IAEA FAO UNESCO UNIDO

INTERNATIONAL SYMPOSIUM ON THE FUTURE OF SCIENTIFIC, TECHNOLOGICAL AND INDUSTRIAL INFORMATION SERVICES

Leningrad, USSR, 28-31 May 1990

IAEA-SM-317/ 16

ACCES MULTILINGUE AUX BASES DE DONNEES EN TEXTE INTEGRAL

Christian FLUHR et Khaled RADWAN

Institut National des Sciences et Techniques Nucléaires (CEA)
CEN/SACLAY, 91191 Gif/Yvette
France

ACCES MULTILINGUE AUX BASES DE DONNEES EN TEXTE INTEGRAL

Résumé:

De nombreuses bases de données en texte intégral sont disponibles dans une seule langue, ou encore ces bases peuvent contenir des documents dans différentes langues.

Même si l'utilisateur est capable de comprendre la langue des documents de la base de données, il lui est souvent plus aisé d'exprimer son besoin dans sa propre langue. Dans le cas de bases contenant des documents dans des langues différentes, il est plus simple d'exprimer la question dans une seule langue et de pouvoir retrouver des documents dans les différentes langues.

Cet article décrit les développements et les premières expérimentations d'interrogation multilingue, appliqués au couple français-anglais, réalisés sur des données textuelles dans le domaine nucléaire en s'appuyant sur le système SPIRIT.

Après avoir rappelé les problèmes généraux de l'interrogation de bases de données en texte intégral au moyen de questions en langage naturel, nous présenterons les méthodes de reformulation utilisées et nous montrerons comment elles peuvent s'étendre à une interrogation multilingue.

Enfin, les premiers résultats obtenus sont présentés sur des données dans le domaine nucléaire (normes AFCEN et résumés INIS).

1 Définition du problème de l'accès à de l'information dans des bases en texte intégral:

La plus grande part des informations que nous manipulons pour décider des actions que nous devons réaliser sont prises dans des données non structurées, textes ou images. Il peut s'agir de publications scientifiques mais aussi de documentations techniques d'un matériel ou d'un logiciel, d'informations de presse, d'informations juridiques ou tout simplement de courriers ou de messagerie électronique.

La disponibilité du texte intégral sous forme traitable est un facteur important qui autorise une indexation automatique des textes et donc une localisation assez précise des informations. Si l'on ne dispose que de l'image digitalisée du document, il faut en extraire le contenu textuel par des systèmes de reconnaissance optique de caractères.

Nous nous intéresserons dans ce qui suit essentiellement aux méthodes qui permettent une localisation précise d'information pour résoudre un problème précis. Cette localisation est suffisamment difficile à réaliser à cause de la diversité d'expression des mêmes choses dans la langue, pour que le résultat ne puisse être considéré comme sûr. Il peut, comme dans les systèmes documentaires plus classiques, y avoir des documents pertinents qui échappent aux efforts de recherche et aussi des documents non pertinents que le système propose de manière impropre. Le contrôle de l'utilisateur sur le fonctionnement du système reste indispensable.

Pour réaliser cette mise en correspondance du besoin exprimé par l'utilisateur et des textes susceptibles de lui fournir des éléments de réponse, on supposera que le besoin de l'utilisateur est exprimé sous forme d'un texte en langage libre (langage naturel). Le problème de la recherche d'information peut dans ce cas se formuler comme un calcul de proximité sémantique entre deux textes.

La proximité sémantique calculée entre texte décrivant le besoin et textes contenus dans la base permet de classer ceux-ci par ordre de pertinence décroissante, elle permet aussi de chaîner, dans un ordre décroissant d'intérêt, les pages pertinentes, dans le cas de documents longs.

De plus, cette proximité sémantique permet aussi de rapprocher deux parties d'un même texte ou de textes différents. L'ensemble des ces chaînages représentent des liens dynamiques d'hypertextes à partir de besoins exprimés sous forme de textes.

Le calcul de cette proximité sémantique nécessite plusieurs outils: un traitement linguistique qui identifie et normalise les unités de représentation sémantique (mots et expressions), un outil de pondération le plus souvent statistique qui permet d'évaluer la pertinence des réponses, un outil de reformulation qui facilite le rapprochement des mêmes idées exprimées avec des mots différents.

2 Rôle du traitement linguistique:

Le traitement linguistique présente un intérêt plus ou moins important suivant les langues et la puissance des mécanismes de comparaison que l'on compte mettre en oeuvre. Cela explique que très longtemps les systèmes travaillant par exemple sur l'anglais aient comporté très peu de traitement linguistique car des palliatifs simples donnent des résultats souvent satisfaisants. On peut citer par exemple le problème de l'assimilation au même concept de chaînes de caractères représentant des mots dérivés d'une même racine. Ce problème se traite assez bien en anglais qui est une langue qui possède peu de dérivation, cela se passe plus mal en français ou en allemand où les dérivations sont plus nombreuses, et cela ne peut se faire simplement dans des langues où les dérivations affectent la forme de la racine comme en arabe.

Nous allons donc faire un bref panorama des problèmes que le traitement linguistique automatique peut contribuer à résoudre. Il faut noter aussi que si l'on veut appliquer un traitement statistique sur les textes pour déterminer une pondération des mots, il est indispensable d'identifier clairement les événements qui vont être comptés, car les chaînes de caractères ne peuvent sans transformation remplir ce rôle. C'est donc le traitement linguistique qui est chargé de déterminer les événements linguistique à compter.

Pour cela un certain nombre de problèmes sont à résoudre:

2.1 Détermination des synonymies:

Il s'agit surtout de reconnaître les différentes chaînes de caractères sous lesquelles une même notion peut être écrite.

ex:

- mots mal orthographiés: nulcear --> nuclear

mots à orthographe peu respectée:
(non-flammable ou nonflammable ou non flammable)

- sigles sous différentes formes: (INIS, I.N.I.S, I.N.I.S., International Nuclear Information System)

- dérivations d'un même mots (conjugaison, féminin, pluriel, cas): (mouse, mice), (load, loaded, loads)

- dérivations avec changement de catégorie grammaticale:

(load, loader), (manual, manually)

- les terminologies doubles (anglais-américain, français-anglais ou deux orthographes autorisées pour un même mot (organization, organisation), (logiciel, software), (clé, clef).

2.2 Détermination des homographes:

Il s'agit dans la mesure du possible de distinguer les différents sens que peut représenter une même chaîne de caractères.

Ce problème n'est pas soluble dans toute sa généralité avec la technique dont nous disposons. Certaines de ces ambiguïtés seront traitées dans la mesure où elles peuvent être prises en compte par des outils morphologiques ou syntaxiques et non sémantiques.

Toutefois, nous verrons que le mécanisme de comparaison question document est un puissant outil de désambiguation sémantique et nous le montrerons en particulier pour la résolution des ambiguïtés de traduction.

Nature des outils permettant de résoudre des homographies:

- Prise en compte de la typographie complète de la langue avec les accents et les majuscules:
Exemple en français (marche, marché)

- Reconnaissance d'expressions idiomatiques:
La prise en compte de "hot dog" comme un mot unique évite la confusion avec "dog" par ailleurs.

- Désambiguation grammaticale:
Par exemple en français: soude (verbe) et soude (substantif)

2.3 Reconnaissances des mots composés et plus généralement des relations de dépendance:

Pour une recherche d'information documentaire, il est important de distinguer l'occurrence d'un mot composé et la co-occurrence sans lien linguistique des mots le composant. C'est la raison pour laquelle les systèmes booléens ont introduit la notion d'adjacence pour traiter le texte intégral.

C'est l'analyse syntaxique des textes comme des requêtes qui va permettre de déterminer les mots qui sont liés par des relations de dépendance et la nature de ces relations. Cela est particulièrement important pour les liens internes au groupe nominal.

2.4 Normalisation (lemmatisation):

Pour permettre un recherche plus rapide et un comptage plus simple, les mots et mots composés sont représentés sous une forme unique quelle que soit la forme d'origine dans le texte ou la question.

ex:

en français:

commissions (substantif) ----> commission

commissions (verbe subjonctif) -----> commettre

A ce moment, on peut éliminer les mots outils en s'appuyant sur le résultat de l'analyse syntaxique, ce qui permet de traiter correctement les mots qui peuvent être ambigus entre mots outil et mots informationnels.

ex:
en anglais:
can: auxiliaire peut être éliminé
can: verbe transitif ou substantif doit être gardé.

en français:
or: conjonction doit être éliminé
or: substantif doit être gardé.

3 Rôle du traitement de pondération:

Dans une interrogation en langage naturel de bases de données en texte intégral, les questions peuvent être des textes longs. Cela présente un avantage certain par rapport à une interrogation booléenne où on est obligé de ne pas trop préciser ce que l'on cherche sous peine de n'obtenir aucun document.

La démarche d'interrogation en langage naturel est, à l'inverse, de donner le plus de détails possible sur ce qui est recherché. En fait une seule question en langage libre correspond à toute une stratégie de recherche qui peut être très longue et comprendre de très nombreuses questions booléennes.

La conséquence de la longueur possible des questions est qu'il y a peu de chance qu'un même document possède la totalité des mots de celle-ci. Il va donc falloir comparer des intersections question-document comportant des mots différents.

Cela est rendu possible par l'attribution d'une pondération pour chaque terme de l'intersection qui mesure le degré d'intérêt que représente ce terme dans cette intersection. On peut donc calculer à partir du poids des mots, un poids associé à chaque intersection et donc les comparer entre elles selon leur degré de pertinence par rapport au problème posé.

Ces pondérations sont le plus souvent calculées par un modèle statistique. Toutefois nous verrons que le mécanisme de reformulation doit pouvoir les modifier sur des considérations de nature linguistique.

Différents modèles de calcul peuvent être utilisés, il est toutefois important que celui-ci autorise une gestion de la base (ajout, modification, suppression de documents) sans avoir à recalculer les pondérations sur toute la base.

4 La reformulation:

La normalisation apportée par le traitement linguistique et la redondance due au texte intégral et à la possibilité de questions longues permettent dans beaucoup de cas d'effectuer le rapprochement entre questions et documents.

Toutefois il reste des cas où la manière dont les notions sont exprimées dans la question et le document sont tellement différentes que le rapprochement n'est pas possible. Cela est particulièrement important dans le cas où une réponse exhaustive est nécessaire.

La reformulation consiste à trouver toutes les formulations équivalentes de la question pour permettre le rapprochement question-document. C'est une tâche difficile que l'on simplifie en essayant de trouver toutes les formulations équivalentes de chacune des notions contenues dans la question.

Cette simplification peut amener à générer des formes qui ne sont pas cohérentes avec le sens de la question générale. Cela peut donc provoquer du bruit.

Une reformulation ne peut être une simple expansion de la question d'origine par ajout de mots équivalents sous peine de noyer les documents pertinents au milieu des documents non pertinents.

On essaiera de s'appuyer sur les pondérations pour distinguer les documents pertinents des autres. C'est donc sur les stratégies de pondération des mots inférés que s'appuiera cette tentative de séparation des documents pertinents et non pertinents.

Les règles de reformulation peuvent être de la forme suivante:

condition de déclenchement (environnement vocabulaire, catégorie grammaticale, type de relation: synonymie, terme générique, terme spécifique, terme associé)

partie gauche: mot ou mot composé normalisé à reformuler.

partie droite: mot ou mot composé normalisé reformulé.

Ces règles de reformulation peuvent être construites automatiquement par un traitement linguistique (stemmatisation) ou reprises d'un thesaurus existant, ou construites pour les besoins de la reformulation.

Exemple de règle de type stemmatisation:

(synonymie) indexer ------> indexation
(terme associé) indexer -----> index

exemple de règles de type thesaurus

(terme spécifique) réacteur nucléaire -----> R.E.P.
nuclear plant -----> P.W.R.

5 Mécanisme de comparaison question document:

Dans ce qui suit nous puiserons nos exemples dans l'utilisation du système SPIRIT.

Le processus de comparaison texte requête-texte de la base, est réalisé dans SPIRIT en cinq étapes.

5.1 Expansion de la question avec métarègles:

Cette étape a pour but de produire à partir des mots et mots composés de la question d'origine, tous les mots qui peuvent être inférés avec différents types de relation sémantiques.

Il n'est pas toujours bon de produire les inférences maximales et certaines règles peuvent ne pas être déclenchées en fonction de règles stratégiques basées sur le contexte d'utilisation.

Par exemple, on déclenchera des reformulations de termes génériques pour des applications où la question est beaucoup plus précise que le texte des documents. Cela peut être le cas dans des applications où une question précise est comparée à des descriptions de contenu de bases de données en ligne pour déterminer à quelle base la question doit être adressée.

A cette étape, pour chaque mot inféré, on aura le mot qui l'a produit et le type de relation utilisée. Cela est indispensable pour établir dans l'étape d'évaluation, la pondération associée aux mots inférés.

5.2 Filtrage par rapport à la base et pondération primaire:

Cette étape a pour but de vérifier que les mots de la question et les mots inférés sont présents dans la base et de déterminer leur poids ainsi que la localisation de leur liste inversée.

5.3 Etablissement d'une stratégie de recherche optimisée:

Cette étape a pour but de trier les termes en fonction de leur poids afin d'élaborer une stratégie optimisée de détermination des documents les plus pertinents. En effet, en commençant la recherche des documents par les mots les plus discriminents, on obtient tout de suite au début les documents qui ont le plus de chance d'être pertinents, ce qui permet d'arrêter la comparaison après un quantum de temps maximum ou un ensemble de réponse maximum sans courir le risque de perdre des documents intéressants.

5.4 Identification des intersections:

Cette étape est l'exécution de la stratégie élaborée à l'étape précédente et a pour but d'obtenir pour les documents les plus pertinents, une description de leur intersection avec les termes de la question d'origine et les termes inférés.

5.5 Pondération des intersections et regroupement par classes:

S'appuyant sur des règles externes, ce mécanisme permet d'évaluer le poids, en particulier des mots inférés qui ne peuvent avoir le même poids que s'ils étaient directement dans la question. De plus, le calcul du poids des mots inférés dépend de la nature de la relation sémantique qui les a inférés.

On pourra trouver une description plus détaillée de ce processus dans [2].

D'autres règles permettent de ne pas prendre en compte deux mots isolés si le mot composé correspondant existe dans le document.

Le système regroupe ensuite les documents par classes caractérisées par la nature des concepts exprimés dans la question qui sont présent dans le document (soit par l'intermédiaire des mots de la question soit par l'intermédiaire de mots inférés).

Remarque:

Ce mécanisme de calcul d'une proximité sémantique peut être utilisé pour trier par ordre décroissant de pertinence les parties de document lors de la visualisation. Cela permet de chaîner les pages informationnelles dans des documents qui peuvent être très longs.

6 Le problème de la reformulation multilingue:

Le problème que nous voulons résoudre est l'interrogation, dans une langue, de données textuelles exprimées dans une autre langue ou dans plusieurs langues.

Il ne s'agit pas pour nous de faire une traduction de la question mais, une fois l'ensemble des termes significatifs déterminés et normalisés par le traitement linguistique, de faire une étape de reformulation supplémentaire avec un nouveau type de relation sémantique, c'est à dire utiliser des règles de type traduction.

Nous nous plaçons résolument dans le cas général car il n'est pas raisonnable d'envisager de construire de règles de traduction pour chaque base de données ou même chaque domaine d'application. On trouvera dans le paragraphe suivant des indications sur les méthodes de construction de ces règles de traduction.

De plus certaines métarègles de nature syntaxiques doivent être utilisées:

Par exemple, pour l'interrogation de français vers l'anglais, on utilisera la règle suivante pour les mots composés:

mot1 (substantif) mot2 (adjectif) --> mot2 (adjectif) mot1 (Substantif)

avant d'appliquer les règles de type traduction sur mot1 et mot2 individuellement. Cela est bien entendu le cas s'il n'existe pas de règle de traduction globale de mot1-mot2:

exemple:

indexation automatique ---> automatique indexation

qui après application de règles de traduction mot à mot donnera : automatic indexing.

Le processus de traduction n'est pas univoque, en effet un même mot peut se traduire de différente manière et en particulier il peut produire différents synonymes, mais aussi, il peut y avoir, à cause de la polysémie du mot de la langue source, plusieurs traductions sémantiquement incompatibles.

Nous verrons dans les premiers exemples d'interrogation multilingue que le processus de comparaison texte question - texte de la base est un puissant outil de sélection des bonnes traductions car les documents avec de mauvaises traductions se retrouvent dans des classes beaucoup moins pertinentes que les documents possédant le mot avec sa bonne traduction.

7 Constitution des règles de reformulation multilingues:

Deux méthodes ont été utilisées:

Tout d'abord pour obtenir un ensemble de règles de base en particulier portant sur les mots simples, il est possible de filtrer des dictionnaires existants pour en extraire des règles du type:

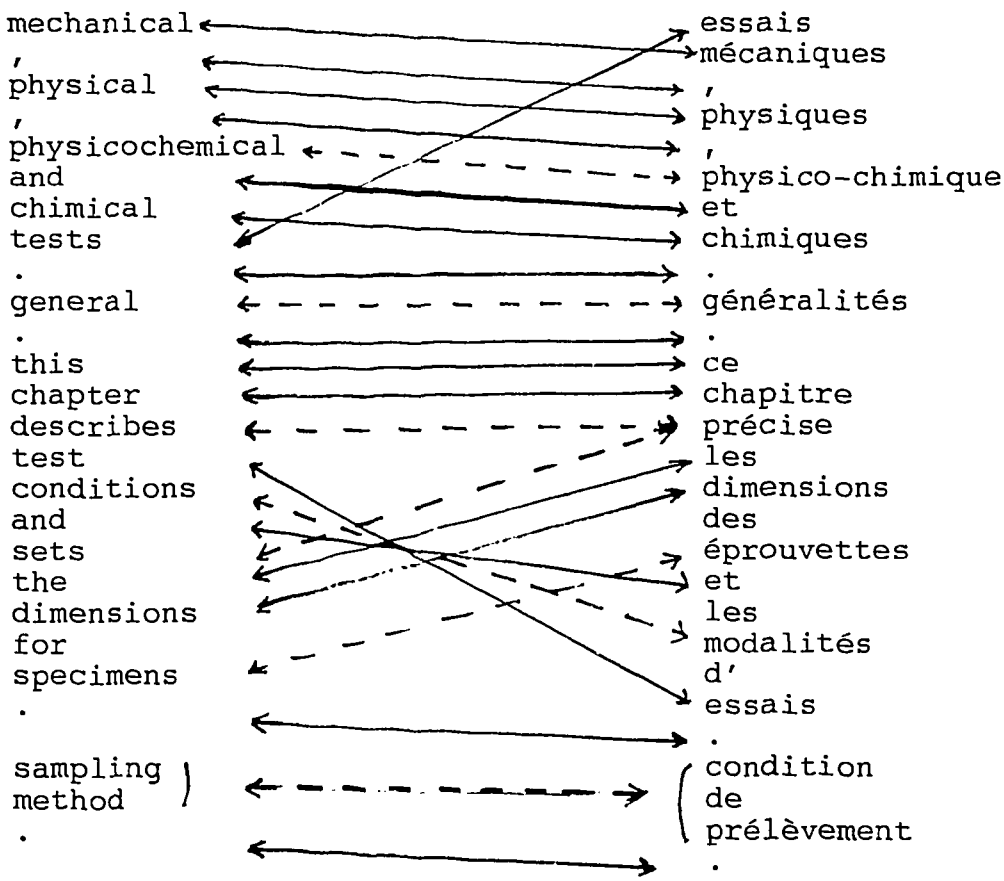
mot langue source (condition éventuelle sur la catégorie) --> mot langue cible.

Cela ne permet pas en général de traduire la terminologie spécialisée du domaine qui est souvent exprimée par des mots composés.

Nous sommes donc en train de mettre au point une méthodologie de construction automatique de règles de traduction d'expressions si l'on dispose déjà dans le domaine de textes traduits.

Pour cela, on analyse les mêmes textes dans les deux langues. Un programme de mise en correspondance permet, à l'aide du dictionnaire général d'établir la correspondance pour une grande partie des mots du texte. Les mots simples qui n'ont pas de traduction ou les mots composés qui ne sont pas traduits mot à mots sont repérés et par des règles syntaxiques assez simples, sont mis eux aussi en correspondances. Ces correspondances permettent de créer de nouvelles règles de traduction propre au domaine de la base.

Ce processus a été expérimenté sur quelques textes de normes de construction de centrales édictées par l'AFCEN. On trouvera dans ce qui suit un exemple de mise en correspondance. Dans l'exemple qui suit, les mises en correspondance automatique à partir du dictionnaire général sont en trait plein, les relations nouvelles inférées sont en pointillé (travail réalisé par S. Elyès).



Nous prévoyons de renouveler cette expérience sur les résumés INIS pour lesquels nous disposons d'une version en français et en anglais. Le résultat de ce travail servira à réaliser une interrogation bilingue de la base INIS. Il devrait aussi permettre une amélioration des lexiques du système de traduction automatique SYSTRAN que nous expérimentons pour la traduction en anglais des résumés destinés à la base INIS.

8 Expérimentation:

La première expérimentation d'interrogation bilingue a été menée sur un échantillon de 73 documents de la base INIS. Nous avons choisi un ensemble de documents pour lesquels nous disposions de résumés en français et en anglais. Deux bases contenant le même ensemble de documents ont été générées, l'une en anglais l'autre en français. Cela devrait permettre de comparer l'accès monolingue et l'accès bilingue à la même information.

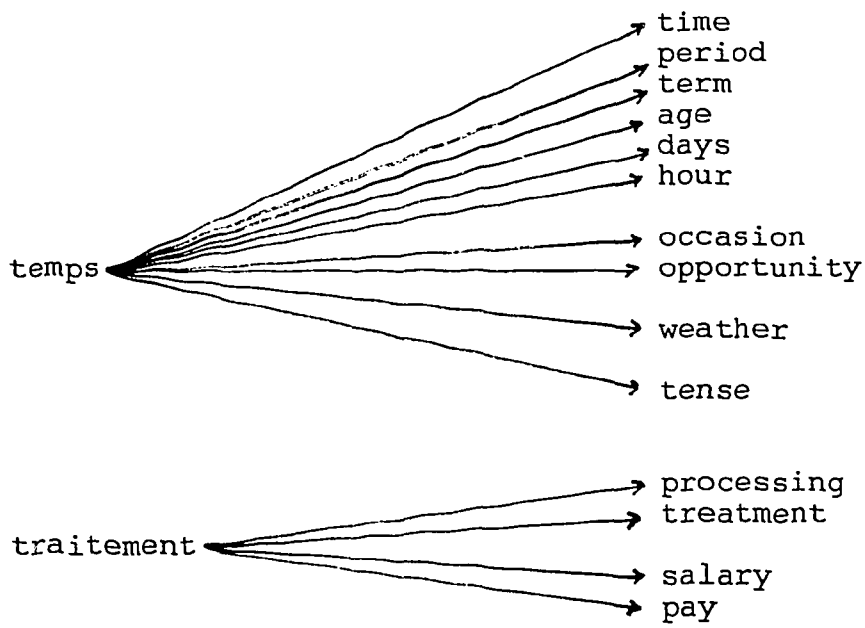
La méthodologie de reformulation est la plus élémentaire possible, aucune stratégie particulière n'a été introduite, les règles de transformation syntaxique n'ont pas encore été implantées, aucun mot ou expression spécifique du domaine n'a été introduit dans les règles de reformulation. Seules les règles issues de dictionnaires généraux ont été employées.

Nous n'avons pas encore pu mener suffisamment d'expérimentations pour faire des mesures significatives de performance comparative.

La première constatation, qui peut sembler surprenante à première vue, est que l'interrogation bilingue est souvent plus performante (moins de silence) que l'interrogation monolingue (sans reformulation).

En fait cela s'explique très bien par la nature des règles de reformulation multilingue qui combinent, par la possibilité de traduire un mot par plusieurs mots différents synomymes, une relation de traduction bilingue et plusieurs relations de reformulation monolingue (synonymies).

exemple: la question en français est : temps de traitement



Le meilleur document possèdait les mots "processing" et "time" dans le même paragraphe mais pas en tant que mot composé.

La deuxième constatation que l'on peut faire est que le système de comparaison pondéré, combiné à la base de textes, se révèle un puissant outil de levée d'ambiguïtés de traduction.

En effet, si un mot donné peut en dehors de tout contexte avoir de nombreuses traductions dont certaines sont incompatibles sémantiquement (d'autres le sont seulement par l'usage), dans le contexte du texte question, on peut dire que la bonne traduction sera celle que l'on trouve dans le document jugé par le système le plus pertinent, c'est-à-dire le plus proche sémantiquement du texte-question.

Ces premières constatations devront être confirmées par une évaluation statistique de performances sur un grand nombre d'essais.

9 Conclusion:

Cette expérimentation n'en est qu'à ses débuts, de nombreuses améliorations peuvent être apportées, ajout des expressions du domaine, utilisation des règles syntaxiques, meilleur ajustement de la pondération des termes traduits.

De plus, il faut étudier la combinaison de la reformulation monolingue et multilingue pour diminuer le silence en essayant de ne pas produire trop de bruit en particulier parmi les documents jugés les plus pertinents.

Il est possible de faire une reformulation dans la langue source, dans la langue cible, ou dans les deux. Il faut comparer les performances et les coûts pour pouvoir prendre les bonnes décisions dans ce domaine.

La possibilité d'introduire des métarègles (stratégie) doit permettre d'avoir un comportement vis-à-vis de ce problème qui peut s'adapter au cas particulier.

On peut dire pour conclure que les premiers résultats sont très encourageants mais qu'ils devront être confirmés par des mesures systématiques.

10 Bibliographie:

[1] ANDREEWSKY A., BINQUET J.P., DEBILI F., FLUHR C.,POUDEROUX B., Linguistic and statistical processing of texts and its application in the field of legal documentation, 6th symposium on Legal Data processing in Europe (Council of Europe), Thessaloniki, july 1981.

[2] DEBILI F., FLUHR C., RADASOA P., About reformulation in full-text IRS, Information Processing and Management Vol. 25, N° 6 1989, pp 647-657.

[3] FLUHR C., Natural language processing for documentary information retrieval, to be published in 1990, in a special issue of Knowledge Engineering on natural language processing applications .

[4] RAY K., DRISCOLL J., New directions for microcomputer based hypertext systems, Database Magazine, july 1990.