INFORMETRICS 87/88
L. Egghe and R. Rousseau (Editors)

REFORMULATION DES QUESTIONS DANS L'INTERROGATION DES BASES DES DONNEES BIBLIOGRAPHIQUES : SYSTEME LEXIQUEST

J. POMIAN

SERPIA-CDST-CNRS, 26 rue Boyer, 75020 Paris, France

Abstract

The purpose of the LEXIQUEST project is to introduce the use of artificial intelligence techniques in information retrieval. LEXIQUEST handles relations between indexing key-words. It helps the user to find out a meaningful formulation of his query, which is stated in free terms.
Most database interfaces provide the user with catalographic tools (lists of titles, authors, key words, and so on), but are unable to detect the user's preoccupation. Some systems, for example DIALECT or EXPRIM, use an interpretation of the user's request to reformulate his query. This is also the purpose of LEXIQUEST, but our system differs in that the procedures and the rules are independent from the field on which the program is run, since they are directly generated by specialised programs. It is therefore possible to run the LEXIQUEST system on any key-word indexed database, without modifying the system.
LEXIQUEST is designed to guide any query made on a key-word indexed database. The system interactively refines and reformulates the initial query, using the relations it finds between key-words. After some iterations, the final formulation is validated by the user, and the relevant documents are retrieved.
The links between key-words are weighted by the "E" coefficient, in order to detect each word's semantic environment. The value of this "E" coefficient is :

$$E = \frac{C_{ij}^2}{C_i \cdot C_j}$$

This "Equivalence" coefficient measures the mutual inclusion between two words i and j, by using their frequencies C_i and C_j, and their co-occurrence C_{ij}.
The system makes an intensive use of the transitivity of the "E" coefficient. Links weighted with this coefficient are interpreted as rules, and are recorded as such in the set of rules used by the system. The LEXIQUEST system furthermore uses a saturation metarule in order to limit the size of the network attached to each word.
The LEXIQUEST system is interesting in two points. First of all, there is no use in creating manually the association network for each database, since it is created automatically. The system can therefore be adapted to any database. Second, there is no need to know the indexing vocabulary in advance, since LEXIQUEST can handle the reformulation of the query.

1. INTRODUCTION

Le recouvrement de l'information dans les bases de données et les bases documentaires prend une importance croissante tant dans les recherches en sciences de l'information qu'en intelligence artificielle. En effet, la récupération de l'information pertinente au maximum dépourvue de bruit, devient un enjeu crucial compte tenu de la multiplication et de la diversification des sources documentaires. C'est dans le but de faciliter la recherce des références bibliographiques dans une base, indexée par mots-clefs que nous avons mis au point le système Lexiquest.
Lexiquest a pour objectif la mise en place des techniques d'intelligence artificielle dans le cadre des recherches documentaires et plus precisément le traitement de l'interrogation de toute base indexée par mots-clefs. Lexiquest travaille à partir des relations entre les mots indexant les documents qui composent la base bibliographique. Il aide l'utilisateur à préciser le sens de sa question, en cherchant le thème central de celle-ci et en tenant compte de son contexte.
La plupart des systèmes actuels offrent soit des fonctions de recherche du type catalographique (titre, nom d'auteur, etc...), soit proposent directement des mots spécifiques, sans interpréter la question, ce que nous faisons précisément. Il existe toutefois des systèmes, tels Dialect [1], Exprim [2], Intellect [3], ou d'autres [4] qui aident l'utilisateur à reformuler sa requête en l'interprétant.
Les fonctionnalités de Lexiquest et des systèmes précités sont semblables, mais, à différence d'autres systèmes, toutes les règles et procédures de Lexiquest sont indépendantes du domaine d'application et générées directement par les programmes eux-mêmes. Ainsi, il est possible d'exploiter Lexiquest sur n'importe quelle base de données indexée par mots-clefs, sans ajouts ni modifications de traitements.
Après avoir décrit brièvement le fonctionnement de Lexiquest (§ 2), nous détaillerons l'analyse et la normalisation des requêtes (§ 3) et la procédure de reformulation inférentielle (§ 4).

2. LE SYSTEME LEXIQUEST

2.1 Lexiquest et les autres systèmes documentaires

Il existe de nombreux systèmes d'aide à l'interrogation, dont certains font également de la reformulation des requêtes. On peut citer deux systèmes français, tels que Dialect et Exprim, et des systèmes américains tels que Intellect, Xcalibur, Orbi ou Irus [5].
Ces systèmes contiennent un interface étendu en langage naturel qui cherche notamment à prendre en compte l'usage des pronoms et des ellipses, ce qui permet à l'utilisateur de se référer au contenu de la question posée précédemment alors que Lexiquest ne fait pas de traitements linguistiques (cf. § 3 infra). De plus, le traitement de la phrase dans Lexiquest ne cherche pas à mettre en évidence des prédicats, mais à reconnaître les mots-clefs relevant du domaine d'application.
Les systèmes cités ci-dessus ont pour objectif premier de faciliter l'accès aux documents, et non de faire de la reformulation. Celle-ci ou l'accès direct, sont réalisés le plus souvent au moyen des règles de production contextuelles étendues, utilisant des coefficients de vraisemblance (Dialect, Satin, système de Kellog et Travis). Les règles sont construites en fonction du domaine. La représentation prédicative de la phrase doit alors être transformée en un "pattern" utilisé pour la recherche des règles pouvant être appliquées.
L'autre moyen employé également, en complément du précédent c'est le réseau sémantique (Exprim), construit en fonction du thesaurus défini par les documentalistes. Ce deuxième point différencie également Lexiquest des autres systèmes documentaires, car les règles utilisées dans Lexiquest ne correspondent pas à une expertise particulière et elles sont générées automatiquement à partir de l'indexation existante.

Lexiquest se différencie donc fondamentalement des systèmes actuellement existants tant par la formalisation des connaissances que par leur utilisation pour la reformulation des requêtes.

2.2 La procédure générale

A partir de la requête initiale et en s'appuyant sur les relations entre les mots de l'indexation, le système procède à l'affinement et au remodelage progressif de la question. Le processus consiste en une succession d'étapes qui aboutissent à la requête finale validée par l'utilisateur et, en fin de compte, à la recherche des documents demandés.

Principales étapes de la procédure :

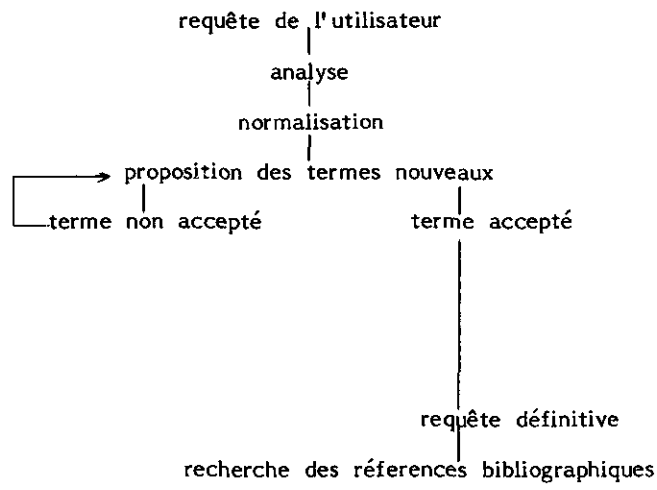


Fig. 1

2.3 Les mots associés

Pour effectuer les recherches et les inférences Lexiquest utilise systématiquement les associations entre les mots de la base indexée construites par le programme Leximappe [6] conçu initialement pour étudier à partir des publications recensées dans les bases bibliographiques la politique de la recherche dans le domaine scientifique et technique. Ce programme construit automatiquement le thesaurus correspondant à une base donnée. Il met en évidence les relations entre les mots utilisés au cours de l'indexation.
Dans le thesaurus Leximappe, chaque mot est décrit par d'autres mots, déterminés en fonction de leur co-occurrence avec le mot d'en-tête. A titre d'exemple on peut obtenir (extrait d'un fichier relatif à l'intelligence artificielle qui nous servira d'exemple tout au long du présent article (*)) :

(*) ce fichier contient 800 notices bibliographiques extraites de la base Pascal, qui regroupe environ 4 millions de références.

14 RECONNAISSANCE PAROLE (FREQUENCE : 34)

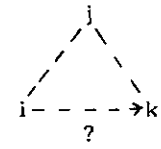
Mots associés :

indice	n° mot	mots associés
0.15	60	RECONNAISSANCE MOT
0.01	40	RECONNAISSANCE AUTOMATIQUE
0.08	71	DIALOGUE HOMME_MACHINE
0.07	61	SYNTHESE PAROLE
0.05	25	COMPREHENSION LANGAGE

Les mots associés sont obtenus avec l'indice d'équivalence, appelé "indice E". Il donne pour chaque mot son environnement lexical qui peut être interprété de façon sémantique. En effet, l'indice E obtenu par la formule suivante :

$$E = \frac{C_{ij} * C_{ij}}{C_i * C_j}$$

mesure l'implication mutuelle ou l'inclusion réciproque de deux mots i et j, en utilisant les co-occurrences C_{ij} entre les mots i et j, et leur fréquences d'apparition C_i et C_j.
La propriété fondamentale du coefficient E est sa transitivité. Il est possible de l'interpréter en disant que "si l'on s'intéresse au mot A, il est probable que l'on s'intéresse également au mot B avec la probabilité e", c'est-à-dire que dans l'exemple donné ci-dessus, il est possible de traduire la relation entre "reconnaissance parole" et "reconnaissance mot" par la règle suivante : "si on s'intéresse au mot "reconnaissance parole" on s'intéressera au mot "reconnaissance mot" avec la probabilité de 0.15". Puisqu'il s'agit d'un coefficient d'inclusion réciproque la règle précédente peut également être exprimée de façon suivante "si on s'intéresse au mot 'reconnaissance mot' on s'intéressera au mot 'reconnaissance parole' avec la probabilité de 0.15".
Les relations entre les mots construites avec le coefficient E sont transcrites sous forme de règles qui, une fois regroupées, constituent une base de règles.
La base de règles permet de s'intéresser à la chaîne des mots que "ramène" un mot donné. Cette chaîne des mots est en réalité le chemin d'un réseau, chemin ayant pour point de départ un mot. Le réseau correspond dans notre cas, à l'ensemble des mots du thesaurus, ou encore à toutes les règles de la base. La longueur d'une chaîne susceptible d'être ramenée par un mot est finie grâce à une métarègle de saturation qui freine la transitivité entre les mots; en effet, si nous avons le mot i lié au mot j et le mot j lié au mot k, i est lié à k à conditon que : $C_{ik} * C_{ij} \leq C_{ik} * C_i$.



La liaison entre le mot i et le mot k dépend donc de la validation de la métarègle de transitivité.
Graphiquement on peut représenter de façon suivante un extrait du réseau de mots associés :

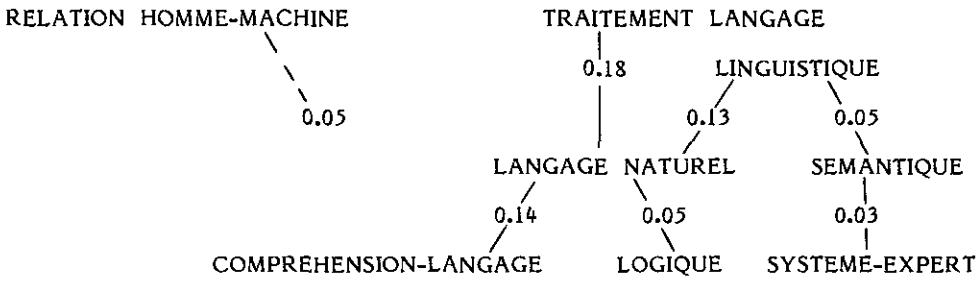


Fig. 2

Le réseau présenté ci-dessus offre une image incomplète des connexions existantes entre les différents noeuds. Ce n'est qu'en appliquant la métarègle de transitivité que l'on met en évidence toutes les liaisons présentes effectivement.
Ainsi toute évaluation du contexte effectuée par Lexiquest est-elle probabiliste. Il s'agit dans ce cas d'un recouvrement de l'information qui ne peut pas donner de solution unique, mais qui propose différentes solutions avec des coefficients de probabilité différents. Le choix définitif est laissé à l'utilisateur qui pose sa question et choisit le (ou les) terme(s) qui la complètent (cf. § 4).

3. ANALYSE ET NORMALISATION DES REQUETES

La question posée au système est formulée en langage libre c'est-à-dire sans qu'il soit nécessaire de consulter au préalable des listes de vocabulaire d'indexation. De même aucune contrainte ne pèse sur la façon de s'exprimer de l'utilisateur. Avant de pouvoir interpréter la requête (§ 4) il faut l'analyser (§ 3.1) et surtout normaliser les mots-clefs qu'elle contient (§ 3.2 et § 3.3).

3.2 Normalisation : règles générales

La normalisation consiste à parcourir la question et à y repérer tous les mots qui sont susceptibles d'être mots-clefs du lexique d'indexation.
Pour procéder à la normalisation on dispose de :
- l'ensemble des mots du lexique d'indexation de la base d'articles,
- quelques règles générales relatives à la normalisation des mots-clefs dans Pascal (*), et des règles créées spécifiquement à cet effet.

Les règles générales de normalisation :
- aucun mot isolé n'est au pluriel,
- en général les mots composés sont au singulier, sauf cas particuliers,
- pas de prépositions,
- pas d'adjectifs : tous les mots-clefs sont des substantifs,
- pas de sigles,
- pas de mots abrégés ou préfixes seuls,
- certains mots, considérés comme vides de sens ne seront jamais des mots-clefs,
- deux concepts distincts donnent deux mots-clefs distincts.

A partir de ces règles on procède à la normalisation de la question. Cependant le but de la normalisation n'est pas l'obtention des mots-clefs "parfaits"; nous venons d'exclure tout recours à un dictionnaire annexe des mots du vocabulaire scientifique et technique. En effet, le vocabulaire traité dans l'exemple d'application de Lexiquest relève d'un domaine d'application particulier - l'intelligence artificielle -, mais une base bibliographique scientifique et technique contient des dizaines des domaines d'application dont les vocabulaires ne se

(*) l'adaption de la normalisation à une autre base bibliographique ne nécisserait aucune modification de fond du programme.

recoupent que partiellement. Ainsi, dans la base Pascal, trouve-t-on environ 100.000 descripteurs contrôlés pour l'ensemble des disciplines prises en compte, ce qui donne une indication sur la taille et la diversité du vocabulaire utilisé. Il en est de même pour les banques d'images sur vidéodisque, indexées aussi par mots-clefs. Ce serait donc une gageure que d'essayer de constituer des listes d'exceptions, qu'il faudrait modifier lors de chaque nouvelle application. Notre but est donc de construire un outil robuste, résistant au bruit et aux mots inconnus.

Certaines règles de normalisation spécifiques portent sur des mots simples, d'autres sur des mots composés.

Les mots composés le sont lorsque plusieurs termes définissent un concept. "Reconnaissance parole" est donc un mot composé. On distingue des mots vides et des mots incomplets. Les mots vides sont des syntagmes sans signification propre qui ne seraient jamais utilisés comme mots-clefs, comme par exemple "analyse" alors que les mots incomplets sont des mots que l'on trouve prefixés dans le lexique d'indexation, et seuls dans la question.

Les règles de normalisation dans Lexiquest :

- les articles et les prépositions sont enlevés,
- pour des mots simples le pluriel est enlevé systématiquement sauf pour des mots étrangers (ex. "process"). L'absence de recours à une liste d'exceptions a pour conséquence la génération de quelques mots "aberrants" comme "acce", "creu", "discour", etc. Leur nombre reste toutefois très faible, et de plus, l'algorithme du traitement du pluriel a été affiné afin de ne laisser que les termes dont il est impossible de savoir s'ils constituent ou non une exception. Ainsi garde-t-on l'expression "accès lexical" sans la modifier, mais il est impossible d'empêcher le génération d'un mot comme "creu", si le terme "creux" n'existe pas dans le lexique de référence,
- les mots vides sont enlevés : est considéré comme mot vide un mot qui dans la question se trouve seul, alors que dans le lexique il est toujours accompagné d'un terme supplémentaire qui est un suffixe. "Analyse" est considéré comme un mot vide, car on trouve dans le lexique "analyse amas", "analyse donnée", etc..., alors que "image" n'est pas un mot vide, car outre "image numérique" on trouve "analyse image", "traitement image", etc...,
- pour les mots composés on garde le pluriel dans le cas où une forme au pluriel est attestée dans le lexique (ex. "représentation connaissances"), sinon on met au singulier tout le mot composé.
- sur les mots composés on opère un traitement particulier : dans le cas où un mot composé l'est de deux termes dont le premier figure dans le lexique, le second est incomplet, et le mot composé n'existe pas dans le lexique en tant que tel, on éclate le mot composé (ex. "segmentation image" devient "segmentation" et "image" et "image" sera traité plus loin comme terme incomplet),
- si les substantifs du mot composé appartiennent séparément au lexique, le mot est éclaté en autant des mots-clefs différents : "représentation sémantique du discours" est éclaté en "représentation sémantique" d'une part et "discour(s)" de l'autre; "génération du langage naturel" devient "génération langage" et "langage naturel",
- on peut rencontrer dans la question des mots qui correspondent incomplètement aux mots du lexique : ce sont des mots incomplets. Ainsi trouve-t-on "parole" dans une question, alors que dans la base d'indexation il y a "reconnaissance parole". Le traitement conjoint des mots vides et des mots incomplets permet de réduire l'écart entre le vocabulaire de la question et celui de la base, sans éliminer des mots abusivement.

Avant de décider si un mot est vide on regarde s'il ne peut pas être remplacé par un mot du lexique de la base d'articles. On regarde donc si dans le lexique pour le mot "m" on trouve des mots de la forme "x-m". Si oui, alors le mot sera examiné en tant que mot incomplet. Pour le moment nous nous intéressons uniquement aux mots-clefs de la forme "x-m", et non aux mots de la forme "m-x". En effet, il semble que dans les mots-clefs du français, le premier terme est toujours plus général que le second, (ex. "analyse", "reconnaissance", "traitement"), et lorsqu'il est employé isolément dans une question il peut soit

être considéré comme un mot vide (ex. "analyse"), soit exister dans le lexique (ex. "application", "vision"). Il va toutefois de soi, que le traitement exposé ci-dessous peut être appliqué sans difficulté aux mots de la forme "m-x", donc généralisé aux mots-clefs d'autres langues, comme l'anglais, par exemple.
L'analyse des mots incomplets utilise les données fournies par le thesaurus généré par Leximappe. Le cas simple est celui où il n'y a qu'un seul terme de la forme "x-m" correspondant à "m" : il remplace alors "m" dans la question. Lorsque plusieurs termes de la forme "x-m"" ont été trouvés, différents cas de figure peuvent se produire.

3.2 Normalisation : remplacement des mots incomplets

3.2.1 Sélection directe des mots incomplets

Soit la question suivante : "image reconnaissance forme segmentation algorithme statistique" déjà analysée et partiellement normalisée. Le terme "image" est examiné comme mot incomplet. "Statistique" ne figure pas dans le lexique, les mots "reconnaissance forme segmentation algorithme" serviront à choisir entre les différents mots possibles se terminant par "image" et qui sont : "analyse image", "interprétation image", et "traitement image". La base des règles E fournit les mots qui leur sont associés :

analyse image : - interprétation image 0.07
- vision ordinateur 0.05
- segmentation 0.04
- numérisation 0.04
- contour 0.03
- détection bord 0.03

interprétation image : - analyse image 0.07

traitement image : - segmentation 0.08
- image numérique 0.08
- reconnaissance forme 0.07
- algorithme 0.06
- image binaire 0.06
- structure binaire 0.04

L'intersection entre les mots présents dans la question et les différents mots associés n'étant pas vide, on peut calculer le poids de chaque mot susceptible d'être utilisé en additionnant les coefficients des mots présents dans la question et dans la liste des mots associés. On obtient :
traitement image : 0.21, analyse image : 0.04, interpretation image : 0.0
ce qui permet de choisir "traitement image", la règle étant de choisir le terme affecté du plus fort poids en mots associés, mots présents dans la question. Un problème se pose cependant si deux mots sont affectés d'un même poids, ou si tous les poids sont nuls. Dans le cas où deux termes ont un même poids non nul, on insère les deux termes sans faire de choix entre eux. Si tous les termes sélectionnés sont affectés des poids nuls, cela veut dire qu'aucun mot de la question n'est parmi leurs mots associés. Il faut donc approfondir l'analyse des contextes d'utilisation de chacun de termes.

3.2.2 Sélection par analyse en profondeur

Pour déterminer le terme remplaçant on utilise la base des règles E, présentée ci-dessus. La procédure générale consiste à extraire les chaînes contextuelles pour chaque mot et à utiliser les mots composant ces chaînes pour déterminer le terme de remplacement.
La procédure est la suivante : pour chaque "remplaçant" potentiel on génère le réseau qu'il ramène par transition avec ses mots associés. On prend les mots obtenus par transition comme étant eux-mêmes associés au terme "remplaçant", à condition qu'ils vérifient la métarègle de transivité. Les mots ont alors un coefficient multiple de deux valeurs d'association. Si les mots de la question figurent parmi les termes associés au second niveau, alors on détermine le terme remplaçant de la même façon que dans le cas de la sélection directe, en

additionnant les poids des mots associés, sinon on continue, en générant un extrait supplémentaire du réseau. Il est possible, en effet, d'étendre la procédure de génération du réseau jusqu'à un niveau de profondeur quelconque. En général, cependant, une solution est trouvée au bout d'une ou deux itérations. Dans le cas où le parcours continue, on s'arrête lorsque les valeurs des coefficients deviennent négligeables, c'est-à-dire inférieures à 0.001.
Soit la question suivante : "image analyse scene région". Les mots "analyse scene" et "région" ne figurent pas parmi les mots associés à "analyse image" ou à "traitement image". On génère donc le réseau associé à "analyse image " qui ramène les mots suivants :

analyse image	
par segmentation	analyse texture
par segmentation	analyse scene (0.002)
par segmentation	reconnaissance forme
traitement image	
par image numérique	classification
par reconnaissance forme	classification
par reconnaissance forme	classification automatique
par reconnaissance forme	segmentation
par reconnaissance forme	algorithme
par reconnaissance forme	analyse scene (0.021)
par algorithme	plus proche voisin
par algorithme	région (0.024)
par algorithme	reconnaissance forme
par segmentation	reconnaissance forme

On obtient donc :
traitement image : 0.045, analyse image : 0.002
et on remplace dans la question le terme "image" par "traitement image".
L'analyse du contexte d'utilisation possible de "traitement image" a montré que celui-ci pouvait comporter des termes comme "analyse scene" et "region" avec une probabilité plus forte que "analyse image".
Si aucun contexte commun entre les mots "remplaçants" et ceux de la question n'est détecté, le mot incomplet ne sera pas remplacé. Il sera traité ultérieurement, par le programme de traitement des mots inconnus, actuellement en cours de développement.
La normalisation est une étape indispensable pour pouvoir traiter une question. Elle permet de passer d'une phrase comme "je m'intéresse aux règles de productions dans les systèmes de diagnostic" à une question transformée en "regle- production diagnostic", ce qui autorise le démarrage de l'activité de reformulation.

4. LA REFORMULATION

4.1 Proposition des termes nouveaux

Une fois la normalisation achevée on peut procéder aux recherches des termes nouveaux susceptibles de préciser le sens de la requête. Dans la suite de l'article nous partirons, pour plus de clarté, des questions déjà analysées et complètement normalisées. La reformulation, ou l'analyse sémantique de la question utilise également la base des règles E. Comme nous l'avons déjà souligné, les termes proposés le seront en fonction de leur probabilité d'apparition commune dans la base d'indexation.
On commence par générer les mots associés à chaque mot de la question, et calculer les intersections éventuelles entre eux. si une telle intersection existe, alors on a réussi à trouver un ou plusieurs mots à proposer à l'utilisatuer.
Toutes les propositions de termes se font en suivant les valeurs de probabilites décroissantes; le premier terme proposé est donc toujours celui qui a la plus forte valeur d'association avec les mots de la question.

Exemple 1

Question : règle production et diagnostic

Réponse : Votre question porte sur : SYSTEM-EXPERT ?

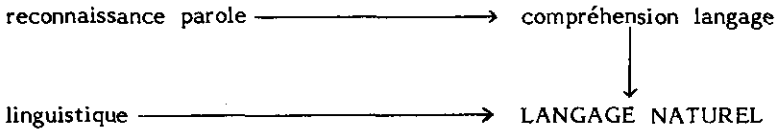
En effet, le mot "systeme expert" figure parmi les mots directement associés à la fois à "règle production" et à "diagnostic".
Il se peut que l'utilisateur ne valide pas immédiatement le terme proposé. La procédure générale consiste alors à parcourir en parallèle les réseaux des termes associés à chacun des mots de la question, et à calculer, après chaque itération, les intersections éventuelles entre tous les mots obtenus. Un mot donné, obtenu par intersection, peut en effet être associé de façon plus forte au mot m1 de la question et moins forte au mot m2. Il se trouvera donc à différents niveaux de l'arborescence de mots.

Exemple 2

Question : reconnaissance parole et linguistique

Réponse : Votre question porte sur : LANGAGE NATUREL

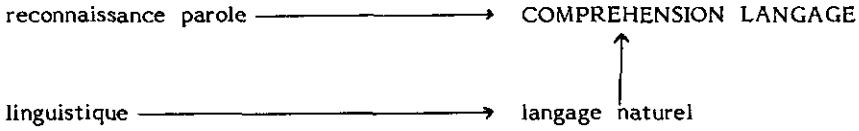
Chaque fois que Lexiquest fournit une réponse en proposant un terme, l'utilisateur peut demander la justification du choix de mot proposé. Il obtient alors de façon dont le terme proposé a été obtenu, à travers la visualisation de l'extrait du réseau parcouru, qui a permis de sélectionner le (ou les) mot(s) proposé(s).
Explication : Nous avons détecté les chaînes contextuelles suivantes :



Ce mini-réseau visualise les relations de transition entre les différents termes. Compte tenu de la transitivite nous savons qu'il y a un lien entre "reconnaissance parole" et "langage naturel", car la métarègle de transitivité a été validée.
Si le premier terme proposé à l'utilisateur n'a pas été approuvé, on en propose d'autres, s'ils existent, ou bien on continue la recherche. Dans cet exemple, si "langage naturel" n'a pas été validé, on propose :

Réponse : Votre question porte sur : COMPREHENSION LANGAGE

Explication : Nous avons détecté les chaînes contextuelles suivantes :



Cet autre extrait du réseau permet de constater d'une part qu'il y a un chemin direct de "reconnaissance parole" à "compréhension langage" et que, d'autre part, "compréhension langage" est un terme auquel partant de "linguistique" on peut aboutir en transitant par "langage naturel".

Si aucun des termes n'a été retenu alors la recherche est poursuivie et le réseau s'étoffe au fur et à mesure. On peut arriver ainsi à une représentation suivante des chaînes contextuelles :

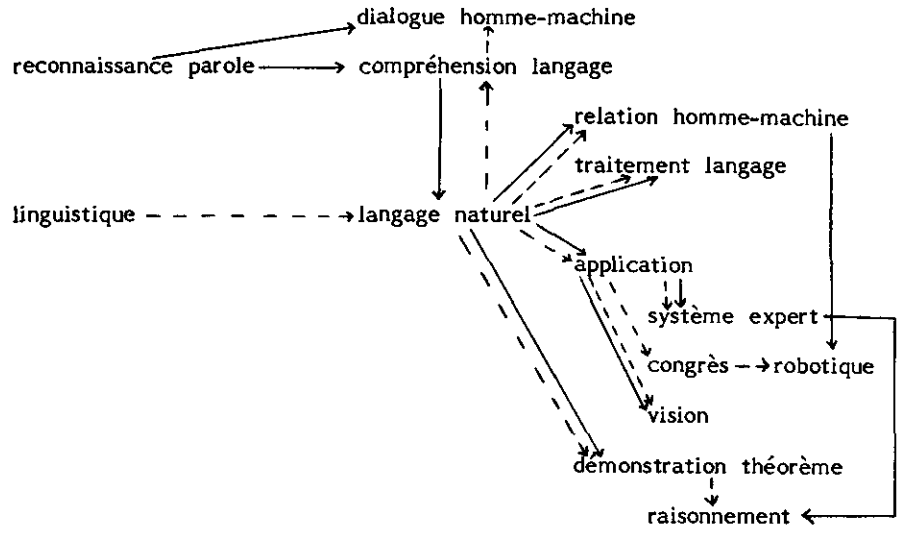


Fig. 3

- - → : chemin suivi en partant de "linguistique
——→ : chemin suivi en partant de "reconnaissance parole"

Ce réseau permet de comprendre l'utilisation des termes qui figurent dans l'indexation. Tout d'abord on observe que l'ensemble de mots, à l'exception du mot "dialogue homme-machine" directement associé à "reconnaissance parole", est obtenu en transitant toujours par deux mots proposées au départ, à savoir : "langage naturel" et "compréhension langage". Nous avons appelé ce type de mots "incontournable" le noyau sémantique dur de la question.
Le noyau sémantique dur définit avec précision le contexte d'utilisation des mots de la question dans la base et permet de prévoir le type d'articles qui seront obtenus. Ainsi, si on valide par exemple le terme "robotique" la question deviendra "reconnaissance parole linguistique robotique" et son interprétation à partir du réseau de mots associés sera la suivante : "on retrouvera des articles sur la reconnaissance de la parole et sur la linguistique écrits dans le cadre des recherches sur la compréhension du langage et le traitement du langage naturel pour les relations homme-machine, et cela dans la cadre des applications décrites par les congrès consacrés à la robotique". Les références obtenues seront donc essentiellement des renvois à des applications présentées dans des congrès.
La visualisation d'un réseau de mots associés offre la possibilité de bien choisir le mot qui est le thème de la question, car à chaque fois le mot proposé apparaît dans un contexte qui renvoie ou non au contexte de la question, et plus particulièrement à l'idée que se fait l'utilisateur de sa question et de son interprétation sémantique.
Lexiquest doit aider l'utilisateur à compléter sa question, ou autrement dit à préciser sa pensée. Par les mots qu'il propose il cherche à déclencher "ce que l'utilisateur a dans la tête", c'est-à-dire à définir au sein de quel réseau qui lui est cognitivement propre l'utilisateur inscrit la question posée.
Nous pensons, en effet, que les mots-clefs sont autant des "lieux de mémoire", des "feux de signalisation" qui permettent à l'utilisateur de se retrouver à l'intérieur d'une domaine donné. Lexiquest tient compte de cette interprétation cognitive des réseaux de mots associés.

4.2 Autres retombées de Lexiquest

Lexiquest permet, entre autres, de détecter la cohérence sémantique de la question, car dans ce cas les mêmes termes permettent, à travers d'autres termes, d'aboutir à deux thèmes différents. Pour la question "apprentissage système cognitif", on propose soit le terme "representation connaissances", soit "raisonnement", en ayant dans les deux cas les réseaux suivants :

apprentissage — — — — — — — — — → inférence
— — — — → représentation connaissances
— — — — → raisonnement

système cognitif — — — — — — — — — — → système expert
— — — — → représentation connaissances
— — — — → raisonnement

Dans ce cas, il apparaît clairement que les termes "apprentissage" et "système cognitif" sont fréquemment utilisés ensemble pour indexer des documents, mais cela a lieu dans deux contextes distincts : celui de l'inférence ou celui des systèmes experts. Il n'y a donc pas dans cet exemple de noyau sémantique dur propre aux mots de la question, car on arrive aux mots proposés en transitant par des chemins qui n'ont pas d'intersection commune permanente, comme c'était le cas dans l'exemple précédent.
Une option intéressante de Lexiquest est la possibilité de détecter des sous-questions dans une question. Dans un cas "trivial" comme : "règle production diagnostic reconnaissance parole linguistique", on signale à l'utilisateur que sa question se compose en réalité de deux questions disjointes. On les identifie car on obtient dans ce cas deux réseaux de mots associés disjoints. On propose alors à l'utilisateur de les traiter séparement. Pour une question comme "apprentissage generation hypothèse système cognitif", on remarquera que "generation hypothèse" n'est jamais utilisé dans le contexte de "apprentissage système cognitif" : compte tenu de la structure de la base d'indexation on obtient également deux questions différentes.
Parfois apparaissent dans les questions des termes qui ne figurent pas dans le lexique mais que l'on peut interpréter comme étant des mots "spécifiques" par rapport aux mots "génériques" de la base d'indexation. Ainsi "linguistique computationnelle" ne figure pas dans la base des termes d'indexation, mais on y trouve le mot "linguistique". On proposera alors à l'utilisateur de transformer sa question pour poursuivre la recherche. La question "parole linguistique computationnelle" sera d'abord transformée en "parole linguistique" puis en "reconnaissance parole linguistique", par remplacement du terme incomplet "parole". Ce traitement a pour but de donner le maximum d'ouverture au système.

5. CONCLUSION

Notre approche a l'avantage de prendre en considération simultanément tous les contextes possibles d'utilisation d'un mot donné et de faire la discrimination entre eux en fonction du contexte propre à la question et des relations de probabilité qui existent entre les différents mots de la base d'indexation.
L'intérêt de Lexiquest est double : il n'y a pas besoin de construire manuellement le réseau de mots d'un domaine, car il l'est automatiquement. Lexiquest peut donc s'adapter à toutes les bases des données bibliographiques existantes. De plus pour interroger une telle base, nul besoin de connaître à l'avance le vocabulaire de l'indexation puisque Lexiquest se charge lui-même de l'analyse et de la reformulation de la question.
Dans la version présente de Lexiquest, écrite en Le-Lisp sur un BM-60 (compatible IMB/AT) nous n'avons pas encore développé au maximum l'interface avec l'utilisateur. Celui-ci devrait s'accompagner d'un analyseur du langage naturel afin d'offrir un mode d'accès *vraiment intelligent*.

La recherche des références à partir de la question définitive n'a pas été développée dans le cadre de Lexiquest car ce système a pour but avant tout de valider la démarche de reformulation des questions dans le cadre d'une approche probabiliste à travers une utilisation particulière des mots associés.
De plus, pour une utilisation appliquée de Lexiquest, nous devrions modifier l'interface utilisateur qui explique les propositions de nouveaux mots, car une explication en langage naturel est plus facile à comprendre que la visualisation de réseau des mots associés, dont l'usager ne connait pas a priori, les principes d'interprétation. Il faudrait munir Lexiquest d'un module de génération d'explications en langage naturel, à partir de l'extrait de réseau obtenu.
Actuellement nous sommes en train de tester Lexiquest sur des microbases telles que les mycotoxines, la géologie marine et la chimie. Parallèlement nous cherchons à valider et à circonscrire la notion du "noyau sémantique dur" en appliquant Lexiquest à une étude du contexte d'apparition de mots dans un texte littéraire. Il s'agit, en occurrence, des "Exercices de Style" de R. Queneau dont la specificité littéraire (même histoire décrite de 99 façons différentes) permet d'étudier les glissements sémantiques d'utilisation de mots.
Nous n'avons pas, pour le moment, traité le problème des mots inconnus qui n'existent pas dans le lexique. Le vocabulaire scientifique et technique évolue au gré des progrès de la recherche et des nouvelles formulations, des glissements sémantiques des termes existants, apparaissent périodiquement. Ce problème est en cours d'étude dans le cadre d'un project de recherche relatif à la reindexation automatique des supports indexés par mots-clefs (SYRENA) qui fait suite à la réalisation de Lexiquest. Couplé avec ce dernier, il offrira un outil capable d'auto-apprentissage ce qui devrait rendre plus facile l'accès aux supports indexés par mots clefs.

BIBLIOGRAPHIE

[1] Bassano, J-C., DIALECT, un système-expert pour la recherche documentaire, (thèse d'Etat, Orsay, 1986).
[2] David, J.M., Créhange, M., "L'activité inférentielle de reformulation de requêtes dans le système iconographique EXPRIM", Congrès de Recherche d'Informations Assistée par Ordinateur, (Grenoble, 1985).
[3] INTELLECT, Artificial Intelligence Corporation, Query systems user's guide, (A.I. Corp. Waltham MA, 1981).
[4] voir § 2.4.
[5] Carbonell, J., Boggs, W., Mauldin, M., Anick, P., "The XCALIBUR project : a natural language interface to expert system", Proceedings of IJCAI, (1983).
Oliveira, E., Peireira, L., Sabatier, P., "ORBI : an expert system for environmental resource evaluation through natural language", Proceeding International conference of Logic Programming, (1987).
Bobrow, D.G., Bates, M., "IRUS : information retrieval using a transportable natural language interface", ACM SIGIR, 17 (4), (1983).
[6] Callon, M., Law, J., Rip, A., Texts and their powers : Mapping the dynamics of science and technology, (Macmillan, Londres, 1986).
Callon, M., Courtial, J.-P., Turner, W.A., Bauin, S., "From translation to problematic networks : an introduction to co-word analysis", Social Science Information, vol. 22, n° 2, (1983), pp. 191-235.