Indexation de textes : l'apprentissage des concepts

C. Enguehard* - P. Malvache- P. Trigano ***

* Université de Technologie de Compiègne
URA CNRS 817 - GI
BP 649
60206 Compiègne CEDEX - FRANCE
EMail : Trigano@FRUTC51.bitnet

** Commissariat à L'Energie Atomique
Centre d'Etudes de Cadarache
13108 Saint-Paul-Lez-Durance - FRANCE

ABSTRACT

In technical fields, many documents go unread due to a lack of awareness of their existence. A system which indexes texts can find all relevant texts in response to a query. The problem is to establish the indexation. At present, advanced full text systems automatically index texts on the complete thesaurus with computed weights. Another way of doing this can be a person choosing the set of relevant concepts. This second solution is better but more costly and dependent on the classification choices made by the operator.
To meet these problems, ANA (Auomatic Natural Acquisition) had been developed. This system automatically extracts relevant concepts from free texts to produce a semantic network. It does not rely on grammar or lexicon but, instead, is based on an original statistical method.
This research brings about two developments : on one hand the system is also capable of extracting the simple grammatical structures it encounters, most often in order to improve its performance, and on the other hand this will lead to an automatic definition of semantic classes of concepts, in order to structure the network.

1 - INTRODUCTION :

Le domaine des grandes bases de connaissances, rassemblant des textes, est apparu vers les années 50 comme une des applications privilégiées de la puissance des ordinateurs. Deux besoins cruciaux ont été identifiés : l'indexation des textes doit être correcte, la recherche doit être efficace en réponse à une simple question.

Au coeur de ces problèmes, se posent le choix des concepts et, plus généralement, la définition de nouveaux thésaurus. Salton avait préconisé dès 1966 l'automatisation de ces tâches car leur réalisation manuelle est coûteuse et non déterministe [SALT 66].

Nous présentons ici le système ANA (Apprentissage Naturel Automatisé) qui sélectionne les concepts (sur lesquels seront indexés les textes de la base), et les structures afin de faciliter les interrogations ultérieures.

Nous avons choisi de travailler avec le minimum de connaissances, sans analyseur syntaxique, sans dictionnaire, uniquement par l'observation statistique des textes. Les concepts sélectionnés sont alors directement issus de la langue employée. A cette exigence de simplicité, nous avons ajouté la robustesse. Le système doit supporter les dysfonctionnements que pourrait causer une lacune dans ses connaissances. Enfin, la simplicité des ressources utilisées permet au système d'auto-découvrir les connaissances dont il a besoin.

Indexation manuelle

Les systèmes les plus simples et les plus répandus sont basés sur la sélection de mots-clés dans les textes. Une question utilisant ces mots donne accès aux textes ainsi sélectionnés. Ces systèmes présentent l'inconvénient d'être très rigides : l'ajout d'un nouveau mot-clé oblige à reparcourir tous les textes déjà indexés pour y rechercher sa présence. Même automatisée, cette procédure est très contraignante. De plus Salton [SALT 86] a démontré les inconvénients de l'indexation manuelle. A titre d'exemple, deux sujets différents ne choisissent qu'à 70% des mots-clés identiques pour indexer un même document à l'aide du même thésaurus. De plus, des informations, qui, à un moment donné, ne semblent pas pertinentes à l'indexeur peuvent jouer un rôle contexte important [ANDRa]

Méthodes statistiques

Le problème du choix des concepts est contourné lorsque l'on utilise le thésaurus en entier.
Des critères purement statistiques, se référant à la valeur des termes d'indexation et non à leur sens [DACH] sont utilisés pour indexer les textes.

Très tôt, Stiles a montré l'intérêt de prendre en compte les occurrences simultanées de termes [STIL 61]. Plus récemment sont apparus les réseaux connexionistes qui permettent de gérer dynamiquement les liens et les coefficients de pondération affectant les termes d'indexation du thésaurus [KIMO 90]. Dans [ANDRc], on utilise les probabilités de Bayes actualisées en fonction des réponses et du poids sémantique des termes dans le thésaurus (ou le dictionnaire). Cette théorie oblige à distinguer homographes et synonymes car ceux-ci peuvent provoquer des biais importants. Turtle tente de simplifier les calculs de probabilité dont la complexité grandit de façon exponentielle avec la taille de la base [TURT 91].

D'autres méthodes sont développées pour représenter le contenu sémantique de chaque document, en particulier à l'aide de matrices : les lignes étant les documents et les colonnes les mots-clés. C'est la méthode de la structuration de la sémantique latente [FURN], [DEER 88], [DEER 90].

Approches mixtes

Entre ces deux extrêmes, l'intervention de l'intelligence humaine dans l'indexation manuelle, et la prise en compte de tout le thésaurus (sans compréhension), l'Intelligence Artificielle oriente ses recherches vers l'automatisation du choix des concepts porteurs de l'indexation. Le problème est alors de définir les critères qui permettront la sélection des concepts.

Certains systèmes utilisent des connaissances lexicales, syntaxiques, parfois sémantiques (les synonymes). S. David pense que l'analyse morpho-syntaxique est une étape indispensable : l'utilisation de patrons catégoriels permet d'isoler les groupes de mots intéressants [DAVI]. Ces approches linguistiques, à priori les plus appropriées, sont aussi les plus difficiles à implanter.

De nombreux systèmes mixtes font intervenir à la fois des outils linguistiques et statistiques. Le système Spirit en est un bon exemple. Les textes y sont analysés dans le but de repérer les éléments articulatoires du langage qu' utilise l'analyse linguistique pour sélectionner les concepts jugés pertinents. Des filtres statistiques évaluent les pondérations [ANDRa].

2 - PRESENTATION

Nous avons choisi d'utiliser l'apprentissage pour acquérir les concepts correspondants aux textes traités. L'apprentissage automatique du langage (russe) par le comptage d'occurrences a déjà été étudié par Andreewsky [ANDRb] mais le but était alors de découvrir la grammaire de le langue à travers l'agencement des déclinaisons.

Notre idée a été de concevoir un système aussi simple que possible avec le minimum de connaissance, même incomplète.

Ce système répond au problème du choix des concepts en n'utilisant ni l'analyse syntaxique ou sémantique ni le dictionnaire.

Nous avons essayé d'évaluer et de réduire autant que possible les connaissances, explicites et implicites, fournies au système. Celui-ci est efficace lorsque les textes se réfèrent à un domaine technique. Ils sont alors généralement écrits dans un langage dit "opératif", un langage précis comportant peu d'homographes ou de synonymes [FALZ].

La mise en oeuvre d'heuristiques très simples permet au système d'acquérir une expérience des objets familiers du domaine qui apparaissent dans les textes fournis. Cette connaissance se réfère directement au langage utilisé dans les textes, même si ceux-ci ne sont pas syntaxiquement corrects ou si les mots employés ont un sens différent de leur définition.

Nous présenterons dans un premier temps les processus mis en oeuvre dans le système ANA. Ensuite, nous

examinerons ses nouvelles fonctionnalités et les extensions que nous lui avons apportées. Enfin, seront présentés les résultats d'un test sur un corpus de 120 000 mots.

Notons que nous utilisons un modèle qui permet de définir, d'instancier et de gérer des classes d'objets et des liens (Property Driven Model [BART 79]).

Cette présentation sera illustrée de nombreux exemples pour lesquels nous nous situerons dans le cadre d'une application domestique.

3 - LE SYSTEME ANA [ENGU 91]:

Le premier objectif est le choix automatique de concepts en vu de l'indexation de textes. Un concept est la forme canonique correspondant à une classe de mots ou de syntagmes. "VERRE", par exemple, identifie les mots "verre", "verres".

• Les connaissances procédurales

Nous avons utilisé un postulat se référant à des aspects statistiques ou surfaciques du langage :

Les évènements fréquents sont significatifs.

Ce postulat peut être appliqué :
- pour rechercher des séquences de mots répétitives,
- pour identifier des configurations dénotant des concepts.

Ces configurations privilégiées sont implantées sous forme de deux modèles symétriques que l'on tentera de faire correspondre avec le texte.

Si l'on rencontre l'une de ces configurations :

mot inconnu	mot spécifiant un schéma	concept connu

concept connu	mot spécifiant un schéma	mot inconnu

Alors le mot inconnu est considéré comme susceptible de devenir un concept.

Les mots spécifiant les schémas sont acquis par apprentissage si le corpus est suffisamment important, ou donnés sous forme déclarative.

•'bootstrap'

Le réseau de concepts est initialisé par un échantillon de concepts que nous appelons 'bootstrap' dans l'esprit de Pitrat [PITR].

Durant le processus, le système essaie de confirmer cet ensemble : un concept est 'confirmé' si, ôté du bootstrap, il est découvert par le système en cours de fonctionnement.

• Les connaissances déclaratives

1 - Liste de mots vides

La liste de mots vides rassemble quelques prépositions, conjonctions, adverbes.... qui sont considérés comme non significatifs. Ils ne pourront ni obtenir le statut de concept à titre individuel, ni figurer au début ou à la fin d'un concept.

2 - Liste de mots fortement liés

Certains mots voisins peuvent se fondre en un unique mot au mépris de la nuance exprimée dans le texte. Par exemple, dans ces phrases, la variation de sens est trop fine pour que nous nous y attachions.

Il a mangé des fraises
ou *Il a mangé toutes les fraises*
La clé de la porte
ou *La clé de cette porte*
J'ai signé toutes ces lettres urgentes
ou *J'ai signé les lettres urgentes.*

Les termes fortement liés, considérés comme un unique terme, sont généralement de la forme préposition-article comme "de la" ou "d'une".

3 - Les schémas

Enfin, les configurations intéressantes signalées ci-dessus (nous les appellerons des 'schémas') prennent la forme d'une simple liste de mots comme "de", "de la", "de l", "en", etc.

• Modèle

Nous utilisons plusieurs classes d'objets :

- La classe des concepts :

Ses instances sont les concepts connus du domaine auxquels viendront s'ajouter les concepts découverts par le système. Ils sont liés par la relation "amont" qui est une première structuration du réseau.

Dans les figures les concepts sont toujours entourés d'un double cadre, dans le texte, ils sont écrits en capitales.

- Les classes des expressions et des candidats :

Ces deux classes correspondent à deux mécanismes de découverte de nouveaux concepts. Ils permettent de stocker les occurrences de textes jugées intéressantes, et de noter la fréquence de ces configurations.

Fonctionnement

1 - Analyse lexicale

L'analyse lexicale se limite à la reconnaissance des concepts connus. Toutes les marques de ponctuation sont éliminées. Cette reconnaissance est tolérante aux fautes d'orthographe et aux différentes flexions qui peuvent être rencontrées.

Le texte ainsi perçu est analysé en appliquant le postulat au contexte local autour des concepts.

2 - Recueil des occurrences

Techniquement, le texte est vu au travers d'une fenêtre de quatre mots. Les mots vides et ceux de moins de deux lettres ne sont pas pris en compte dans le calcul de l'empan de cette fenêtre.

La fenêtre est déplacée tout le long du texte, son contenu est recueilli suivant trois voies différentes en fonction de sa nature.

Cas 1

Lorsque le système voit deux concepts, il note l'occurrence, c'est à dire l'extrait de texte que laisse voir la fenêtre, dans un objet du type "expression" particulier à ces deux concepts.

ex : Soit le texte : *"je voudrais un VERRE d'EAU ou de ..."*

L'occurrence *"VERRE d EAU"* est écrite dans l'objet expression correspondant.

Cas 2

S'il ne voit qu'un concept (ici "VERRE"), le contexte local est analysé pour repérer un schéma, et donc un mot potentiellement intéressant ("lait"). Un objet de type candidat portant son nom recueille l'occurrence.

ex : Soit le texte *"j'ai renversé mon VERRE de lait devant..."*

L'occurrence *"VERRE de lait"* est écrite dans le candidat "lait"

Cas 3

Si l'examen du contexte local ne fait apparaître aucun schéma connu, l'occurrence est également conservée dans un champ spécifique. Elle sera traitée différemment.

ex : Soit le texte *"Voici de l'EAU minérale"*

L'occurrence *"Voici de l'EAU minérale"* est écrite dans le candidat "EAU"

3 - Analyse des occurrences

Cette phase de lecture est suivie de l'examen des informations recueillies. Seuls les objets ayant recueilli plus de n occurrences sont examinés.

Les expressions

Si la même configuration, aux variations morpho-syntaxiques près, se présente n fois au moins, elle devient un concept sous sa forme la plus fréquente.

ex : Voici les occurrences de l'expression rassemblant "VERRE" et "EAU" :

"je voudrais un VERRE d'EAU ou de..."
"Bois un VERRE d'EAU pour faire..."
"aspirine dans ton VERRE d'EAU..."

L'analyse va qualifier le nouveau concept "VERRE D'EAU"

Les candidats et les schémas

Les candidats dont la fréquence est supérieure au seuil m deviennent eux-mêmes des concepts sous la forme morpho-syntaxique la plus fréquente.

ex : Voici les occurrences du candidat "lait" :

"j'ai renversé mon POT de lait devant..."
"distribuer un VERRE de lait à chacun..."
"Boire un VERRE de lait c'est..."
"Je préfère un VERRE de lait nature..."
"J'ai vidé la BOUTEILLE de lait qui était..."

L'analyse va qualifier le nouveau concept "LAIT"

Les candidats sans schéma

Les concepts existants présentant n fois le même contexte local engendrent un nouveau concept intégrant ce contexte.

ex : Voici les occurrences sans schéma du candidat "VERRE" :

"Bois un grand VERRE cela ira mieux..."
"J'ai acheté un VERRE à bière..."
"Voici ce grand VERRE dont je t'ai parlé..."

L'analyse va qualifier le nouveau concept "GRAND VERRE"

Les seuils n et m sont arbitrairement fixés aux valeurs 3 et 5 qui se sont expérimentalement révélées correctes pour des corpus de 40 000 à 200 000 mots. Cependant il semblerait nécessaire de les rendre adaptatifs quand le corpus devient très grand.

Le réseau obtenu

Nous représentons les résultats obtenus sur les exemples précédents :

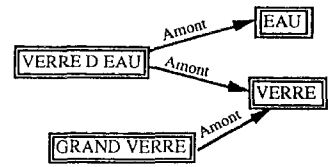

figure 1

4 - Les résultats :

Ce système répond de façon très satisfaisante à nos attentes.

Voici les résultats de son fonctionnement sur des textes totalisant environ 120 000 mots et provenant d'interviews relatives au retour d'expérience du démarrage du réacteur à neutrons rapides Super-Phénix.

La base initiale comprenait 350 concepts effectivement utilisés dans les textes analysés.

L'analyse a donné lieu à la découverte de 700 nouveaux concepts dont les deux-tiers ont été jugés qualitativement très bons. D'autre part, 260 des concepts du bootstrap ont été confirmés.

D'autres résultats sont détaillés dans [ENGU 91].

4 - LES EXTENSIONS

Nous abordons l'apprentissage des connaissances utilisées pour l'apprentissage ! Nous avons vu comment découvrir des concepts. Le système va maintenant apprendre une partie des connaissances nécessaires à ce premier apprentissage, c'est à dire les connaissances déclaratives : la liste des mots vides, la liste de mots fortement liés et les mots spécifiant les schémas.

Les résultats de cet apprentissage, les listes que le système va établir, ne seront pas exactement identiques aux listes fixées à l'avance qui, jusqu'à présent, lui étaient fournies. Nous nous attendons à ce que son fonctionnement en soit amélioré : le processus va négliger certains schémas, rares dans l'échantillon, en mettre de nouveaux à jour auxquels nous n'avions pas pensé. Bref, l'adéquation à la langue manipulée dans les textes sera meilleure.

Les extensions de l'apprentissage

Le postulat est appliqué à la structure interne des concepts afin de découvrir la façon dont ils sont formés.

Les configurations les plus fréquentes fourniront des généralisations qui serviront à dégager les schémas de découverte des nouveaux concepts.

Examinons l'apprentissage des connaissances déclaratives qui auparavant étaient fournies au système : la liste des mots vides, la liste de mots fortement liés et les mots spécifiant les schémas.

L'apprentissage des connaissances déclaratives

Afin de modéliser la structure interne des concepts, nous définissons une nouvelle classe d'objets.

Une nouvelle classe d'objets : les termes

Les termes sont les mots composant les concepts. Ils sont liés entre eux par la relation "voisin" qui mémorise la fréquence de chaque association. De chaque terme nous connaissons le nombre d'occurrences et le fait qu'il soit, ou non, concept à titre individuel. Les termes sont entourés d'un simple cadre dans les représentations graphiques.

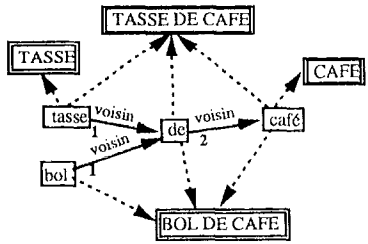

figure 2

Cet exemple montre la décomposition des concepts "TASSE DE CAFE" et "BOL DE CAFE" en trois termes chacun. Remarquons que les termes "TASSE" et "CAFE" sont eux-mêmes des concepts, alors que "bol" et "de" n'en sont pas.

Détermination de la liste de mots vides

Pour obtenir une liste de bonne qualité, il est nécessaire d'utiliser un échantillon de textes d'au moins 40 000 mots, soit environ 100 pages (minimum issu de l'examen de différents corpus).

Le système lit l'échantillon et compte tous les mots qu'il rencontre. Un terme est ici strictement défini par sa forme, par la chaîne ordonnée de caractères qui le composent. Ainsi, "chaise" et "chaises" sont considérés comme deux termes différents.

Les différents termes sont ensuite classés en fonction de leurs fréquences décroissantes et affectés d'un numéro correspondant à leur rang.

La courbe, fréquence = f (log (rang)), est seuillée au rang s tel que l'aire As (définie par la courbe, l'axe des abscisses, l'axe des ordonnées et la droite x = s), approche 95 % de l'aire totale A (définie par la courbe et l'axe des abscisses).

Soient : n, le nombre de termes de l'échantillon,

$$A = \sum_{x=1}^{n} f(x).\log(x) \quad ,$$

$$A_s = \sum_{x=1}^{s} f(x).\log(x)$$

Cherchons s tel que

$$A_s < 0{,}95\ A < A_{s+1}$$

Dès lors, tous les mots de rang $x \leq s$ sont des mots vides. Ils sont écrits dans la liste adéquate.

Détermination des mots fortement liés et des mots de schémas

Les mots caractérisant les schémas ont la propriété de lier des concepts. Nous utilisons cette particularité pour les isoler.

A l'initialisation du système nous disposons de l'ensemble des concepts donnés dans le bootstrap.

Dans un premier temps, éliminons les concepts composés de plusieurs termes, ceux-ci risqueraient de biaiser notre analyse future, et travaillons avec les seuls concepts simples.

La première opération utilise ces concepts et un échantillon de textes pour en déduire des concepts composés par la collecte d'occurrences associées à des expressions. A ce stade, aucune connaissance n'intervient, nous ne faisons qu'appliquer le postulat pour regrouper les concepts présents afin d'en former de plus complexes.

Au fur et à mesure de leur création, ces concepts sont décomposés en termes. Nous utilisons une information cruciale attachée à chaque terme : Est-il un concept de façon individuelle ?

Alors, les listes que nous cherchons peuvent être établies
- les mots fortement liés :
Ce sont les couples de termes voisins qui ne sont ni l'un ni l'autre des concepts à titre individuel.
- les mots de schéma :
Pour accéder à ce statut, un terme doit vérifier plusieurs critères :
- Il n'est pas un concept de façon individuel,
- Ses fréquences de voisinage droit et gauche sont du même ordre de grandeur,
- Il lie souvent des termes qui, eux, sont des concepts.

Quelques résultats

Nous avons appliqué ce nouveau processus à un échantillon de **60 000 mots.**

L'analyse statistique établit une liste de 35 **mots vides** :
"a", "au", "avait", "c", "ce", "cela", "d", "dans", "de", "des", "donc", "du", "en", "est", "et", "était", "fait", "il", "je", "l", "la", "le", "les", "n", "ne", "on", "pas", "pour", "qu", "que", "qui", "sur", "un", "une", "y".

Nous obtenons **110 nouveaux concepts** dont voici quelques exemples :

"CAPTEURS DE DEPLACEMENT"
"CIRCUIT DE VIDANGE DE L'INTERCUVE"
"CODES DE CALCULS"
"CONTROLE COMMANDE"
"CUVE DE RETENTION"

L'analyse de ces concepts permet :

- d'**unifier les termes** :
  "à la", "de l", "de la".

- de qualifier les **termes caractéristiques de schémas** :
  "de la", "d", "des", "de", "du".

Nous constatons que les mots de schémas retrouvés par le système sont les plus productifs quant aux nouveaux concepts qu'ils sont susceptibles de découvrir.

D'autres résultats seront exposés durant la conférence.

5 - CONCLUSION

Le contrôle des connaissances de notre système ainsi que leur introduction sous forme déclarative nous ont permis d'exploiter le réseau de concepts et de termes.

Toutefois, il nous reste à explorer de nouvelles extensions vers une plus grande structuration du réseau : la définition automatique de classes de mots.

Le processus d'induction de ces classes sera basé sur l'examen des contextes droits et gauches des termes composant les concepts. L'utilisation des termes dans le langage reflétant la manipulation des objets dans le monde physique. Cet isomorphisme présupposé des structures, des termes et des objets, correspond à la théorie psychologique de capture des classes par prototypage.

6 - BIBLIOGRAPHIE

[ANDRa] Andreewsky A., Fluhr C., "Indexation automatique - Construction automatique des thésaurus - classification automatique", Note CEA-N-1795

[ANDRb] Andreewsky A., Fluhr C., "Le problème de l'identification automatique des concepts", Note CEA-N-1816

[ANDRc] Andreewski A., Debili F., Fluhr C., Hlal Y., Nicaud L., "Résumé des problèmes de l'indexation automatique tels qu'ils sont abordés par le groupe de recherche en linguistique automatique"

[BART 79] Barthes J.P., Vayssade M., Znamierovska M., "Property Driven Databases", IJCAI'79, Tokyo, 1979

[DACH] Dachelet R., "Etat de l'art de la recherche en informatique documentaire : la représentation des documents et l'accès à l'information", Rapport n°1201 - 32 pages - Programme 8 - Communication homme-machine, INRIA

[DAVI] David S., Plante P., "De la nécessité d'une approche morpho-syntaxique en analyse de textes"

[DEER 88] Deerwester S., Dumais S.T., Furnas G., Landaeur T.K., Harshman R., "Using latent semantic analysis to improve access to textual information", CHI'88, pp : 281 - 286

[DEER 90] Deerwester S., Dumais S.T., Furnas G., Landaeur T.K., Harshman R., "Indexing by latent semantic analysis", Journal of the american society for information science, pp : 391 - 407, n° 41, 1990

[ENGU 91] Enguehard C., Malvache P., "Apprentissage Naturel Automatisé", Convention IA 91, pp : 145 - 163, 1991

[FALZ 89] Falzon P., "Ergonomie cognitive du dialogue", Presses Universitaires de Grenoble, Sciences et Technologies de la connaissance, chapitre 4, 1989

[FURN] Furnas G.W., "Information retrieval using a singular value decomposition model of latent semantic structure", 11th ACM International Conference on research and development in information retrieval, pp : 465 - 480

[KIMO 90] Kimoto H., Iwadera T., "Construction of a dynamic thesaurus and its use for associated information retrieval", 13th International Conference on research and development in information retrieval, pp : 227 - 240, 1990

[PITR] Pitrat J., "Textes, ordinateurs et compréhension", Eyrolles, 1985

[SALT 66] Salton G., "Information dissemination and automatic information systems", Proc. IEEEE, 54, 12, December, 1966

[SALT 86] Salton G., "Another look at aotomatic text-retrieval systems", Communications of the ACM, 29 (7), pp : 648 - 656, 1986

[STIL 61] Stiles H.F., "The association factor in information retrieval", journal of the ACM, vol. 8, pp : 271-279, 1961

[SPIR] Spirit, Présentation + Manuel utilisateur

[TURT 91] Turtle H.R , Croft W.B, "Efficient probalistic inference for text retrieval", RIAO'91, p : 644