Les stratégies de traduction de requêtes en recherche d'information par croisement de langues

Nawel NASSR, Mohand BOUGHANEM
IRIT/SIG
Campus Univ. Toulouse III
118, Route de Narbonne
F-31062 Toulouse Cedex 4
E-mail : {boughane,nassr}@irit.fr

1 Introduction

Le multilinguisme, collections de documents multilingue, devient de plus en plus une réalité incontournable. En effet, le développement explosif de l'Internet avec ses collections d'information disséminées aux quatre coins de la planète, conduit le réseau des réseaux naturellement vers un multilinguisme de fait. Eliminer les barrières de la langue en permettant à un Système de Recherche d'Information (SRI) de retrouver des documents pertinents exprimés dans une langue autre que celle de la requête, est une tâche importante à laquelle la communauté de la Recherche d'Information (RI) s'intéresse de plus en plus.

L'évaluation de requêtes dans les SRI actuels privilégie la présence ou l'absence d'un mot dans un document, même si les modèles et les mesures de poids différent. En ce sens, il est peu probable que des documents qui ne contiennent aucun terme de la requête (initiale ou étendue) soient sélectionnés. Ce problème est encore plus crucial dans le cas de collections de

documents multilingues, car les requêtes et les documents peuvent être exprimés dans des langues différentes. Dans ce cas précis, nous avons peu de chance de sélectionner les documents pertinents écrits dans une langue différente de celle de la requête.

La notion de multilinguisme en RI peut se présenter sous différentes facettes dont les deux principales sont les suivantes :
- dans la première, le multilinguisme peut être vu comme une collection de documents multilingues que l'on interroge avec des requêtes pouvant être exprimées dans différentes langues, mais, les documents sélectionnés sont écrits dans la langue de la requête. Dans ce cas on restreint l'espace des documents recherchés à la langue de la requête. Ceci constitue en fait une recherche d'information monolingue et ne pose pas de problèmes particuliers,

- dans la deuxième, le multilinguisme concerne la possibilité offerte à un SRI de sélectionner des documents exprimés dans une langue différente de celle de la requête. C'est ce que l'on appelle la recherche d'information par *croisement de langues* (cross-language) [12]. Ceci nécessite donc la considération et le choix entre trois niveaux de traductions :

- *traduction au niveau de la requête* : il s'agit dans ce cas de traduire la requête vers la langue des documents, et de présenter au moteur de recherche les différentes traductions de la requête dans les différentes langues souhaitées. Le système récupérera par la suite les différents documents correspondants à chaque traduction.
- *traduction au niveau des documents* : ici à l'inverse, ce sont les documents qui sont traduits vers la langue de la requête. Le système procède ensuite à une simple interrogation monolingue. Cette méthode emploie des outils de traduction automatique de documents. Son principal inconvénient est lié à la taille de la base. En effet, le nombre de documents à traduire est

important et il faut pouvoir stocker les différentes traductions. De plus le temps de traduction est considérable dans les bases volumineuse. C'est pourquoi cette méthode, bien que souvent décrite, ne trouve pas de réelles applications par manque sensible de faisabilité.
- traduction des documents et de la requêete : dans ce cas, il s'agit de trouver une représentation dans laquelle on décrit la requête et les documents. Pour cela, il faut faire une traduction des deux dans une représentation commune.

Actuellement, La recherche dans ce domaine se base principalement sur la traduction de requête car elle est moins coûteuse [17].
C'est précisément, cette deuxième facette qui nous intéresse. En effet, l'objectif de cet article est d'étudier deux techniques de traduction de requêtes:
- une première technique basée sur l'utilisation d'un dictionnaire,
- une deuxième technique basée sur l'utilisation des associations entre termes. Ces dernières sont établies à partir d'une collection de documents parallèle.

La section 2 de cet article présente un bref état de l'art sur le croisement de langues en recherche d'information. La section 3 décrit les trois techniques que nous avons proposées pour la traduction de requête. La section 4 présente les expérimentations effectuées sur la base Amaryllis et les différents résultats obtenus.

2 Croisement de langues en RI: Etat de l'art

Les travaux effectués dans le domaine de la recherche d'information par croisement de langues se sont principalement focalisés sur la traduction de requête. Dans ce contexte, ces systèmes tentent de résoudre trois problèmes majeurs.

- Le premier problème concerne *la traduction* des termes de la requête. Dans ce cas on essaie de substituer chaque terme exprimé dans la langue source (ls) par un ou plusieurs terme(s) censés le représenter dans la langue cible (lc).
  - Le second problème est la *désambiguisation*. Ce dernier traduit le fait qu'un terme dans la langue source ls possède plusieurs traductions. Il consiste à choisir la ou les meilleure(s) traduction(s) dans la langue cible lc.
  - Le troisième problème concerne la *pondération* des termes traduits.

Les méthodes de traduction de requêtes proposées sont basées sur l'utilisation de :

Dictionnaires bilingues: l'idée principale de techniques basées sur les dictionnaires[2],[1], [12] est de remplacer chaque terme de la requête par le ou les terme(s) approprié(s) dans la langue cible. 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 aussi pour cela que les méthodes basées sur des dictionnaires sont les plus utilisées dans la recherche d'information par croisement de langues.

Corpus alignés (parallèles ou comparables): les méthodes basées sur le corpus [15], [16],[6] et [5], utilisent directement le contenu d'un ensemble de documents, regroupés dans un corpus, pour déduire des relations entre les termes de langues différentes. L'alignement de documents consiste à mettre en correspondance des documents de langues différentes selon un critère donné. Nous distinguons deux types d'alignement : alignement parallèle et alignement comparable.

- L'alignement parallèle consiste à mettre en correspondance chaque document d'une langue l1 avec le document représentant sa traduction dans la langue l2. Dans ce cas on peut réduire la taille des

parties étudiées comparativement afin de mieux cibler les correspondances, ainsi l'alignement peut être fait sur: le document, les paragraphes, les phrases ou les termes.*

- L'alignement comparable plus délicat à réaliser [22], revient à mettre en correspondance des documents en se basant sur des critères comme par exemple la présence de mêmes dates, de mêmes noms de personnes dans des documents de langues différentes [12], [17],. Dans ce cas, on n'utilise plus des documents et leurs traductions, mais des documents de différentes langues dont le contenu est proche; en effet l'alignement se fait selon le contexte des documents et ne peut être réalisé que sur le document en entier.

- Traducteurs automatiques : les techniques basées sur les traducteurs automatiques sont principalement employées lors de la traduction des documents. Ces systèmes sont généralement plus complexes et loin d'être parfaits[14],[18] et [19] car ils s'appuient sur des grammaires et autres méthodes linguistiques et 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 succès du fait que ces dernières sont souvent courtes et exprimées par des mots indépendants.

L'utilisation d'un dictionnaire, quand une version électronique de celui-ci existe et est facilement exploitable, est le moyen le plus simple pour réaliser la traduction de requêtes. De nombreux travaux ont exploré cette direction [1],[3] [13],[7] et [8]. Ces travaux diffèrent souvent par leur fa\c{c}on d'aborder la désambiguisation.

Plus précisément, Davis [8],[9] exploite une version électronique du Collins pour réaliser la traduction des termes de la requête de l'anglais vers l'espagnol. Son approche de désambiguisation est basée sur l'utilisation d'un corpus parallèle. Elle consiste \`a sélectionner pour chaque terme en anglais de la requête

source le meilleur terme en espagnol parmi les substitutions possibles. Ce terme en espagnol est trouvé de la fa\c{c}on suivante. La requête en anglais et la requête en espagnol sont toutes les deux évaluées sur un corpus parallèle. En se basant sur les documents sélectionnés par ces requêtes, une similarité est calculée entre le vecteur documents de chaque terme anglais et les vecteurs documents de ces substituants possibles. Le meilleur substituant du terme anglais est celui de plus grande valeur de similarité.

Ballestros [1],[3] aborde le problème de désambiguisation en se basant sur la co-occurrence entre termes. Des valeurs de co-occurrences sont calculées entre les termes anglais et espagnols en se basant sur un corpus aligné (document espagnol aligné avec document anglais). La traduction des termes de la requête est effectuée en se basant sur une version électronique du Collins (Anglais-Espagnol). La désambiguisation consiste \`a retenir pour chaque terme anglais le terme espagnol le plus co-occurrent parmi les substitutions possibles.

Dans les méthodes basées sur le corpus, les termes de la requête sont traduits en utilisant des associations entre termes dérivées à partir de corpus alignés.

Davis [7] remplace les termes de la requête source (Anglais) par les 100 termes les plus fréquents provenant des 100 premiers documents espagnol. Ces derniers sont alignés avec les documents en anglais répondant à la requête.

Sheridan [22] construit automatiquement un thesaurus de similarité ou de co-occurrence, à partir d'un corpus comparable, produisant pour chaque terme dans la langue source une liste ordonnée de termes jugés similaires dans la langue cible. Ce thesaurus est utilisé par la suite pour effectuer la traduction de requête.

Yamabana [24] a développé une méthode de désambiguisation utilisant un corpus comparable. L'approche proposée consiste à calculer automatiquement à partir de ce corpus comparable l'ensemble de valeurs de cooccurrence entre les terme de la langue source et

les termes de la langue cible. Ce thesaurus est utilisé pour sélectionner la meilleure traductions en langue cible.\\

Les travaux basés sur la traduction automatique de [23],[17] et [11] ont montré des performances plus faibles que les techniques de traduction citées ci-dessus. 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 [20].

3 Description générale du processus de recherche d'information par croisement de langues

L'approche générale que nous adoptons en RI par croisement de langues est représentée schématiquement par la figure 1

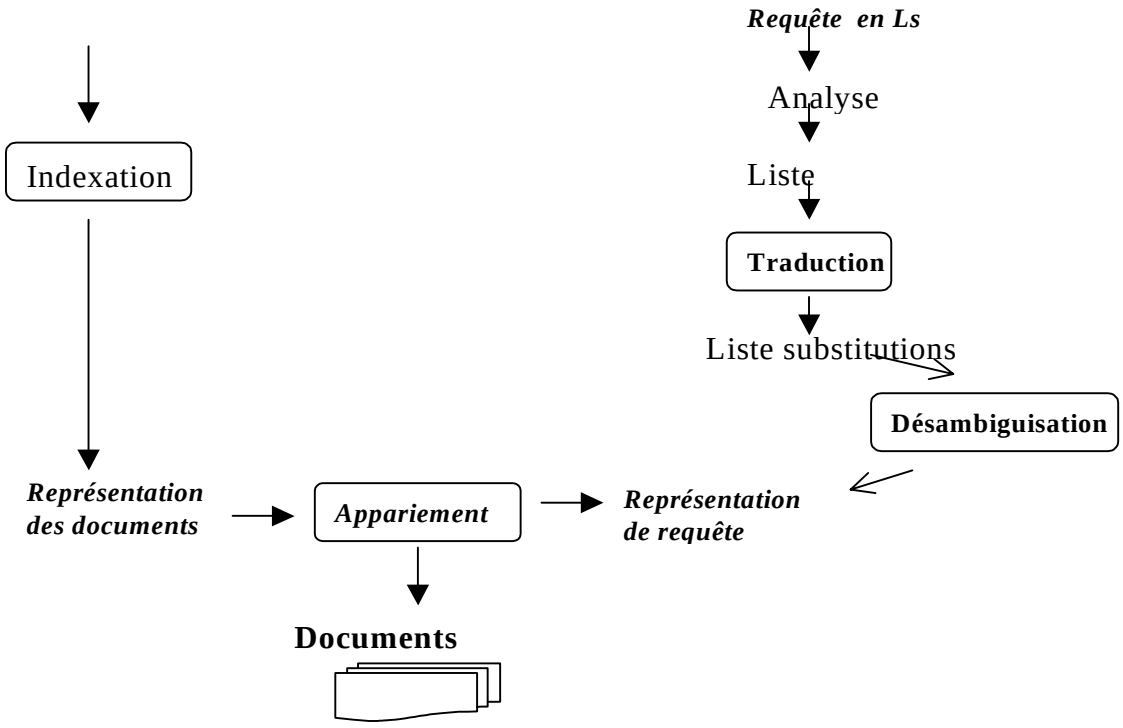


FIG. 1 : Approche générale pour CLIR

On y distingue 4 fonctions principales:

- L'indexation.
- L'analyse de la requête.
- La traduction.
- L'appariement requête-document.

3.1 L'indexation :

L'opération d'indexation automatique consiste d'une façon générale, à extraire les mots d'un document puis à les radicaliser, soit par troncature pour le français, l'italien et l'allemand ou bien par l'algorithme de Porter pour les textes en anglais. Cette opération a été testée seulement sur ces quatre langues. L'élément intéressant dans cette opération est la reconnaissance automatique de la langue. Elle

consiste d'une façon générale à comparer la liste des mots simples extraits d'un document à chacune des listes de mots vides disponibles, soit dans notre cas une liste pour chacune des langues traitée français, anglais, italien, allemand. La comparaison revient à compter le nombre de mots communs entre le document et chacune des listes de mots vides. La langue du document correspond à celle de la liste des mots vides avec laquelle il a le plus grand nombre de mots communs.

3.2 L'analyse de la requête :
Le texte libre de la requête source est analysé afin d'extraire tous les mots clés. Les mots vides de la requête sont aussi éliminés. Le résultat de cette opération est une liste de mots significatifs avec un poids associé à chaque terme.

3.3 La traduction :
Cette opération traduit chaque terme de la requête source en une ou plusieurs traductions. Trois techniques de traduction sont proposées.

- Une première basée sur l'utilisation des dictionnaires.
- Une deuxième basée sur l'utilisation des associations entre termes.
- Une troisième basée sur le contexte local.

3.3.1 Technique basée sur le dictionnaire
Cette approche utilise un dictionnaire bilingue pour réaliser le processus de traduction. Ce dernier est un ensemble de termes en langue L1 alignés avec d'autres termes en langue L2

Dans ce cas, la traduction de la requête consiste à rechercher dans ce dictionnaire et à remplacer chaque terme de la requête source (ie. requête initiale exprimée dans la langue source) par le/les terme(s) les plus adéquat(s) dans la langue cible. Le résultat de cette traduction dépend principalement de la qualité du dictionnaire utilisé pour la traduction.

3.3.2 Technique basée sur les associations entre termes

L'idée sous jacente à cette approche qui s'inscrit dans le cadre des approches basées sur le corpus, est la possibilité de déduire automatiquement des associations entre termes de la langue source et ceux de la langue cible à partir de leur distribution dans des documents alignés. L'alignement des documents détermine la façon de faire correspondre deux documents. Plus l'alignement est fin, c'est à dire plus la taille des blocs (terme, phrase, paragraphe ou document) à aligner n'est pas importante, plus la qualité de traductions devrait être bonne. Notons que l'alignement par terme est l'alignement le plus fin à effectuer. En effet, on fait correspondre un terme avec une liste plus réduite de traductions possibles, ce qui sur l'ensemble du corpus, doit permettre de trouver le terme qui est employé dans les documents traduits.

Les associations qui sont construites automatiquement à partir du corpus aligné sont utilisées comme moyen pour traduire les termes de la requête. Chaque terme de la requête source est remplacé par le(s) terme(s) avec lesquels il est associé.
Notre travail consiste à construire ces associations automatiquement en se basant sur la co-occurrence des termes dans le corpus aligné. Cette co-occurrence est mesurée seulement entre les termes issus des documents de la langue source (l_s) et ceux issus des documents de la langue cible (l_c). Ces associations sont établies en utilisant la formulation suivante :

$$\mathrm{co}\,(t_i^{ls},t_j^{lc}) = \frac{\sum_{k=1}^{N}\min(d_{ik}^{ls},d_{jk}^{lc})}{\sum_{k=1}^{N}d_{ik}^{ls}+\sum_{k=1}^{N}d_{jk}^{lc}-\sum\min(d_{ik}^{ls},d_{jk}^{lc})}$$

Où

t_i^{ls},t_j^{lc} : respectivement le terme i dans la langue source (l_s) et le terme j dans la langue cible (l_c),

d_{ik}^{ls} : poids du terme i dans le document d_k^{ls}. Noter que d_k^{ls} est aligné avec d_k^{lc},
N : le nombre de documents du corpus pour une langue.

Cette formule permet de calculer un rapport entre l'apparition d'une paire de termes t_i^{ls} et son éventuelle traduction t_j^{lc} et les apparitions de chacun des termes dans la base. Le résultat obtenu est dans l'intervalle [0 1], une cooccurence qui vaut 1 indiquant une correspondance parfaite entre deux terme, c'est a dire qu'un terme traduit apparaît à chaque fois que le terme source apparaît. Dans cette approche, la meilleure traduction à retenir pour chaque terme dans la langue source est le terme de la langue cible le plus co-occurent parmi les substituants possibles.

3.3.3 Technique basée sur le contexte local

Cette technique exploite les propriétés des corpus parallèles, pour traduire la requête de la langue source L_1 en langue cible L_2. La figure 2 décrit notre technique.

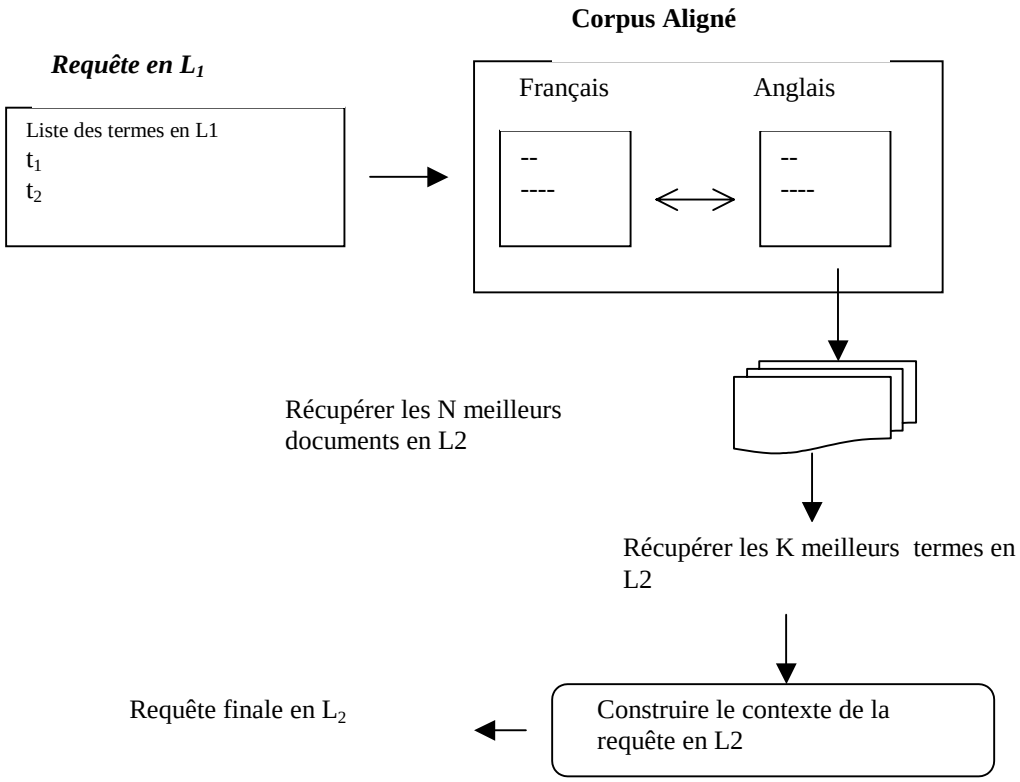


FIG. 2 *: Technique basée sur le contexte local*

Les points suivants illustrent notre technique :

1. Chaque requête en langue source est évaluée sur les documents en langue source du corpus parallèle. Les N meilleurs documents pertinent en langue source sont sélectionnés.

2. On récupère les N documents en langue cible dans le corpus parallèle. Ces documents sont alignés avec les N documents en langue source .

3. Pour les N documents en langue cible, on récupère les K meilleurs termes qui les représentent.

4. La requête en langue cible, est alors représentée par l'ensemble des K meilleurs termes sélectionnés dans la langue cible.

3.4 Appariement requête-document:

L'appariement s'effectue sur la base des représentations des documents et des besoins en informations; il permet de sélectionner des documents à l'utilisateur en réponse à son besoin. Cette opération permet d'associer une mesure appelée pertinence système, supposée représenter la pertinence d'un document vis a vis d'une requête. Cette pertinence est calculée à partir d'une fonction de similarité, notée RSV(Q,d) (Retrieval Status Value), où Q est une requête et d un document. Cette mesure tient compte des poids des termes déterminés en fonction d'analyses statistiques et probabilistes. Notons que ce processus est étroitement lié aux représentations des documents et des requêtes. En effet, si l'opération d'indexation est la même dans la plupart des modèles de recherche d'information, ces derniers diffèrent souvent par rapport aux fonctions utilisées pour la mesure des poids et pour l'appariement requête-document.

4 Expérimentations et résultats

Le but de nos expérimentations est de montrer la faisabilité des deux techniques de traduction. L'expérimentation a été effectuée sur un corpus de documents issus du programme Amaryllis[1] Ce corpus est composé de deux collections de documents EldaFr (les documents en français) et EldaAng (les documents en anglais). Le tableau 1 montre les caractéristiques de ces deux collections.

[1] Amaryllis: projet français de test et d'évaluation de systèmes de Recherche d'information , http://www.inist.fr/ama/pages/corp.html

Collection Amaryllis	**EldaFr**	**EldaAn**
Taille de la base.	3511	3511
Nombre de termes	16917	14797
Taille moyenne du Document(terme)	164	263

TAB. 1:*Les caractéristiques de la collection Amaryllis*

Pour nos tests, nous nous sommes restreints à un croisement sur la paire de langues Français-Anglais. Les expérimentations effectuées consistent à sélectionner les documents en anglais pour des requêtes exprimées en français. Les tests ont été effectués sur 15 requêtes issues d'Amaryllis. La sélection des documents est effectuée par le système Mercure développé au sein de l'équipe SIG de l'IRIT[4].

Les documents sont indexés par le SRI Mercure[4] où chaque document est représenté par une liste de termes pondérés. Le processus de recherche de document pertinent consiste à comparer la liste des termes qui représente la requête avec celles qui représentent les documents. Une mesure de similaritée est calculée entre ces listes pour sélectionner une liste de documents ordonnée pour la requête exprimée dans la langue source.

Trois groupes de tests ont été réalisés. Le premier concerne l'utilisation d'un dictionnaire pour traduire les requêtes. Un dictionnaire bilingue (français-anglais) obtenu gratuitement via Internet sur le site http://www.freedict.com est utilisé. Il est représenté par une liste simple de 35200 termes en français alignés avec d'autres termes en anglais.

Le deuxième groupe concerne l'utilisation des associations entre termes établies automatiquement à

partir des documents parallèles issus d'Amaryllis. Ces associations sont utilisées pour traduire chaque terme de la requête du français vers l'anglais. Le terme en anglais le plus co-occurent avec le terme en français de la requête initiale est retenu.

Le troisième groupe concerne l'utilisation de la technique basée sur le contexte local et qui utilise les documents parallèles issus d'Amaryllis pour construire le contexte de la requête en anglais.

Dans le but de mesurer l'efficacité des techniques de traduction, nous avons comparé les résultats obtenus par ces deux techniques avec ceux obtenus par le test monolingue anglais. Le test monolingue consiste à utiliser 15 requêtes en anglais correspondant à la traduction exacte des 15 requêtes en français. Ces requêtes qui ont été fournis dans Amaryllis sont comparées à la collection de documents en anglais.

L'évaluation des performances des deux techniques est effectuée sur l'ensemble des documents sélectionnés pour les 15 requêtes. Pour chaque requête on retient les 250 premiers documents sélectionnés. Cette évaluation est effectuée en se basant sur les mesures de rappels et de précision. Les précisions à différents points p5, p10, p15, p20, p30, p100 représentant le nombre de documents pertinents parmi les 5, 10, 15, 20, 30, 100 premiers documents, et une précision moyenne (AvgPr) sur l'ensemble des documents sélectionnés.

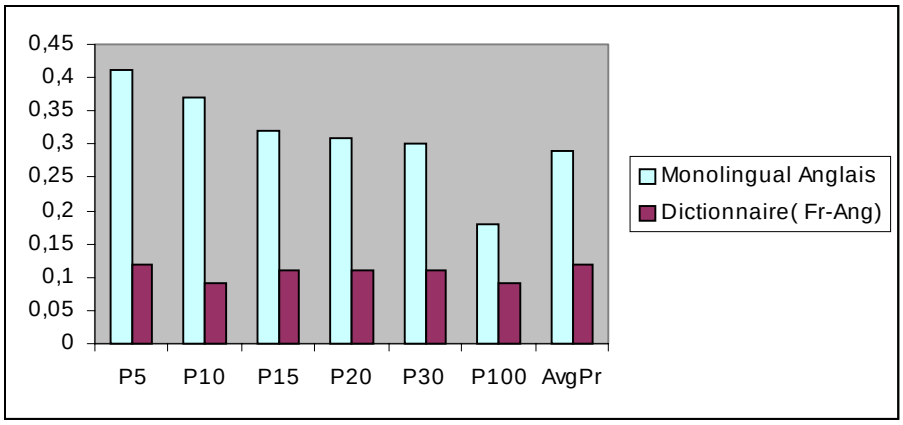
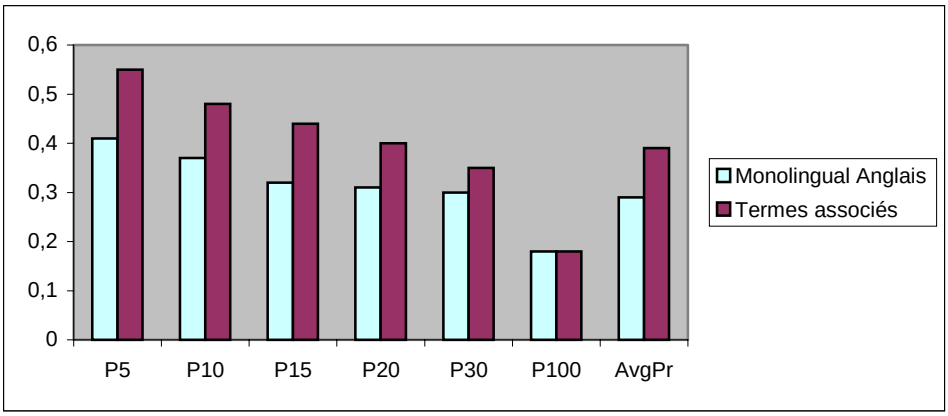


FIG.3 *Comparaison entre le dictionnaire et le monolingue anglais*

La figure 3 compare les résultats du monolingue à ceux obtenus par le dictionnaire. On remarque clairement que les résultats du monolingue sur toutes les précisions sont meilleurs que ceux obtenus par le dictionnaire. Plus précisément, la précision moyenne (AvgPr) est de 0.12 contre 0.29 pour le monolingue. Ce résultat est dû à l'existence de plusieurs traductions pour chaque terme source (le problème d'ambiguité) et aux termes du dictionnaire qui sont totalement indépendants des



termes des documents de la collection de test. De plus, le dictionnaire utilisé est général et ne traite pas le même domaine que les documents de la collection de test.

FIG. 4 : *Comparaison entre les associations entre termes et le monolingue anglais*

La figure 4 compare le monolingue anglais aux associations entre termes. On y constate que les résultats obtenus par les associations entre termes sur toutes les précisions sont meilleurs que ceux obtenus par le test monolingue. La précision moyenne (AvgPr) est de 0.39 contre 0.29 pour le monolingue. Ce résultat peut être expliqué par le fait que les associations entre les termes sont déduites à partir d'un corpus de documents qui traitent les mêmes sujets que les documents de la collection de test. L'ambiguité lors de la traduction est alors diminuée.

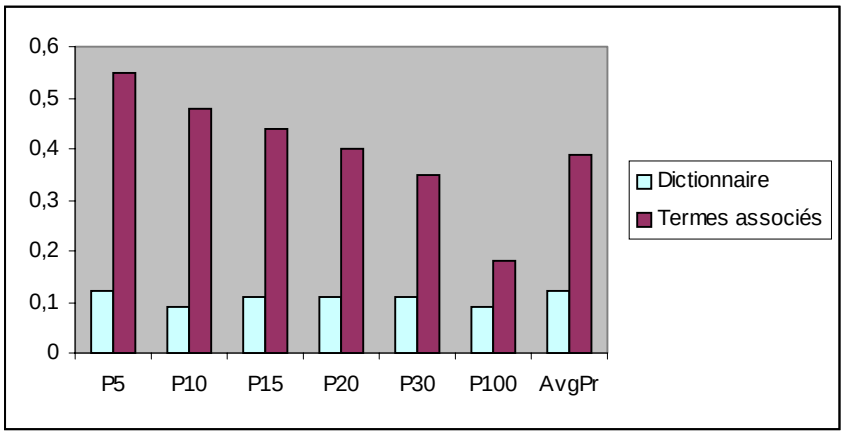


FIG. 5 : *Comparaison entre le dictionnaire et les associations entre termes*

La figure 5 compare les deux techniques de traductions. On remarque clairement que l'approche basée sur les associations entre termes est meilleure que celle basée sur le dictionnaire. La précision moyenne (AvgPr) est de 0.12 pour le dictionnaire contre 0.39 pour les associations entre termes. Les résultats obtenus sur toutes les précisions par les associations entre termes sont meilleurs que ceux obtenus par le dictionnaire.

Cette différence se traduit par la différence entre les ressources utilisées dans les deux techniques. Le dictionnaire utilisé est général et ne traite pas le même domaine que les documents de la base alors que le domaine des associations entre termes est lié au domaine des documents de la collection de test.

id-Exec	P5 P10 P15 P30	Exact Prec.Moy
Monolingue Anglais	0.4133 0.3667 0.3156 0.3000	0.3113 0.2945
Termes associés	0.5467 0.4800 0.4400 0.3489	0.3933 0.3854
Amélioration (\%)	32.27 23.60 28.27 14.01	20.84 23

TAB. 2 : *Impact de la technique basée sur les associations entre termes*

Le tableau 2 compare les différentes valeurs trouvées par le test monolingue anglais avec celles trouvées par les termes associés sur plusieurs précisions de documents (5 docs, 10 docs, 15 docs, 30 docs). On remarque que sur toutes les précisions (P5, P10, P15, P30, Exact, Prec.Moy), notre technique donne une amélioration importante par rapport au test monolingue. Plus précisément la précision moyenne augmente de 23\%.

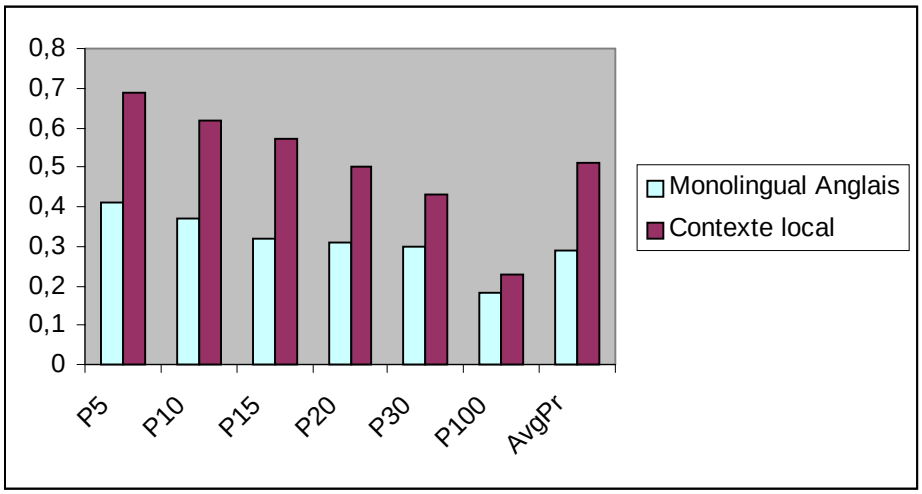


FIG. 6 *: Comparaison entre le contexte local et le monolingue anglais*

La figure 6 compare le monolingue anglais à la technique basée sur le contexte local. On y constate que les résultats obtenus par la technique basée sur le contexte local donnent de meilleurs performance par rapport au monolingue anglais sur toutes les précisions de documents. La précision moyenne (AvgPr) est de 0.5 contre 0.29 pour le monolingue. Ce résultat peut être expliqué par le fait que les documents du corpus traitent les mêmes sujets que les documents de la collection de test.

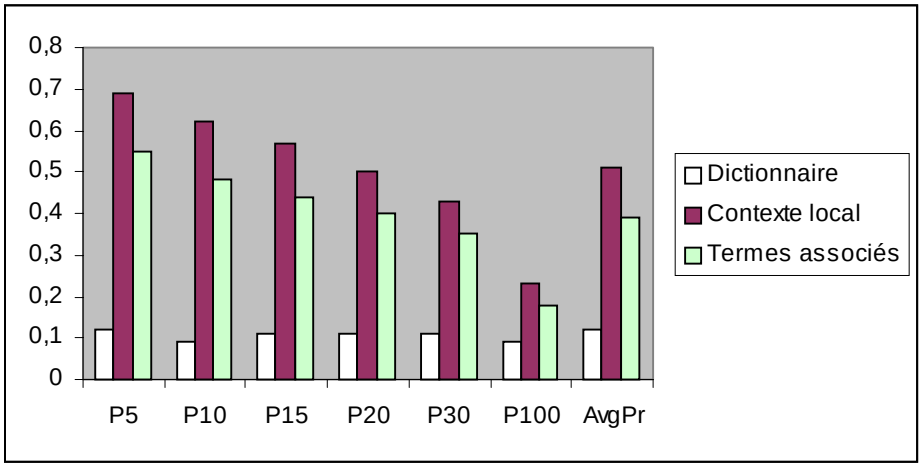


FIG. 7 : *Comparaison entre les trois techniques*

La figure 7 compare les trois techniques de traduction. On remarque clairement que l'approche basée sur le contexte local est meilleure que les associations entre termes et la technique basée sur le dictionnaire. La précision moyenne (AvgPr) est de 0.5 pour le contexte local contre 0.39 pour les associations entre termes et 0.12 pour le dictionnaire.

5 Conclusion

Cet article présente trois techniques de traduction de requêtes en recherche d'information par croisement de langues.

- La première concerne l'utilisation d'un dictionnaire bilingue (Français-Anglais),
- La deuxième concerne l'utilisation des associations entre termes calculées automatiquement à partir d'un corpus parallèle d' Amaryllis.
- La troisième est basée sur le contexte local.

Ces techniques ont été testées en utilisant le moteur de recherche Mercure.
Nous avons montré à l'issue de ces tests que l'utilisation du dictionnaire bilingue (Français-Anglais) et l'utilisation des corpus parallèles sont des solutions viables pour réaliser le croisement de langues.

Les résultats obtenus par la technique du dictionnaire sont moins performamts que ceux obtenus par le test monolingue anglais et se traduit par le fait qu'on a utilisé un dictionnaire bilingue qui est général et incomplet et par le fait qu'un terme donné en langue cible peut avoir plusieurs traductions en langue cible (problème d'ambiguité).
Le meilleur résultat par contre est obtenu par la technique des associations entre termes et la technique basée sur le contexte local. Ce résultat se traduit par la relation qui existe entre les documents du corpus et les documents de la base. Le corpus traite le même domaine que les documents de la collection de test alors que le domaine du dictionnaire est totalement indépendant du domaine des documents de cette collection.

6-Perspectives

Les techniques que nous avons proposées dans cet article et plus particulièrement les résultats obtenus par la technique basée sur les dictionnaires sont très insuffisants. Cette insuffisance est dû au fait qu'un terme dans la langue source peut avoir plusieurs termes en langue cible (problème d'ambiguïté).
Pour résoudre ce problème et pour améliorer les résultats du dictionnaire, d'autres techniques de traductions qui combine le dictionnaire et les corpus parallèles sont proposées et en cours de test:

- La technique basée sur les mesures de similarités: dans ce cas nous allons tester deux mesures de similarité. La première utilise l'ordre

lexicographique et la deuxième utilise les coefficients de corrélation. Ces mesures de similarité sont calculées automatiquement à partir d'un corpus.

- La technique basée sur la traduction Bi-Directionnelle. Dans ce cas un dictionnaire bilingue est utilisé pour résoudre le problème d'ambiguïté.

Ces deux techniques qui sont en cours de test s'inscrivent dans le cadre d'un projet européen e-Court (http://laplace.intrasoft-intl.com/e-court/) auquel nous participons depuis juin 2000.

Références Bibliographiques

[1] Ballesteros L., Croft W. (1996). Dictionary methods for cross-langual information retrieval. In Proceedings of DEXA'96, pages 791-801.

[2] Ballesteros L., Croft W. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. in Proceedings of ACM SIGIR'97, pages: 84-91.

[3] Ballesteros L., Croft W. (1998). Resolving Ambiguity for Cross-Language Retrieval}. in Proceedings of the 21st ACM SIGIR'98, pages 64-71.

[4] Nassr.N, Boughanem M.(2000). Mercure at CLEF-1. In Proceedings of CLEF-2000 Cross Language Evaluation Forum, Lecture Note in Computer Science 2069, Spring Verlag, pages:202-210

[5] Braschler M, Krause J, Peter C, Schauble P. (1998). Cross-Language Information Retrieval (CLIR) Track Overview. in Proceedings of TREC-7.

[6] Carbonell J, Yiming Y,Frederking R,Brown R, Yibing G, Danny L. (1997). Translingual information retrieval: a comparative evaluation. in Proceedings of the Fifteenth International Joint conference on Artificial Intelligence, pages: 708-714 available in http://www.cs.cmu.edu/~yiming/papers.yy/ijcai97-6.ps

[7] Davis M., Dunning T. E. (1996). A TREC evaluation of query translation methods for multi-lingual text retrieval. In Proceedings of TREC-4, pages 483-497.\footnote{Ces proceedings sont dispondibles sur le site http://trec.nist.gov/publications} Conference.

[8] Davis M., (1997). New Experiments in cross-langage test retrieval. In Proceedings of REC-5, pages: 447-454.

[9] Davis M. (1998). On The Effective Use of Large Parallel Corpora in Cross-Language Text Retrieval. In Cross-Language Information Retrieval, Edited by Gregory Grefenstette, Kluwer Academic Publishers, pages: 11-21.

[10] Denjean P. (1989). Interrogation d'un système vidéotext arborescent : l'indexation des textes}. Thèse de doctorat de l'université Paul Sabatier.

[11] Gey F. (1999). Manual Queries and Machine Translation in Cross-Language Retrieval and Interactive Retrieval with Cheshire II. In Proceedings of TREC-7, page: 527-540.

[12] Grefenstette G. (1998). The Problem of Cross-language Information Retrieval. In Cross-Language Information Retrieval, Edited by Gregory Grefenstette, Kluwer Academic Publishers, pages: 1-9.

[13] Hull D., Grefenstette G. (1996). Querying across languages. A dictionary-based approach to multilingual information retrieval. In Proceedings of ACM-SIGIR'96, pages: 49-57.

[14] Gachot.D, Yang S, Lang E.(1998). The Systran NLP browser : An application of machine translation technique in Multilingual Information Retrieval. In Edited by Gregory Grefenstette, Kluwer academic, pages:105-117

[15] Landauer, Littman M.(1990). Fully Automatic cross-language document retrieval using latent semantic indexing. In Proceedings of the Sixth

Annual Conference of UW Center for the New OED and Text Research, pages: 31-38.

[16] Nie j-y.(1998). Trec-7using a Probabilistic Translation Model}. In NIST Special Publication, the Seventh Text Retrieval Conference (TREC-7).

[17] Oard D., Dorr B. (1996). A Survey of Multilingual Text Retrieval. Report UMIACS-TR-96-19 CS-TR-3615 (http://www.clis.umd.edu/dlrg/filter/papers/mlir.ps).

[18] Oard D. (1998). A comparative study of query and document translation for Cross Language Information Retrieval. In Farwell (ed), Report AMTA'98. Proceeding. Springer-Verlag berlin.

[19] Oard D., Hackette B. (1997). Document translation for Cross-Language Text Retrieval at the University of Maryland. In Proceedings of TREC-6.

[20] Pirkola A. (1998). The Effects of Query Structure and Dictionary Setups in Dictionay-Based Cross-laguage Information Retrieval. In Proceedings of ACM-SIGIR'98, pages: 55-63.

[21] Privat R. (1999). Conception et Réalisation d'un Système de Croisement de langues}. Rapport de DEA, 2IIL de l'université Paul Sabatier de Toulouse.

[22] Sheridan P., Ballerini J. P. (1996). Experiments in multilingual information retrieval using SPIDER system. In Proceedings of ACM SIGIR'96, pages: 58-65.

[23] Yamabana K., Muraki F., Doi S., Kamei S. (1996). A Language Conversion Front-end for Cross-Linguistic Information Retrieval. In Proceedings of ACM-SIGIR'96, pages: 43-39.

[24] Yamabana K., Muraki F., Doi S., Kamei S. (1998). A Language Conversion Front-end for Cross-Linguistic Information Retrieval. In G.Grefenstette(ed) Cross-Language Information Retrieval, Chapter 8. Kluwer Academic Publisher, Boston 1998.

Revue
D'Information
Scientifique &
Technique

Centre de Recherche sur l'Information Scientifique et Technique

FORMULAIRE CONFIDENTIEL D'EVALUATION
D'UN PROJET DE PUBLICATION

N° de code : ……
Date d'envoi : ……
Nom du référé : ……
Intitulé : ……
……

EVALUATION DE L'ARTICLE

Mettre une croix dans l'une des cases suivantes :
(A – très bien, B- Bien, C- Assez bien, D- Passable, E- Médiocre)

	A	B	C	D	E
01- Originalité du travail ……..……..	☐	☐	☐	☐	☐
02- Pertinence par rapport à la connaissance scientifique dans le domaine ……..……	☐	☐	☐	☐	☐
03- Fondements théoriques et hypothèses…. …….	☐	☐	☐	☐	☐
04 – Méthodologie : (description adéquate et appropriée des matériaux et des méthodes)……..	☐	☐	☐	☐	☐
05-Adéquation hypothèses / résultats ……..……..…..	☐	☐	☐	☐	☐
06- Qualité des références bibliographiques…..…… (75% des références doivent dater de la dernière décennie)	☐	☐	☐	☐	☐
07- Qualité et clarté des tableaux et schémas…..……..	☐	☐	☐	☐	☐
08- Maîtrise de la langue utilisée……	☐	☐	☐	☐	☐

09- Style…… ☐ ☐ ☐ ☐ ☐

Avis du référé :
(cocher l'une des cases suivantes)

- Accepté dans la forme présentée ……..…….. ☐
- Accepté après révision mineure …… ☐
- Accepté après révision majeure ……. ☐
(Assez long. Le manuscrit doit être condensé davantage).
- Rejeté …… ☐

Signature du référé **Date** : ……

Commentaires destinés aux auteurs **N° de Code** :
Date :

Outre des commentaires sur l'intérêt scientifique et l'originalité de l'article, vous pouvez exprimer vos recommandations sur la présentation, le style employé, l'organisation de l'étude, donner des indications pour réduire éventuellement l'article, l'adéquation et la disponibilité des références bibliographiques,… Préciser les éventuels points qui doivent faire l'objet de modification ou d'amélioration pour que le thème soit publiable.
(Utilisez éventuellement une feuille supplémentaire) Attention ne pas signer ce document.

……
……
……
……
……
…..……
……
……
……
……
……
……
……
……
……
……
……
……
……
……

Commentaire destinés au Rédacteur en Chef

N° de code : …………………..

(Facultatifs)

Ces commentaires doivent être considérés comme confidentiels et ne peuvent être communiqués aux auteurs. (Utilisés éventuellement une feuille supplémentaire)

………………………………………………………………………………
………………………………………………………………………………
………………………………………………………………………………
………………………………………………………………………………
………………………………………………………………………………
………………………………………………………………………………
………………………………………………………………………………
………………………………………………………………………………
………………………………………………………………………………
………………………………………………………………………………
………………………………………………………………………………
………………………………………………………………………………
………………………………………………………………………………
………………………………………………………………………………
………………………………………………………………………………
………………………………………………………………………………
………………………………………………………………………………
………………………………………………………………………………
………………………………………………………………………………
………………………………………………………………………………
………………………………………………………………………………
………………………………………………………………………………
………………………………………………………………………………
………………………………………………………………………………
………………………………………………………………………………
………………………………………………………………………………
………………………………………………………………………………
………………………………………………………………………………

Signature eu référé **Date** : …………………..