ETUDE D'UN ANALYSEUR DE SURFACE DE LA LANGUE NATURELLE POUR UN SYSTEME DE RECHERCHE DOCUMENTAIRE

Patrick PALMER, Catherine BERRUT
Groupe "Systèmes Intelligents de Recherche d'Informations"
Laboratoire Génie Informatique, Institut IMAG Grenoble
BP 68, 38402 ST MARTIN D'HERES CEDEX, FRANCE

RESUME

Dans le cadre d'un système de recherche documentaire, nous étudions une stratégie d'analyse automatique du Français destinée à reconnaître des entitées conceptuelles structurées, issues de syntagmes nominaux, pour une indexation de documents. Nous nous sommes principalement attachés à privilégier l'aspect automatique de cette analyse, en intégrant un enrichissement automatique du vocabulaire, et à en minimiser le coût en nous limitant à une analyse syntaxique partielle et en utilisant une organisation de dictionnaire appropriée.

I - INTRODUCTION

Nous présentons la définition d'un système d'analyse partielle de textes en langue naturelle (Français), pour une indexation automatique de documents. Nous opérons dans le cadre d'une base de données textuelles. Il nous faut donc définir des outils spécifiques d'analyse de textes pour des corpus importants, d'où notre orientation vers une analyse de surface de la langue naturelle.

Ce travail se situe en amont des traitements d'indexation et constitue la composante linguistique d'un système documentaire. Le but visé est l'extraction, à partir d'un texte en langue naturelle, d'éléments structurés (nommés Groupes Conceptuels), contenant l'essentiel de la connaissance véhiculée.

Pour la réalisation de cet objectif, nous dégageons trois axes principaux:

- Reconnaissance des Groupes Conceptuels (GC), prédéfinis syntaxiquement, sans faire appel à des stratégies d'analyse coûteuses.
- Enrichissement automatique du vocabulaire, afin de ne pas entraver le processus d'analyse par des interventions de l'utilisateur lors de la rencontre d'un mot inconnu.
- Compatibilité des algorithmes et structures avec une microprogrammation, envisagée ultérieurement, afin d'améliorer les performances du système.

II - DEFINITION DES GROUPES CONCEPTUELS (GC)

Nous désirons dépasser le stade d'indexation par "mots-clés", d'aide à l'indexation (manuelle) des systèmes classiques (Mistral, Golem, Stairs), pour arriver à une indexation automatique à l'aide de "Groupes Conceptuels" (G.C.), (Kerkouba, 1981).

Pour définir ces groupes, nous avons été amenés à rechercher une entité conceptuelle reconnaissable lors d'une analyse de surface de la langue naturelle. Ces entités doivent représenter les concepts véhiculés par le texte. La sélection des concepts représentatifs d'un document est effectuée lors d'une phase ultérieure par un module d'indexation indépendant (Kerkouba, 1985). Il est généralement admis dans le domaine que l'information conceptuelle nécessaire à l'indexation est en grande partie portée par les éléments constitutifs des syntagmes nominaux, ceci étant particulièrement vérifié en Français (Le Guern, 1982). On peut distinguer à l'intérieur des syntagmes nominaux, selon leurs catégories grammaticales, les éléments porteurs sémantiquement, des éléments jouant le rôle d'outils syntaxiques. L'objectif de notre système d'analyse est donc la détection de ces syntagmes nominaux pour en extraire un sous-ensemble "porteur" constituant le groupe conceptuel.

Nous avons déterminé les groupes conceptuels que nous souhaitons reconnaître en établissant la syntaxe suivante, cette syntaxe-cible définit en quelque sorte un squelette de syntagme nominal, ordonné.

```
GC    -->  ARG RELP ARG  /  NOM GQ  /  ARG
ARG   -->  NOM  /  VINF  /  NOM NOM
GQ    -->  CPN  /  ADJ+ [GADJ]  /  GADJ
CPN   -->  PDE ARG  /  PDE NOM GQ
GADJ  -->  ADJ PREP VINF  /  ADJ PREP NOM ADJ*
PREP  -->  PDE  /  RELP
NOM   -->  Substantif commun  /  Substantif propre
ADJ   -->  Adjectif qualificatif  /  Adjectif numéral ordinal
           Adjectif numéral cardinal  /  Participe passé
VINF  -->  Verbe à l'infinitif
PDE   -->  Prépositions: de, du, des, d'
RELP  -->  Autres prépositions
```

Commentaires

```
ARG  =  ARGument
GQ   =  Groupe Qualificatif
CPN  =  ComPlément de Nom
GADJ =  Groupe ADJectival
PREP =  PREPosition
```

Remarques:

- a) Le groupe adjectival, GADJ, n'est reconnu qu'à partir d'un adjectif en position d'attribut.

Ex: "Le programme entier est difficile pour les étudiants."

```
GADJ  -->  difficile pour étudiant
GC    -->  programme   entier   difficile pour étudiant
                NOM       ADJ            GADJ
```

- b) Dans les titres, on peut trouver une RELP entre deux noms.

Ex: "Le prêt dans les bibliothèques."

```
GC    -->  prêt   dans   bibliothèque
```

Cette syntaxe-cible permet de représenter trois catégories d'éléments:

- Le mot isolé ARG. La structure NOM NOM est considérée comme un mot isolé (nom composé).

Ex: "Plan Barre" "bleu ciel"

- Le groupe conceptuel simple GCS dont la syntaxe est:

GCS --> NOM ADJ / NOM PDE ARG / ARG RELP ARG

On peut remarquer que certaines structures permises par la syntaxe-cible ne seront jamais rencontrées en Français,

EX: NOM PDE VINF

celà n'est pas génant, dans la mesure où notre but est la reconnaissance

de structures présentes dans les textes.

- Le groupe conceptuel complexe GCC: C'est tout autre groupe (GC), n'appartenant pas aux deux catégories précédentes.

Cette distinction sera nécessaire pour la sélection, lors de la phase d'indexation, des groupements jugés représentatifs d'un texte (Kerkouba, 1984). Pour réaliser cette sélection, nous considérons d'abord les groupes conceptuels les plus longs. S'ils ne sont pas retenus, une cassure syntaxique est alors effectuée, suivant cette distinction, pour considérer les groupes issus du groupement rejeté.

Relations entre les groupes conceptuels

La définition des groupes conceptuels reste extensible. En particulier lors d'un développement ultérieur, le rajout d'un sous-ensemble des syntagmes verbaux reste possible. Ce rajout deviendrait nécessaire si l'on s'orientait vers une représentation de la connaissance.

Une sélection composée uniquement de groupes conceptuels perdrait une partie du sens ou de la connaissance, contenus dans la phrase. C'est pourquoi, lors du processus dindexation, les groupes conceptuels sélectionnés pourront être connectés par un ensemble de relations reconnues au cours de l'analyse textuelle. Ces relations interprètent en quelque sorte les syntagmes verbaux.

Dans un premier temps, nous nous sommes limités à trois types de relations:

- Les relations prépositionnelles. Ces relations sont introduites par les prépositions, suivant ou précédant, directement le verbe. Elles traduisent donc la notion de complément d'objet indirect.

Ex: "L'oiseau niche sur la branche."
==> oiseau RP_sur branche

- La relation générique AUX. Cette relation est introduite par l'auxillaire "être" lorsqu'il n'est pas suivi d'une préposition, c'est à dire lorsqu'il n'introduit pas de complément d'objet indirect (cas précédent).

Ex: "La rose est une fleur."
==> rose AUX fleur

- La relation de proximité PROX. Cette relation permet de connecter un groupe conceptuel sujet avec un groupe conceptuel complément d'objet direct.

Ex: "L'enfant mange un gâteau."
==> enfant PROX gâteau

Ces relations permettent de représenter un noyau d'informations extrait lors de l'analyse du texte, automatiquement et sans en interpréter le sens.

Le choix de ces relations prépositionnelles et de la relation générique doit permettre, au cours d'une phase ultérieure, par une étude

de leurs propriétés de réaliser des inférences permettant de compléter le travail classique effectué au niveau du thésaurus.

La reconnaissance de ces groupes conceptuels répond aux impératifs fixés par l'évolution du domaine vers la constitution de bases de données textuelles où le texte est entièrement saisi, d'où la définition d'outils spécifiques de traitement de textes en langue naturelle pour des corpus importants. Dans ce sens, l'automatisation de cette analyse devra être la plus poussée possible afin de limiter au maximum les interventions manuelles entravant le processus d'analyse.

Dans le but d'assurer une certaine généralité au système, nous avons renoncé à reprendre la démarche classique consistant à définir un environnement sémantique des concepts a priori. Nous n'établissons pas de liste de mots vides (sans information sémantique), par contre, nous considérons que certaines catégories grammaticales (articles, pronoms,...) ne représentent que des outils, utiles pour la levée des ambiguïtés grammaticales (homographies). Ces ambiguïtés inhérentes au langage naturel constituent le principal obstacle à tout traitement automatique de textes en langue naturelle (Courtin, 1977). Une des caractéristiques essentielle de notre analyse est de ne tenter la levée de ces ambiguïtés que dans la mesure où elles nous serviront pour l'extraction des groupes conceptuels.

III - COMPOSANTE MORPHOLOGIQUE

Nous effectuons une analyse morphologique classique de gauche à droite consistant en une segmentation d'une forme en racine + désinence. Les racines se trouvent dans le dictionnaire d'analyse, les désinences dans de multiples petites tables spécifiques. Pour une même forme, toutes les segmentations possibles connues sont mémorisées.

Nous utilisons un dictionnaire d'analyse où toutes les racines sont factorisées sous forme d'arbre lexicographique et rangées selon un critère de fréquence pour optimiser les accès (Palmer, 1981). Chaque fin de racine est matérialisée par le positionnement d'un indicateur. Au cours de l'analyse, la rencontre d'un tel indicateur valide un ensemble de tables de désinences. La reconnaissance est alors poursuivie parallèlement dans ces tables et dans le dictionnaire sans "back-tracking".

Cette organisation est compatible avec un enrichissement automatique du vocabulaire, toute insertion d'une nouvelle racine se faisant facilement soit en "feuille" dans l'arbre, soit par le positionnement d'un indicateur de fin de racine, si elle est préfixe d'une racine déjà insérée. Cet enrichissement du vocabulaire contribue fortement à l'automatisation de l'analyse en limitant les interventions manuelles pour la reconnaissance et l'intégration des formes inconnues, contrairement aux systèmes classiques. Cette possibilité est particulièrement intéressante dans le domaine de l'informatique documentaire où les corpus à traiter sont volumineux.

Pour réaliser cette acquisition, il est nécessaire de pouvoir déterminer la catégorie grammaticale du mot inconnu. Cette information est obtenue au niveau de la composante syntaxique, d'où nécessité d'un parallélisme entre syntaxe et morphologie.

Nous nous sommes restreints à l'enrichissement automatique des formes régulières des classes ouvertes du Français, en ayant au préalable consigné lors d'une phase d'initialisation du dictionnaire les classes fermées et l'ensemble des formes irrégulières. Nous pouvons remarquer que l'ensemble des irrégularités du Français provient d'un "héritage" de la langue et que, par conséquent, nous pouvons considérer ces irrégularités également comme une classe fermée du Français: l'évolution actuelle de la langue Française régularisant systématiquement les mots nouveaux (Grevisse, 1980).

La catégorie grammaticale d'une forme inconnue ne pourra donc être que l'une des catégories suivantes: Substantif, Adjectif qualificatif, Verbe, Adverbe. Ces quatre catégories sont automatiquement attribuées par défaut aux formes inconnues du dictionnaire d'analyse.

Lors du choix des catégories grammaticales, nous avons relativement diversifié les classes dites fermées, puisqu'elles sont présentes dans le dictionnaire, ce qui permet d'affiner leur rôle syntaxique.

Dans un tel système dont le but est l'extraction de groupes conceptuels, la non-reconnaissance d'un mot en raison d'une mauvaise orthographe ne porte pas à conséquence; le mot erroné peut être appris mais ne sera pas réutilisé. Pour une meilleure cohérence, il est possible d'envisager périodiquement des sessions de vérification du nouveau vocabulaire enregistré.

IV - COMPOSANTE SYNTAXIQUE

A la sortie de l'analyse morphologique, nous obtenons pour une phrase un réseau dont les noeuds sont consitués des formes rencontrées, auxquelles sont associées une ou plusieurs listes de propriétés grammaticales. Ces listes peuvent éventuellement être vides. Ces formes peuvent être identifiées de manière non-ambiguë, ou constituer des homographies, ou encore être inconnues.

Dans ce réseau, certaines formes parfaitement identifiées peuvent constituer des points d'ancrage intéressants pour les levées d'ambiguïtés.

Ces levées d'ambiguïtés ne seront pas systématiques; en effet, il n'est pas utile de disposer de la structure, ou des différentes possibilités de structures, syntaxiques complètes d'une phrase (mécanisme d'analyse très coûteux) pour en reconnaître les entités susceptibles de constituer les groupes conceptuels.

Nous procédons de la façon suivante :

1 - Utilisation d'un filtre syntaxique.

Ce filtre syntaxique est basé sur les relations positionnelles des mots dans une phrase, représentées synthétiquement sous forme d'une matrice de précédence booléenne. Chaque élément de la matrice indique la possibilité pour deux catégories grammaticales de se trouver successivement dans un texte (Andreewsky, 1973), (Fluhr, 1977). Les impossibilités reconnues permettent de simplifier la combinatoire. Cette matrice a été constituée de deux manières:

- par apprentissage sur un texte résolu manuellement,
- en s'intéressant plus particulièrement aux catégories jouant un rôle syntaxique intéressant dans la composition des groupes nominaux.

Nous pouvons envisager d'étudier chaque case de la matrice séparément, vu le nombre relativement limité des éléments (matrice carrée 40 X 40). L'emploi de ce mécanisme permet d'augmenter le nombre de points d'ancrage (noeuds non-ambigus du réseau).

2 - Détection des ambiguïtés intéressantes et des points d'ancrage pouvant aider à leur résolution.

Nous entendons par ambiguïté intéressante toute ambiguïté dont la résolution peut déboucher sur un composant d'un groupe conceptuel, c'est-à-dire sur une catégorie grammaticale intervenant dans la syntaxe-cible. Pour cela, nous avons établi une liste des éléments constitutifs susceptibles d'intervenir dans les groupes nominaux. Un point d'ancrage appartenant à cette liste est un point de départ prépondérant pour le système d'extraction des groupes conceptuels.

Deux classes de règles permettent la levée de ces ambiguïtés:

- Un ensemble de règles syntaxiques classiques portant sur les compatibilités des propriétés grammaticales (genre, nombre, ...).
- Un ensemble de règles de constitution des éléments des classes ambiguës. Ces règles sont constituées de manière ad-hoc en fonction des ambiguïtés les plus fréquentes dans les textes d'apprentissage (Merle, 1982).

Nous disposons en sus des listes de terminaisons possibles pour certaines catégories grammaticales (verbe, adverbe) et de règles syntaxiques de regroupement pour les locutions non rentrées dans le dictionnaire d'analyse.

Cette composante syntaxique se décompose en deux phases:

1 - Simplification de la combinatoire du réseau constitué par les différentes solutions grammaticales des formes analysées, menée en parallèle avec la morphologie, y compris la détermination de la catégorie grammaticale des mots inconnus. Cette simplification est effectuée par l'utilisation de la matrice de précédence booléenne.

2 - Processus de levée d'ambiguïtés grammaticales déclenché uniquement pour les ambiguïtés intéressantes pour lesquelles la résolution paraît devoir aboutir (nombreux points d'ancrage à proximité).

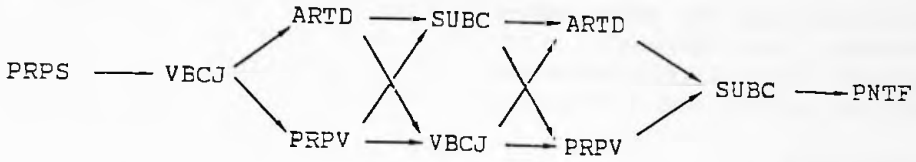
Exemple: Simplification de la combinatoire du réseau.

Soit la phrase: "Il ouvre les portes le matin."

==> Listes de catégories grammaticales suivantes:

il	ouvre	les	portes	le	matin	
PRPS	VBCJ	ARTD	SUBC	ARTD	SUBC	PNTF
		PRPV	VBCJ	PRPV		

==> on obtient le réseau ambigu suivant:



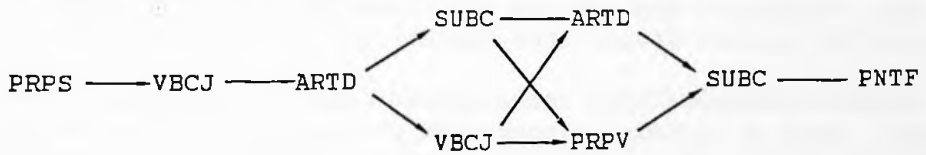
Nous pouvons simplifier ce réseau en utilisant le sous-ensemble de la matrice de précédence suivant:

	VBCJ	PRPV	SUBC	ARTD
VBCJ	0	0	0	1
PRPV	1	1	0	0
SUBC	1	1	1	1
ARTD	0	0	1	0

Dans laquelle nous pouvons lire:

VBCJ-ARTD = 1 (permis) et VBCJ-PRPV = 0 (interdit)

==> Le réseau simplifié suivant:



Puis, aprés application du même type de règles, nous obtenons le réseau final non-ambigu suivant:

il	ouvre	les	portes	le	matin	
PRPS	VBCJ	ARTD	SUBC	ARTD	SUBC	PNTF

V - STRUCTURE DU DICTIONNAIRE

La structure du dictionnaire utilisé permet une analyse caractère par caractère sans "back-tracking". En effet, la factorisation des racines autorise l'exploration de toutes les décompositions possibles d'une forme en une seule passe. Afin d'améliorer les performances de l'analyse, notre effort a également porté sur l'optimisation des accés au dictionnaire permettant un classement fréquentiel des racines. Pour cela, nous nous sommes inspirés de la technique "Median Split Tree"

(Sheil, 1978) qui consiste en l'utilisation d'arbres binaires organisés suivant deux relations d'ordre. Dans notre application, la première relation est basée sur les fréquences des caractères, la seconde est l'ordre lexicographique. Lors de l'initialisation du dictionnaire avec les classes fermées et les irrégularités du Français, chaque racine a été rentrée avec un poids calculé à partir d'un dictionnaire de fréquences (T.L.F., 1971) constitué sur un très gros corpus de textes représentatifs de la littérature française. Un algorithme de réorganisation permet de réarranger périodiquement le dictionnaire.

Nous envisageons lors d'une phase ultérieure une microprogrammation des algorithmes et structures afin d'en améliorer les performances. La technique d'analyse employée et l'organisation du dictionnaire ont été influencées par cet objectif.

VI - CONCLUSION

Cet analyseur doit fournir en sortie un document restreint aux groupes conceptuels reconnus, éventuellement connectés par des relations (prépositionnelles, génériques, ou de proximité). Une version simplifiée de la syntaxe des Groupes Conceptuels a été testée sur un échantillonage de textes pleins (Kerkouba, 1984), (Bruandet, 1985) et a donné des résultats prometteurs pour une indexation de documents.

Certaines parties de cet analyseur, telles que les modules de gestion du dictionnaire et l'analyseur morphologique, ont été implémentées en PASCAL. Nous avons utilisé les langages PROLOG et LISP pour une évaluation qualitative de la syntaxe des Groupes Conceptuels, la version définitive devant être implémentée en PASCAL.

L'étape suivante de notre étude est l'intégration de notre analyseur dans un système d'indexation automatique. La syntaxe proposée restant extensible, nous envisageons de compléter cet analyseur pour l'adapter à la reconnaissance de la structure syntaxique des requêtes d'interrogation en langue naturelle (Defude, 1984).

REFERENCES BIBLIOGRAPHIQUES

ANDREEWSKY A. , FLUHR C.
Apprentissage, analyse automatique du langage, application à la documentation
Dunod, Doc. de linguistique quantitative n° 21, 1973

BRUANDET M.F.
Partial knowledge model for an information retrieval system
Conférence RIAO85, Grenoble 18-20 mars 1985

COURTIN J.
Algorithmes pour le traitement interactif des langues naturelles
Thèse d'état, Grenoble 1977

DEFUDE B.
Knowledge based system versus thesaurus: an architecture problem about expert system design
3rd ACM and BCS Symposium, Research and development in information retrieval, Cambridge 2-6 july 1984

FLUHR C.
Algorithmes à apprentissage et traitement automatique des langues
Thèse d'état, Paris 1977

GREVISSE M.
Le bon usage
Hatier, Paris dernière éd. 1980

LE GUERN M.
Les descripteurs d'un système de documentation. Essai de définition
Colloque Traitement automatique des langues naturelles et systèmes documentaires, Clermont-Ferrand 6-7 mai 1982

KERKOUBA D.
Incidence du thésaurus dans les systèmes documentaires
Rapport de DEA, Grenoble 1981

KERKOUBA D.
Indexation et propriétés structurelles de documents dans un système de recherche d'informations
Thèse de Docteur-Ingénieur, Grenoble 1984

KERKOUBA D.
Automatic indexing and structural properties of texts
Conférence RIAO85, Grenoble 18-20 mars 1985

MERLE A.
Un analyseur pré-syntaxique pour la levée des ambiguïtés dans des documents écrits en langue naturelle: Application à l'indexation automatique
Thèse de Docteur-Ingénieur, Grenoble 1982

PALMER P.
Etude de l'organisation d'un dictionnaire pour l'analyse du Français
Rapport de DEA, Grenoble 1981

SHEIL B.A.
Median split tree: A fast lookup technique for frequently occuring keys
Communications of the ACM, november 1978

T.L.F.
Dictionnaire des fréquences
CNRS T.L.F., Nancy 1971