (c'est-à-dire en fixant le niveau des hyponymes à un niveau bas et en faisant varier le niveau des hyperonymes à un niveau moyen puis élevé).

–Au final et après avoir comparés les résultats, nous concluons que les meilleurs résultats sont fournis par la combinaison de la relation d'hyperonymie (niveau 1) et de

la relation d'hyponymie (niveau 0), en d'autres termes c'est le cas ou l'expansion se fait uniquement à l'aide des hyperonymes (niveau 1), en effet la précision est de 0.6071. Nous expliquons cela par le fait que les termes issus de l'expansion à l'aide des hyperonymes du niveau 1 sont ceux qui se rapprochent le plus des termes de la requête. Ces résultats sont ensuite comparés à ceux fournis par les synonymes (tableau III.5).

	Hyperonymes (niveau 1)	**Synonymes**
Précision moyenne	**0.6071**	**0.9100**
Rappel moyen	**0.9501**	**0.9408**

Tableau III.5 Comparaison entre l'expansion faite à l'aide des hyperonymes (niveau1) et l'expansion faite à l'aide des synonymes

Le tableau ci-dessus montre clairement que la relation de synonymie présente les meilleurs résultats. C'est donc les résultats de la stratégie d'expansion basé sur la relation de synonymie que nous allons comparer aux résultats du test de traduction sans désambigüisation et expansion.

4.2.2. Evaluation finale

	1) Traduction sans expansion	**2) Synonymes**	**Amélioration (2-1)**
Précision moyenne	**0.8773**	**0.9100**	**3,72%**
Rappel moyen	**0.7628**	**0.9408**	**23,33%**

Tableau III.6 Impact de la désambiguïsation et de l'expansion basée sur les synonymes

D'après le tableau III.6, on remarque clairement que les résultats obtenus par la combinaison de la stratégie d'expansion basée sur la relation de synonymie à la traduction sont meilleurs que les résultats obtenue par la traduction des requêtes uniquement. En effet l'amélioration est de 3,72% pour la précision et de 23,33% pour le rappel.

Cette amélioration s'explique par le fait que dans le second cas c'est à dire celui ou les termes de la requête passent par une étape de désambigüisation et d'expansion basée sur la relation de synonymie, les mots représentants les synonymes et utilisés pour l'expansion ont tendance à être plus précis que les mots de la requête initiale ce qui amène le traducteur automatique à fournir les traductions exactes pour ces mots et donc d'améliorer la qualité des résultats obtenus.

Concernant le rappel et après avoir passé en revu toutes nos évaluations, nous remarquons une importante amélioration de ce dernier dans le cas de l'expansion

comparé au test de traduction sans désambigüisation et expansion. En effet, il est au environ de 94% contre 76%. Cela s'explique par le fait que l'expansion permet de restituer plus de documents pertinents.

5. Conclusion

Dans ce chapitre, nous avons présenté une technique d'expansion de requête en recherche d'information par croisement de langues basée sur l'utilisation des concepts issus d'une base de données lexicographique externe (WordNet 2.0). Le but de cette expansion est d'améliorer la qualité des requêtes traduites par le traducteur automatique.

La faisabilité de cette démarche a été testée en utilisant un moteur de recherche que nous avons élaboré en se basant sur le modèle vectoriel. La démarche proposée permet d'une part, de se focaliser sur le sens dominant de ces requêtes, et d'autre part, de les enrichir avec des termes reliés sémantiquement à ceux des requêtes.

En effet nous avons choisit d'effectuer une désambigüisation manuelle, en permettant à l'utilisateur de choisir le sens dénoté par les mots de la requête et ceux afin tirer profit du contexte initiale de la requête. Puis nous nous sommes intéressés à l'étude de stratégies d'expansion de la requête, ces stratégies sont basées sur les relations sémantiques existantes dans la base lexicographique WordNet 2.0 (hyperonymie, hyponymie, synonymie). Pour déterminer la meilleure stratégie d'expansion, nous avons effectué une série de tests (évaluations). L'évaluation concerne en premier lieu l'expansion à l'aide des deux relations hyperonymie et hyponymie, étant donné que ces dernières se présentent sous forme d'hiérarchie (niveaux). A l'issus de cette première évaluation nous concluons que les meilleurs résultats sont fournis par la combinaison de la relation d'hyperonymie (niveau 1) et de la relation d'hyponymie (niveau 0), c'est-à-dire le cas ou l'expansion se fait uniquement à l'aide des hyperonymes (niveau 1), en effet la précision est de 0.6071. Nous avons poursuivi notre évaluation en comparants ces résultats à ceux fournis par les synonymes, et ce sont ces derniers qui offrent les meilleurs.

Au final et afin de déterminer l'impact de la stratégie d'expansion basée sur la relation de synonymie, nous avons comparé ces résultats à ceux du test de traduction sans désambigüisation et sans expansion. Nous avons clairement remarqué que le premier cas offre les meilleurs résultats. En effet l'amélioration est de 3,72% pour la précision et de 23,33% pour le rappel.

Ce qu'il faut tirer des différentes expérimentations que nous avons réalisées est qu'il faut toujours placer la requête dans son contexte. En effet, il faut arriver à extraire les concepts dominants véhiculés par la requête. De plus ce sont ces concepts ou ce contexte qu'il faut traduire ou lui trouver un équivalent dans la langue cible. Nous avons montré à l'issue de ces tests, que l'utilisation des concepts d'une ressource sémantique externe, est une solution fiable pour la désambigüisation et l'expansion des requêtes.

Cette solution nous a permis d'améliorer non seulement la qualité des requêtes traduites mais aussi la qualité des résultats recensés.

Conclusion Générale

Les travaux développés dans ce mémoire s'inscrivent dans le cadre de la conception des systèmes de recherche d'information capables de rechercher des documents indépendamment de la langue de la requête. Rappelons que le but des ces systèmes de recherche d'information par croisement de langues est de récupérer des documents pertinents répondants a un besoin d'utilisateur exprimé dans une langue différente de celles de la requête.

Notre manuscrit s'articule en trois chapitres, le premier chapitre introduit un cadre général ou nous avons présenté les points cruciaux du domaine de la recherche d'information, dans le second chapitre nous avons décrit les caractéristiques principales des systèmes de recherche d'information par croisement de langues, les différentes ressources multilingues étant donné que la qualité de ces systèmes en dépend puis nous avons passé en revue les différents travaux du croisement de langues, plus particulièrement, nous nous sommes intéressés a la description des techniques basées sur la traduction des requêtes qui représentent le cadre de notre travail, nous avons ensuite mis l'accent sur les différents problèmes induits par cette traduction, a savoir, la perte de sens qui engendre une dégradation des rendements du système de recherche d'information. Dans le chapitre 3 nous avons proposé une approche permettant d'améliorer la qualité des requêtes traduites. Pour cela nous avons choisit d'effectuer l'étape de désambigüisation et d'expansion de la requête avant l'étape de traduction et ceux afin d'améliorer le résultat de cette dernière.

La méthode de désambigüisation est basée sur une désambigüisation manuelle, c'est-à-dire que l'utilisateur doit intervenir pour choisir le sens dénoté par les mots de la requête(les sens sont tirés de la base lexicographique WordNet 2.0). Nous avons évalué les différentes stratégies d'expansion basées sur les relations sémantiques (hyperonymie, hyponymie et synonymie) et ceux, pour déterminer la meilleure d'entre elles. Nous avons comparé en premier lieu les hyperonymes et les hyponymes et étant donné que ces deux derniers se présentent sous forme d'hiérarchie (niveaux), nous avons du les combiner pour pouvoir effectué l'évaluation. Cette évaluation consiste à faire varier le niveau afin de déterminer l'impact de ce dernier sur la qualité des

Rapport-gratuit.com
LE NUMERO 1 MONDIAL DU MÉMOIRES

résultats recensés. A l'issus de cette première évaluation nous concluons les meilleurs résultats sont fournis par la combinaison de la relation d'hyperonymie (niveau 1) et de la relation d'hyponymie (niveau 0), c'est-à-dire le cas ou l'expansion se fait uniquement à l'aide des hyperonymes (niveau 1), en effet la précision est de 0.6071. Nous avons poursuivi notre évaluation en comparants ces résultats à ceux fournis par les synonymes, et ce sont ces dernier qui offrent les meilleurs.

Au final et afin de déterminer l'impact de la stratégie d'expansion basée sur la relation de synonymie, nous avons comparé ces résultats à ceux du test de traduction sans désambigüisation et sans expansion. Nous avons clairement remarqué que le premier cas offre les meilleurs résultats. En effet l'amélioration est de 3,72% pour la précision et de 23,33% pour le rappel.

L'amélioration relative à cette démarche a permit d'une part, de se focaliser sur le sens dominant de ces requêtes et de les enrichir avec des termes reliés sémantiquement à ceux des requêtes et d'autre part, d'améliorer la qualité des requêtes traduites et donc d'améliorer la qualité des résultats recensés.

Ce qu'il faut tirer des différentes expérimentations que nous avons réalisées est qu'il faut essayer au départ de tirer profit du contexte de la requête afin d'en extraire les concepts dominants et ceux, avant d'effectuer la traduction. Dans le cas contraire, c'est-à-dire celui ou nous effectuons d'abord une traduction puis une désambigüisation, le risque de perdre les concepts de base véhiculés par la requête est très élevé.

A l'issus des travaux élaborés dans ce manuscrit, notre perspective dans un premier temps est de consolider la démarche proposée en l'évaluant sur d'autre collections, puis élargir notre domaine d'investigation vers d'autres paires de langues.

Rapport-Gratuit.com

References bibliographiques

[Allan & al, 98] J. Allan & J. Callan & M. Sanderson & J. Xu & S. Wegmann, INQUERY at TREC-7, Proceedings of TREC-7, 1998.

[Ballestros & al, 96] L. Ballesteros & W. Croft, Dictionary methods for cross-lingual information retrieval in proceedings of DEXA'96, 1996.

[Ballestros & al, 97] L. Ballesteros & W Croft, Phrasal translation and query expansion techniques for cross-language information retrieval In Proceedings of ACM SIGIR'97, 1997.

[Ballestros, 96] Ballesteros L & Croft W, Dictionary methods for cross-langual information retrieval, In Proceedings of DEXA'96, 1996.

[Ballestros, 98] Ballesteros L & Croft W, Resolving Ambiguity for Cross-Language Retrieval, In Proceedings of the 21st ACM SIGIR'98, 1998.

[Baziz, Boughanem & Nassr, 04] Mustapha Baziz, Mohand Boughanem & Nawel Nassr,La recherche d'information multilingue : désambiguïsation et expansion de requêtes basées sur WordNet, Laboratoire IRIT/SIG, Campus Universitaire Toulouse III, 2004.

[Boucham, 09] Boucham Souhila, Une approche basée Ontologies pour l'indexation automatique et la recherche d'information Multilingue, mémoire de magister, Université M'hamed Bougara Boumerdes, 2009.

[Boughanem, 00] Boughanem M & Julien C & Mothe J & Soule-Dupuy C, Mercure at TREC8 In Proceedings of TREC-8, 2000.

[Caro, 97] Budi Yuwono, Savio L.Y. Lam, Jerry H. Ying, Dik L. Lee, A World Wide Web Ressource Discovery System, 1997.

[Davis, 96] Davis M., Dunning T. E, A TREC evaluation of query translation methods for multi-lingual text retrieval, In Proceedings of TREC-4, 1996.

[Davis, 97] Davis M, New Experiments in cross-langage test retrieval In Proceedings of TREC-5, 1997.

[Davis, 98] M.Davis, On the effective use of large parallel corpora in cross-Language Text Retrieval in Grefenstette (ED) Cross language information retrieval, Kluwer Academic Publisher Boston, 1998.

[Deerwester & al, 90] S. Deerwester, S. Dumais, S. Furnas, G. Landauer & R.Harshman, Indexing by Latent Semantic Analysis: Journal of the American Society for Information Science, 1990.

[Diekema & al, 98] A. Diekema & F. Oroumchian & P. Sheridan & E. Liddy TREC-7 Evaluation of conceptual interlangua document retrieval (CINDOR) in English and French, In Proceedings of TREC-7, NIST Special Publication, 1998.

[Gey & al, 97] F. Gey & H. Jiang & A. Chen & R. Larson, Manual Queries and Machine Translation in Cross-Language Retrieval and Interactive Retrieval with Cheshire II In Proceedings of TREC-7, NIST Special Publication 500-242, 1997.

Rapport-gratuit.com
LE NUMERO 1 MONDIAL DU MÉMOIRES