Extraction des connaissances terminologiques au moyen des Grammaires Catégorielles : un modèle hybride

Ismaïl Biskri ; Jean-Guy Meunier ; Georges Nault
Laboratoire de l'ANalyse Cognitive de l'Information
Université du Québec à Montréal
biskri@pluton.lanci.uqam.ca
meunier.jean-guy@uqam.ca
nault@pluton.lanci.uqam.ca

Actes de 4e Jour. Int. d'Analyse Statistique des Données Textuelles *Nice. 1988p, 123- 135*

Résumé :

Une des recherches de pointe menée actuellement est l'extraction des connaissances dans un texte électronique (Textual data mining). Ce thème de recherche est de première importance pour les technologies de l'information qui sont confrontées à des marées de documents électroniques non formatés. Pour résoudre ce problème, plusieurs stratégies sont possibles, les unes sont mathématiques et les autres sont linguistiques computationnelles. Nous présenterons dans cet article une approche pour un modèle hybride pour l'extraction des connaissances. Notre modèle s'inspirera des modèles neuronaux et de la Grammaire Catégorielle Combinatoire Applicative (GCCA) (Biskri, 1995) (Desclés, Biskri, 1996).

1. Introduction :

De nos jours, un nombre croissant d'institutions accumulent très rapidement des quantités de documents qui ne sont souvent classés ou catégorisés que très sommairement. Très vite, les tâches de dépistage, d'exploration et de récupération de l'information présente dans ces textes, c'est-à-dire des "connaissances", deviennent extrêmement ardues, sinon impossibles. Pour y faire face, il devient nécessaire d'explorer de nouvelles approches d'aide à la lecture et à l'analyse de texte assistées par ordinateur (LATAO)
Du point de vue méthodologique, la question de l'extraction des connaissances dans les textes rencontre des difficultés épistémologiques sérieuses. En raison de sa nature sémiotique et langagière, le traitement informatique traditionnel d'un texte est de nature linguistique. Un texte est vu comme une suite de phrases qu'on doit soumettre à des analyseurs linguistiques. Cette approche semble tout à fait naturelle, elle correspond théoriquement au processus naturel de lecture d'un humain. Cependant , cette approche s'avère problématique dès lors qu'il s'agit d'une grande masse de données textuelles.
Dans ce cadre, le traitement d'un texte par ordinateur en appelle à des dépôts de connaissances préconstruites acquises via des enquêtes cognitives (analyse de protocole) auprès des experts ou puisées dans le répertoire encyclopédique du savoir partagé. Ceux-ci sont alors utilisées comme gabarit dans le dépistage et la reconnaissance. De plus les systèmes experts qui opèrent dans ce domaine doivent être dotés des mécanismes habituels (moteur d'inférence, maintien de cohérences, tests de plausibilité, etc.) leur permettant d'effectuer des déductions et des tests d'hypothèses avec un haut niveau de confiance et de réussite. Les connaissances comportent des représentations d'objets, de propriétés, de relations d'événements et de situations propres à l'objet à traiter, en l'occurrence le contenu informationnel du texte. En possession de ce savoir, ce système informatique de type expert pourrait alors réussir ˆ "comprendre" le texte et donc en extraire les connaissances. De nombreuses recherches ont d'ailleurs montré la nécessité d'avoir les connaissances de multiples niveaux (syntaxiques, psycholinguistiques, lexicales, sémantiques, encyclopédiques, etc.). (Regoczei, Plantinga, 1988 Gaines & Shaw 1988, Jacobs & Zernik 1989; Moulin et Rousseau, 1990, Zarri 1990.)

Du point de vue de la lecture et de l'analyse de texte assistées par ordinateur (LATAO), le problème de l'extraction des connaissances d'un corpus textuel se présente de manière totalement différente. Il est en effet délicat de donner à priori à l'ordinateur, les connaissances que le texte avait pour fonction de transmettre sauf peut-être, pour celles qui sont de nature générale, encyclopédique ou technique. Dans le cadre de LATAO, la connaissance se trouve dans le texte lui-même et doit en être extraite. Et les techniques qui ont donné des résultats intéressants en IA sur de petits textes bien maîtrisés (scénario de restaurant, etc.) s'avèrent vite problématiques lorsqu'elles sont appliquées à des domaines dont on ignore en partie ou en totalité la teneur. Un texte contient normalement de nombreux énoncés

originaux qui n'ont pas encore été lus et dont le contenu tant lexical, sémantique qu'encyclopédique est inconnu au préalable par le lecteur, et qu'il découvrira dans le parcours du texte lui-même.

Le deuxième problème est de nature plus technique. Même si on possédait des analyseurs linguistiques raffinés et robustes pouvant décrire un texte selon ses diverses catégories linguistiques (morphologiques, syntaxiques, sémantiques, discursives) il faudrait prévoir que ce traitement prenne un certain temps. Dans la meilleure des situations, la technologie actuelle ne permet guère d'analyser des phrases en deçà de quelques 10 à 20 secondes par phrase. On peut imaginer le temps requis pour traiter des milliers de pages. La situation de LATAO ne permet pas ce type de traitement. Il faut modifier l'approche. Des stratégies, peut-être plus grossières dans leurs approches premières, permettent ultimement des extractions fines de connaissances. C'est dans cette perspective que nous explorons les approches par classification numérique et plus particulièrement les classifieurs de type connexionniste. Il nous semble que, dans le traitement de grande masse d'informations, il faut y aller comme en archéologie. Un bon archéologue, ne commence pas directement sa fouille par le plus fin et le plus précis de ses outils. Au contraire, il commence sa recherche par un parcours général de son territoire. Il utilise pour ce faire des outils généraux (sonar, résonance magnétique, géomatique, etc.). Ce n'est qu'une fois qu'il a cerné le lieu potentiel des vestiges archéologiques qu'il en appelle à des outils plus fins. La pelle, la cuillère, la brosse, etc. Et ce n'est qu'à la fin qu'il prendra son microscope électronique. En d'autres termes deux grandes étapes sont nécessaires, une première étape utilisant un outil que nous dirons `bulldozer` pour classifier d'une manière grossière les données textuelles et ainsi permettre à un utilisateur de sélectionner dans une deuxième étape les parties du texte sur lesquels il veut extraire des connaissances d'une manière plus fine et ce au moyen de méthodes linguistiques.

2. Stratégies numériques :

La littérature technique relative au traitement de l'information textuelle a montré qu'il était possible d'explorer des outils d'extraction des connaissances dans des textes (data mining). Or, l'extraction de connaissances peut être vue sous plusieurs angles. Dans notre perspective, elle n'est pas une "compréhension" du texte, ni une paraphrase, ni un rappel d'information, mais un processus de traitement classificatoire qui identifie des segments de textes qui contiennent un "même" type d'information. Autrement dit, l'extraction des connaissances est définie comme résultant d'une opération de classification fondée sur l'un ou l'autre critère d'équivalence.
Pour les chercheurs dans le domaine de LATAO, cette problématique n'est pas nouvelle. Dans la recherche antérieure, plusieurs techniques et méthodes ont déjà été proposées pour tenter d'organiser le contenu d'un texte en des configurations interprétables. Ces méthodes, souvent moins fines certes que les approches linguistiques et conceptuelles n'en permettent pas moins un premier parcours général et robuste du texte. Elles sont en mesure, par exemple, d'identifier dans un corpus des classes ou des groupes de lexèmes qui entretiennent entre eux des associations dites de cooccurrence et donc de détecter leurs réseaux sémantiques. Et les recherches actuelles commencent d'ailleurs à les privilégier de plus en plus (Church 1989, 1991, Reinheirt 1994, Salem et Lebart 1994, Pustejovski 1995, Wilks 1996, Salton 1989 etc.). Parmi les modèles les plus couramment utilisés, on trouve habituellement l'analyse des cooccurrences, l'analyse corrélationnelle, l'analyse en composante principale, l'analyse en groupe, l'analyse factorielle, l'analyse discriminante, etc. Malgré le succès qu'elles ont obtenu, on a dû constater que ces méthodes particulières posent deux problèmes importants. Premièrement, les modèles classiques ne peuvent traiter que des corpus stables. Toute modification du corpus exige une reprise de l'analyse numérique. Ceci devient un problème majeur dans des situations où le corpus est en constante modification (par exemple les reposoirs de l'autoroute électronique). Deuxièmement, les types de résultats qu'ils produisent ne sont pas sans problèmes théoriques. Ils posent des problèmes d'interprétation linguistique importants (Church, 1990). Les associations des mots dans les classes ne sont pas toujours facilement interprétables. Pourtant, malgré leurs limites, ces approches ont été reconnues des plus utiles pour l'extraction des connaissances et plus particulièrement les connaissances terminologiques. D'une part, ces stratégies classificatoires permettent une immense économie de temps dans le parcours exploratoire d'un corpus, et à ce titre, elles sont incontournables lorsqu'on est confronté à de vastes corpus textuels. D'autre part, elles servent d'indices pour détecter rapidement certains liens sémantiques et textuels. Cependant, lorsqu'associées ^ des stratégies linguistiques plus fines et intégrées dans des systèmes hybrides (i. e., avec analyseurs linguistiques d'appoint), elles livrent une assistance précieuse pour des analyses globales. Elles permettent un premier déblaiement général du texte. Peuvent alors suivre des analyses plus fines.

Les recherches récentes permettent de penser qu'on peut améliorer ces techniques de classification de l'information. En effet, de nouveaux modèles classifieurs dits émergentistes commencent à être explorés pour ce type de tâche. Ils ont pour fondement théorique que le traitement "intelligent" de l'information est avant tout associatif et surtout adaptatif. Parmi ces modèles dits "de computation émergente" on distingue les modèles "génétiques" (Holland 1973), markoviens (R. Kindermann et L. Snell, 1980; Bouchaffra et Meunier, 1993) et surtout connexionnistes. Parmi ces derniers, on trouve une grande variété de modèles, entre autres, les modèles matriciels linéaires et non linéaires (Anderson, Silverstein, Ritz et Jones, 1977; Kohonen, 1989; Murdock, 1982), les modèles thermodynamiques (Hinton et Sejnowski, 1986), et les modèles basés tantôt sur la compétition, tantôt sur la rétropropagation, mais surtout sur des règles complexes d'activation et d'apprentissage (Kohonen, 1989 ; Rumelhart et McClelland, 1986). Les principaux avantages de ces modèles tiennent au fait que leur structure parallèle leur permet de satisfaire un ensemble de contraintes qui peuvent être faibles et même, dans certains cas, contradictoires et de généraliser leur comportement à des situations nouvelles (le filtrage), de détecter des régularités et ce, même en présence de bruit (Reggia et Sutton, 1990). Outre les propriétés de généralisation et de robustesse, la possibilité pour ces modèles de répondre par un état stable à un ensemble d'inputs variables repose sur une capacité interne de classification de l'information.
Cependant, tous ces modèles classifieurs émergentistes opèrent sur des données bien contrôlés et qui toutes doivent être présentes au début et tout au long du traitement. De plus, ils exigent souvent divers paramètres d'ajustement qui relèvent souvent d'une description statistique du domaine. Il s'en suit que les résultats de classification obtenus sont valides pour autant qu'ils portent sur les données bien contrôlées où peu de modification sont possibles. Si, après la période d'apprentissage, pour quelque raison que ce soit, les systèmes sont confrontés à des données qui n'étaient pas prévues dans les données de départ, ils auront tendance à les classer dans les prototypes déjà construits, donc à produire une sous-classification.
Or, le domaine dans lequel nous opérons, à savoir le texte, présente précisément ce type de problème. Chaque nouvelle page peut contenir des informations que le système peut ne jamais avoir rencontrées, et donc qu'il ne peut se permettre de classer dans ses prototypes antérieurement construits. Il faut donc, outre la dynamicité de l'apprentissage, un système qui soit aussi plastique.

3. La Grammaire Catégorielle Combinatoire Applicative dans le cadre de la Grammaire Applicative et Cognitive :

La Grammaire Applicative et Cognitive (Desclés, 1990) postule trois niveaux de description des langues :
a- le niveau phénotypique (ou le phénotype) où sont représentées les caractéristiques particulières des langues naturelles (par exemple l'ordre des mots, les cas morphologiques, etc....). Les expressions linguistiques de ce niveau sont des unités linguistiques concaténées, la concaténation est notée par : 'u_1-u_2-...-u_n'.
b- le niveau génotypique (ou le génotype) où sont exprimés les invariants grammaticaux et les structures sous-jacentes aux énoncés du niveau phénotypique. Le niveau génotypique est structuré comme un langage formel appelé "Langage génotype" ; il est décrit par une grammaire appelée "Grammaire applicative".
c- le niveau cognitif où sont représentées les significations des prédicats lexicaux par des schèmes sémantico cognitifs.
Les trois niveaux font appel à des formalismes applicatifs typés où l'opération d'application d'un opérateur à un opérande est considérée comme primitive. Les niveaux deux et trois s'expriment dans le formalisme de la logique combinatoire typée de H.B. Curry (1958). Cette logique fait appel à des opérateurs abstraits - appelés "combinateurs" - qui permettent de composer intrinsèquement des opérateurs plus élémentaires entre eux (Desclés, 1990). Les combinateurs sont associés à des règles d'introduction et d'élimination. Ceux que nous utiliserons dans cet article [1] sont **B**, $\mathbf{C}_*$, avec les règles d'élimination (β-réduction) suivantes (U_1, U_2, U_3 sont des expressions applicatives typées) :

$((\mathbf{B}\ U_1\ U_2)\ U_3) > (U_1\ (U_2\ U_3))$
$((\mathbf{C}_*\ U_1)\ U_2) > (U_2\ U_1)$

[1] Le combinateur $\mathbf{C}_*$ est souvent noté T.

Le modèle de la Grammaire Catégorielle Combinatoire Applicative (GCCA) relie explicitement les expressions phénotypiques à leurs représentations sous-jacente dans le génotype[2]. Le système consiste en :

(i) une analyse syntaxique des expressions concaténées du phénotype par une Grammaire Catégorielle Combinatoire.
(ii) une construction à partir du résultat de l'analyse syntaxique d'une interprétation sémantique fonctionnelle des expressions phénotypiques.

Les Grammaires Catégorielles assignent des catégories syntaxiques à chaque unité linguistique. Les catégories syntaxiques sont des types orientés engendrés à partir de types de base et de deux opérateurs constructifs '/' et '\'.

(i) N (syntagme nominal) et S (phrase) sont des types de base.
(ii) Si X et Y sont des types orientés alors X/Y et X\Y sont des types orientés[3].
Une unité linguistique u de type orienté X sera désigné par '[X : u]'.
Les deux règles d'application (avant et arrière) sont notées:

$$\frac{[X/Y : u_1] \quad [Y : u_2]}{[X : (u_1\ u_2)]} > \qquad ; \qquad \frac{[Y : u_1] \quad [X\backslash Y : u_2]}{[X : (u_2\ u_1)]} <$$

Les prémisses dans chaque règle sont des concaténations d'unités linguistiques à types orientés considérées comme étant des opérateurs ou des opérandes, la conséquence de chaque règle est une expression applicative avec un type orienté.
La Grammaire Catégorielle Combinatoire (Steedman, 1989) généralise les Grammaires Catégorielles classiques en introduisant des opérations de changement de type et des opérations de composition des types fonctionnels. Dans la GCCA les règles de la Grammaire Catégorielle Combinatoire de Steedman introduisent les combinateurs **B**, $\mathbf{C_*}$ dans la séquence syntagmatique. Cette introduction permet de passer d'une structure concaténée à une structure applicative. Les règles de la GCCA sont :

Règles de changement de type :

$$\frac{[X : u]}{[Y/(Y\backslash X) : (\mathbf{C_*}\ u)]}\ \mathbf{T} \qquad ; \qquad \frac{[X : u]}{[Y\backslash(Y/X) : (\mathbf{C_*}\ u)]}\ \mathbf{T}<$$

$$\frac{[X : u]}{[Y/(Y/X) : (\mathbf{C_*}\ u)]}\ \mathbf{Tx} \qquad ; \qquad \frac{[X : u]}{[Y\backslash(Y\backslash X) : (\mathbf{C_*}\ u)]}\ \mathbf{Tx}<$$

Règles de composition fonctionnelle :

$$\frac{[X/Y : u_1]\text{-}[Y/Z : u_2]}{[X/Z : (\mathbf{B}\ u_1\ u_2)]}\ \mathbf{B} \qquad ; \qquad \frac{[Y\backslash Z : u_1]\text{-}[X\backslash Y : u_2]}{[X\backslash Z : (\mathbf{B}\ u_2\ u_1\ )]}\ \mathbf{B}<$$

$$\frac{[X/Y : u_1]\text{-}[Y\backslash Z : u_2]}{[X\backslash Z : (\mathbf{B}\ u_1\ u_2)]}\ \mathbf{Bx} \qquad ; \qquad \frac{[Y/Z : u_1]\text{-}[X\backslash Y : u_2]}{[X/Z : (\mathbf{B}\ u_2\ u_1)]}\ \mathbf{Bx}<$$

Les prémisses des règles sont des expressions concaténées typées ; les résultats sont des expressions applicatives (typées) avec éventuellement introduction d'un combinateur. Le

[2]Dans le phénotype, les expressions linguistiques sont concaténées selon les règles syntagmatiques propre à la langue. Dans le génotype, les expressions sont agencées selon l'ordre applicatif.
[3]Nous choisissons ici la notation de Steedman (1989) : X/Y et X\Y sont des types orientés fonctionnels. Une unité linguistique 'u' avec le type X/Y (respectivement X\Y) est considérée comme un opérateur (ou une fonction) dont l'opérande de type Y est positionné ^ droite (respectivement ^ gauche) de l'opérateur.

changement de type d'une unité u introduit le combinateur $\mathbf{C}_{*}$; la composition de deux unités concaténées introduit le combinateur **B**.

Pour l'exemple suivant *La liberté renforce la démocratie* nous avons l'analyse suivante :

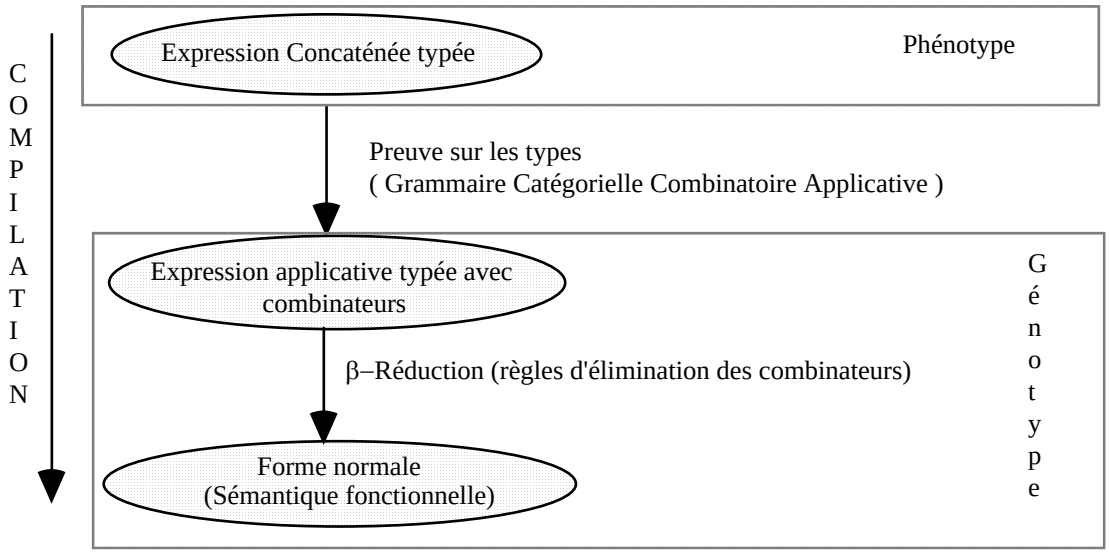
1. [N/N : *la*]-[N : *liberté*]-[(S\N)/N : *renforce*]-[N/N : *la*]-[N : *démocratie*]
2. [N : (*la liberté*)]-[(S\N)/N : *renforce*]-[N/N : *la*]-[N : *démocratie*] **(>)**
3. [S/(S\N) : (**C*** (*la liberté*))]-[(S\N)/N : *renforce*]-[N/N : *la*]-[N : *démocratie*] **(>T)**
4. [S/N : (**B** (**C*** (*la liberté)) renforce)*]-[N/N : *la*]-[N : *démocratie*] **(>B)**
5. [S/N : (**B** (**B** (**C*** (*la liberté)) renforce) la)*]-[N : *démocratie*] **(>B)**
6. [S : ((**B** (**B** (**C*** (*la liberté)) renforce) la) démocratie*)] **(>)**

7. [S : ((**B** (**B** (**C*** (*la liberté)) renforce) la) démocratie*)]
8. [S : ((**B** (**C*** (*la liberté)) renforce)* (*la démocratie*))] **B**
9. [S : ((**C*** (*la liberté))* (*renforce* (*la démocratie*)))] **B**
10. [S : ((*renforce* (*la démocratie*)) (*la liberté*)))] **C***
11. [S : *renforce* (*la démocratie*) (*la liberté*)]

Ainsi pour cet exemple, à l'étape 1 des types catégoriels sont assignés aux unités linguistiques. À l'étape 2, la règle (>) est appliqué aux unités linguistiques *la* et *liberté.* À l'étape 3 une règle de changement de type (**>T**) est déclenchée pour construire un opérateur (**C*** (*la liberté*)) à partir de l'opérande (*la liberté*). Cet opérateur est composé avec l'opérateur *renforce* à l'étape 4 par une opération de composition (**>B**) de façon à former un opérateur complexe (**B** (**C*** (*la liberté)) renforce)*. Deux autres opérations respectivement (**>B**) et (>) suivent.
À l'étape 7 commence la réduction des combinateurs. Cette série de réduction se fait dans le génotype.
À l'étape 11 nous produisons la structure prédicative dont nous aurons besoin pour aider le terminologue.

Un traitement complet basé sur la Grammaire Catégorielle Combinatoire Applicative s'effectue en deux grandes étapes :
(i) la première étape s'illustre par la vérification de la bonne connexion syntaxique et la construction de structures prédicatives avec des combinateurs introduits à certaines positions de la chaîne syntagmatique,
(ii) la deuxième étape consiste à utiliser les règles de β-réduction des combinateurs de façon à former une structure prédicative sous-jacente à l'expression phénotypique. L'expression obtenue est applicative et appartient au langage génotype. La GCCA engendre des processus qui associent une structure applicative à une expression concaténée du phénotype. Il nous reste à éliminer les combinateurs de l'expression obtenue de façon à construire la "forme normale" (au sens technique de la β-réduction) qui exprime l'interprétation sémantique fonctionnelle. Ce calcul s'effectue entièrement dans le génotype. Les formes applicatives obtenus en bout de parcours seront retenues pour être stockées dans des bases de données à des fins d'aide terminologique à l'utilisateur.

Le traitement fondé sur la GCCA prend la forme d'une compilation dont les étapes sont résumées dans la figure 1 :



4. Le modèle Hybride :

Dans sa forme concrète le modèle hybride que nous proposons consiste en deux grandes étapes :
- Un filtrage numérique grossier du corpus qui permet de classifier et de structurer le corpus en des classes de termes qui serviront d'indices de régularités d'associations lexicales que le terminologue peut utiliser comme tremplin pour approfondir les étapes ultérieures d'interprétation, de construction de réseaux sémantiques, et finalement d'élaboration de ses fiches terminologiques. Une plate-forme réalisée au LANCI, en l'occurrence, la plate-forme ALADIN (Meunier, J.G., Seffah, A., 1995) permet d'exécuter une chaîne de traitement qui réalise un tel filtrage. La chaîne présente les étapes suivantes : elle commence par une gestion du document, suit alors une description morphologique (lemmatisation) et une transformation matricielle du corpus. Vient ensuite une extraction classificatoire par réseaux de neurones FUZZYART. Ainsi dans la première étape de sa gestion, le texte est reçu et traité par des modules d'analyse de la plate-forme ALADIN-TEXTE. Cette plate-forme est un atelier qui utilise des modules spécialisés dans l'analyse d'un texte. Dans un premier temps, un filtrage sur le lexique du texte est fait. Par divers critères de discrimination, on élimine du texte certains mots accessoires (mots fonctionnels ou statistiquement insignifiants, etc.) ou ceux qui ne sont pas porteurs de sens d'un point de vue strictement sémantique, et dont la présence pourrait nuire au processus de catégorisation, soit parce qu'ils alourdiraient indûment la représentation matricielle, soit parce que leur présence nuirait au processus interprétatif qui suit la tâche de catégorisation. Vient ensuite une description morphologique minimale de type lemmatisation.
Puis une transformation est opérée pour obtenir une représentation matricielle du texte. Cette transformation est encore effectuée par des modules d'ALADIN explicitement dédiés à cette fin. On produit ainsi un fichier indiquant pour tout lemme choisi sa fréquence dans chaque segment du texte. Suit ensuite un post-traitement pour construire une matrice dans un format acceptable par le réseau de neurone FUZZYART[4]
Le réseau neuronal génère une matrice de résultats qui représentent la classification trouvée. Chaque ligne (ou vecteur) de cette matrice est constituée d'éléments binaires ordonnés. La ligne indique pour chaque terme du lexique original s'il fait ou non partie du prototype de la classe. Ainsi est créé un "prototype" pour chacune des classes identifiées. On dira alors que la classe no. X est "caractérisée" par la présence d'un certain nombre de termes. Autrement dit, chaque classe identifie quels sont les termes qui se retrouvent dans les segments de textes qui présentent, selon le réseau de neurones une certaine similarité. Ainsi, les classes créées sont caractérisées, arbitrairement, par les termes qui sont présents également dans tous les segments du texte qui ont été "classifiés" dans une même classe.
Les résultats du réseau de neurones se présentent donc (avant interprétation) sous la forme d'une séquence de classes que l'on dira "caractérisées" par des termes donnés et incluant un certain nombre de segments.

[4] le réseau de neurones FUZZYART utilisé pour l'expérimentation d'ALADIN a été développé sur une plate-forme de programmation matricielle disponible sur le grand marché appelé MATLAB (Crespo & Savaria, 1995).

- Un traitement linguistique plus fin des segments sélectionnés selon les thèmes choisis par le terminologue. Ce dernier sélectionne donc des segments dont il veut une analyse plus fine et en extraire une représentation des connaissances plus structurée. Le terminologue peut décider pour un segment donné de focaliser son attention sur un terme donné et en construire son réseau sémantique. La Grammaire Catégorielle Combinatoire Applicative peut organiser les phrases dans lesquelles apparaît le terme, choisi par le terminologue, sous forme de structures prédicatives `Prédicat argument1 argument2... argumentn`. Ainsi pour une sélection de phrases on peut engendrer une liste d'expressions prédicatives. Nous pouvons avoir dans cette liste des structures prédicatives ayant des arguments en commun.
Par exemple le cas suivant :
Prédicat1 argument1 argument2
Prédicat2 argument3 argument1
et là un terminologue comprendrait la relation sémantique entre les arguments 2 et 3 par rapport à l'argument 1.

Nous pouvons conserver une liste d'expression de la forme `Prédicat argument1 argument2... argumentn` dans une base de données.
Le terminologue la consultant peut déduire le sens sémantique de chaque argument (donc terme) en ce sens qu'il peut en construire le réseau sémantique.

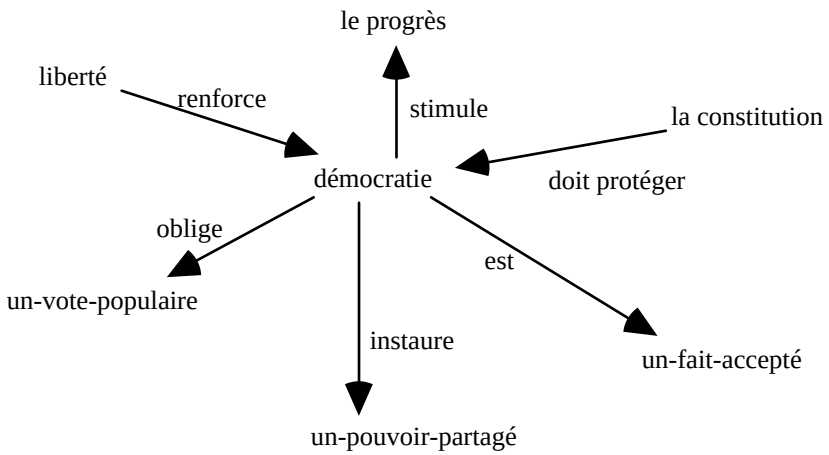
Prenons les phrases suivantes :

La liberté renforce la démocratie.
La démocratie stimule le progrès.
La démocratie oblige un vote populaire.
La démocratie est un fait accepté.
La constitution doit protéger la démocratie.
La démocratie instaure un pouvoir partagé.

Les structures prédicatives de ces phrases obtenues par une analyse linguistique catégorielle sont respectivement :

renforce (la démocratie) (la liberté)
stimule (le progrès) (la démocratie)
oblige (un-vote-populaire) (la démocratie)
est (un-fait-accepté) (la démocratie)
(doit protéger) (la démocratie) (la constitution)
instaure (un-pouvoir-partagé) (la démocratie)

un terminologue ayant dans sa base de données ces deux structures prédicatives pourra représenter le réseau sémantique de (la démocratie).



Remarque :
De tels exemples cependant ne mettent pas en évidence la pertinence de l'approche par classifieurs numériques car il n'y a pas d'ambiguïté dans les exemples. Le mot pole *démocratie* a un sens unique.

Prenons le mot pole *ferme*

Nous pouvons avoir trois groupes possibles :

Groupe 1
Une démocratie ferme est l'espoir d'un peuple.
Un vote ferme est passé à l'assemblée.
C'est une position ferme du gouvernement.

Groupe 2
Les paysans ont occupé les fermes et les villages.
Toutes les fermes ont abandonné leur récolte.
Les fermes sont laissées à l'abandon.

Groupe 3
Le vote ferme la discussion.
La décision du gouvernement ferme les options.
Le président de l'assemblée ferme le vote.

C'est le traitement classifieur qui a permis de séparer les champs lexicaux en trois groupes. Mais c'est au moyen de la grammaire catégorielle combinatoire applicative qu'une analyse en profondeur sera faite.

5. Conclusion :

Nous venons de présenter un modèle d'extraction des connaissances pour terminologues. L'idée d'associer des modèles linguistiques à des modèles numériques est très prometteuse. Elle est également très pertinente, en ce sens qu'elle associe la finesse d'analyse des méthodes linguistiques à la capacité des méthodes numériques d'absorber de gros corpus. L'ordre stratégique d'appliquer une méthode numérique avant de faire intervenir une méthode linguistique résulte du compromis nécessaire pour faire `cohabiter` ces deux approches. En effet, la méthode numérique est plus à même de `débroussailler` un gros texte et de permettre à un terminologue de soumettre des segments choisis à l'analyseur linguistique plus fin.
Ce type d'approche permet de solutionner une des critiques importantes qu'on fait classiquement aux approches de cooccurrence et collocation : leurs difficultés d'interprétation. Notre approche permet d'entrevoir des outils de raffinement de l'analyse des résultats livrés par les approches numériques trop grossières et générales. Il y a compensation entre les deux. Les Grammaires Catégorielles sont trop fines et donc trop lentes sur un corpus ample. Mais bien placées elles ne travaillent que sur des sous-corpus qui ont effectué un premier travail de désambiguisation.
En fait cette approche oblige a effectuer la désambiguisation dans un traitement différent de celui de la grammaire. La désambiguisation joue sur la différentialité des contextes de l'ensemble d'un corpus (relation paradigmatique) alors que l'analyse catégorielle opère sur la dépendance des contextes immédiats (relation syntagmatique). Ainsi , le système n'est pas obligé de faire les deux analyses en même temps ou dans une même passe ce qui, pensons nous, le rend plus efficace.
Enfin la configuration des résultats telle que permise par les expressions prédicatives permet au lecteur analyste d'avoir une organisation plus limpide sur le plan ergonomique et cognitif. On envisage la possibilité de livre ces résultats dans des réseaux structurés mais aussi dans des reposoirs structurés genre dictionnaires, thesaurus, bases de connaissances, etc.

Bibliographie

Biskri, I., 1995, La Grammaire Catégorielle Combinatoire Applicative dans le cadre de la Grammaire Applicative et Cognitive, Thèse de Doctorat, EHESS, Paris.

Biskri, I., Descles, J.P., 1995, "Applicative and Combinatory Catégorial Grammar (from syntax to functional semantics)", Acte du *Colloque RANLP, Bulgarie 1995*.
Burr, D. J. (1987). "Experiments with a connectionnist text reader". IEEE First International Conference on Neural Networks, San Diego, 717-24
Carpenter, G. , & Grossberg, G. (1991). "An Adaptive resonnance Algorithm for Rapid Category Learning and Recognition". Neural Networks 4, 493-504.
Cheeseman, P. , Self, M. , Kelly, J. , StutzJ, Taylor, W. , & Freeman, D. (1988). "Bayesian Classification". Proceedings of AAAI 88, Minneapolis, 607 -611
Curry, B. H., Feys, R., 1958, *Combinatory logic* , Vol. I, North-Holland.
Delany, & P. Landow (Ed.), The Digital Word: Text Based Computing in the Humanities. Cambridge, Mass: MIT Press.
Delisle, S. (1994). Text Processing without a priori domain knowledge: semi automatic linguistic analysis for incremental knowledge acquisition. PH Thesis, Ottawa University. :
Descles, J. P., 1990 *Langages applicatifs, langues naturelles et cognition*, Hermes, Paris.
Descles, J.P., Biskri, I., 1996, "Logique combinatoire et linguistique : Grammaire Catégorielle Combinatoire Applicative" Revue Mathématiques, Informatiques et Sciences Humaines. Paris.
Frey, S. , Reyle, U. , & Rohrer, C. (1983). "Automatic Construction of a Knowledge Base by Analysing Texts in Natural Language". Pro. c of IJCAI, 83 727-729,
Garnham, A. (1981). "Mental models and representation of texts". Memory and Cognition 9 (560-565),
Grefenstette, G. (1992). "Sextant: Exploring Unexplored Contexts for Semantic Extraction from Syntactic Analysis". Proc of the 30th Annual Meeting fo the ACL 324- 326,
Grefenstette, G. (1992). "Use of syntactic Context to Produce Term Association Lists for Text Retrieval". Proc of SIGIR 92 ACM, Copenhagen, june 21-24,
Grossberg, S. , & Carpenter, S. (1987). "Self Organization of Stable Category Recognition Codes for Analog Input Patterns". Applied Optics 26, 4919- 4930.
Jacobs, P. , & Zernik. U. (1988). "Acquiring Lexical Knowledge from Text A case Study". Proceedings of AAA1 88 (St Paul. Min.),
Jansen, S. , Olesen, J. , Prebensen, H. , & Tharne, T. (1992). Computanional approaches to text Understanding. in Copenhaguen: Museum Tuscalanum Press,
Jouis, C., 1993, *Contributions ˆ la conceptualisation et ˆ la modélisation des connaissances ˆ partir d'une analyse linguistique de textes. Réalisation d'un prototype : le système SEEK*, Thèse de doctorat, EHESS, Paris 1993.
Kahonen, T. (1982). "Clustering, taxonomy and topological Maps of Patterns". IEEE Sixth International Conference on Pattern Recognition, 114-122
Lebart, L. , & Salem, A. (1988). Analyse statistique des données textuelles. Paris: Dunod.
Lin, X. , Soergel, D. , & Marchionini, G. (1991). "A Self Organizing Semantic Map for Information Retrieval". SIGIR 91, Chicago, Illinois,
Meunier,J.G (1996) Théorie cognitive:son impact sur le traitement de l'information textuelle.in V.Riale et D. *Fisette Penser L'esprit ,Des sciences de la cognition a une philosophie cognitive*. Presses de Université de Grenoble. 1996 289-305
Moulin B, & Rousseau, D. (1990). "Un outil pour l'acquisition des connaissances a partir de textes prescriptifs". ICO, Québec 3 (2), 108-120.
Recoczei, S. , & E. P. O, P. (1988). "Creating the Domain of Discourse: Ontology and Inventory". In J. &. B. G. Boose (Ed.), Knowledge Acquisition Tools for Experts and Novices. Academic Press:
Regoczei, S. , & Hirst, G. (1989). On extracting knowledge from Text. Modeling the Architecture of Language Users. (TR CSRI 225). Computer Systems Research Institute University of Toronto.
Salton, G. (1988). "On the Use of Spreading Activation". Communications of the ACM vol 31 (2),
Salton, G. , Allan, J. , & Buckley, C. (1994). "Automatic Stucturing and Retrieval of Large Text File". Communications of the ACM 37 (2), 97-107.
Shaumyan, S. K., 1987, *A Semiotic Theory of Natural Language*, Bloomington, Indiana Univ. Press.
Shaw, M. L. G. , & B. R. , G. (1988). "Knowledge Initiation and Transfer Tools for Expert ad Novices". In J. B. &. B. Gaines (Ed.), Knowledge Acquisition Tools for Expert Systems. Academic Press.
Steedman, M., 1989, *Work in progress: Combinators and grammars in natural language understanding*, Summer institute of linguistic, Tucson University.
Tapiero, I. (1993). Traitement cognitif du texte narratif et expositif et connexionnisme: expérimentations et simulations. in Université de Paris VIII,
Thrane, T. (1992). "Dynamic Text Comprehension". In J. O. S. Jansen H Prebensen, T. Thrane (Ed.), Copenhaguen: Museum Tuscalanum Press.

Veronnis, J. , Ide, N. M. , & Harie, S. (1990). "Utilisation de grands réseaux de neurones comme modèles de représentations sémantiques". Neuronimes,
Virbel, J. (1993). "Reading and Managing Texts on the Bibliothèque de France Stations". In P. Delany, & P. Landow (Ed.), The Digital Word: Text Based Computing in the Humanities. Cambridge, Mass: MIT Press.
Williams, M. (1990). " Connectionist Models and Information Retrieval". 25, 209-259.
Young, T. , & Calvert, T. (1987). Classification, Estimation, and Pattern Recognition. Amsterdam: Elsivier.
Zarri, G. P. (1990). "Représentation des connaissances pour effectuer des traitements inférentiels complexes sur des documents en langage naturel. ". In Office de la langue française (Ed. Les industries de la langue. Perspectives 1990. Gouvernement du Québec.