Exploration de classifieurs connexionnistes pour l'analyse de textes assistée par ordinateur

Jean-Guy Meunier, Ismaïl Biskri, Georges Nault,
Moses Nyongwa

* LANCI-UQAM
CP 8888, succursale Centre-Ville
Montréal (Québec) Canada H3C 3P8
Tel: (514) 987 3000 poste 0339
Meunier.Jean-Guy@uqam.ca
{biskri,nault,nyongwa}@pluton.lanci.uqam.ca

Résumé

L'extraction de divers types de connaissances à partir de bases de données textuelles pose des problèmes. Comment maintenir un processus d'extraction opérationnel malgré la quantité massive de textes à traiter en constante modification ? Une approche heuristique se trouve dans les méthodes classificatoires statistico-numériques. Nous avons explorer parmi celles-ci les classifieurs connexionnistes. Le domaine particulier d'application choisi pour cette expérience est l'identification de réseaux lexicaux susceptibles d'interprétation sémantique par un terminologue. Le corpus choisi est un texte légal de la Compagnie Hydro Québec et un numéro d'une revue en sciences de l'éducation (SPIRALE). Le traitement connexionniste est intégré à des processus linéaires rapides de pré-traitement et de post-traitement du texte. La chaîne de traitement ainsi produite allie dynamicité et rapidité pour fournir à l'usager terminologue une aide dans sa tâche d'extraction des connaissances.

1. Présentation générale :

Cet article présente un volet de la recherche et du développement effectués dans le cadre du projet Franco-Québécois intitulé "Les classifieurs émergentistes et le traitement de l'information", L'Objectif général de ce projet est d'explorer les approches classificatoires statistico numériques pour l'analyse de information se présentant en langue naturelle, texte, fiche, documents etc. (leur pertinence également) . Dans le présent travail, les classifieurs explorés ont été les réseaux de neurones, les analyses multidimensionnelles, les algorithmes génétiques, les champs de Markov. Concrètement ,le but visé par ces recherches est de

trouver une méthode efficace et économique d'appariement de tels types dedonnées informationnelles. Dans le cas du texte, il s'agit de formation de classes de segments textuels qui se ressemblent en raison d'un critère particulier choisi dans l'expérimentation. Ces classes peuvent alors soumises à d'autres analyseurs qui eux ouvrent la voie à des fonctions d'analyses plus complexes. Celles-ci deviennent alors des moteurs de fouille du type hypertextualisation automatique, extraction de connaissances, indexation, analyse terminologique etc. Dans le cas de fiches descriptives (genre : dossiers de patients), il s'agit aussi de formation de classes de fiches présentant des similarités , par exemple entre patients ou entre symptômes. Ces classes sont ensuite soumises à des analyses plus fines et intégrées dans des systèmes d'assistance dynamique aux diagnostics.

Un des résultats périphériques mais des plus prometteurs de cette collaboration est l'entente explicite d'un certain nombre de chercheurs pour collaborer à un protocole ou une plate-forme commune afin de permettre aux divers logiciels francophones utilisés dans ce projet d'entrer en interaction. Trop de logiciels français restent sur les tablettes faute de plate-forme pour les mettre en interaction. Le LANCI propose une plate-forme, ALADIN (Seffah et Meunier, 1995), qui constitue une première tentative, prototypale, pour mettre concrètement en interaction ces logiciels. Les laboratoires LANCI et TIMC-IMAG travaillent dans le cadre du présent projet ‡ la réalisation de cet objectif.Dans cet article, nous présentons brièvement l'exploration et expérimentation d'une approche neuronale pour l'analyse terminologie dans les grand corpus.

2. Collaboration franco-québécoise

Ce projet a été réalise dans le cadre d'un programme FRANCO QUEBECOIS , sous la co-tutelle du ministère de l'Enseignement supérieur et de la Recherche pour la partie française (M.E.N.E.S.R.I.P., Délégation à l'Information Scientifique et Technique, Programme Ingénierie linguistique et de la connaissance) et du ministère de la Recherche et du

ministère des Affaires Internationales, de l'Immigration et des communautés Culturelles pour la partie Québécoise.

Du coté français, ce projet s'est déployé en une collaboration portant sur l'assistance au diagnostique médical dynamique où participent Dr. Danel, du CHU de Grenoble -, V.Rialle (et son équipe) du TIMC IMAG de Grenoble.

Du coté Québécois, en raison de la dimension textuelle, ce projet a permis de dynamiser grandement l'intégration de l'équipe à plusieurs projets AUPELF UREF FRANCIL. sur les systèmes logiciels d'assistance a la terminologie (ARC A3). Ce qui s'est a son tour déployé en une collaboration intense et suivie avec l'université de Paris-Sorbonne (Laboratoire CAMS-LALIC : dir J.P Desclés) ainsi que l'IDIST de Lille 3.

3. Classifieurs neuronaux et assistance à la terminologie

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)

Extraction et classification.

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). 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 , Reinhert 1994, 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 particulière 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 à 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 souvent 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 du texte, nous sommes confrontés a des corpus en constante mdofication. Chaque nouvelle page peut possiblement 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 (incrémentialité) de l'apprentissage, un système qui soit aussi plastique.c'est-à-dire qui s'adapte à de nouvelles données. Et on voit apparaître depuis quelque temps des recherches qui sont de plus en plus sensibles a cette dimension (Burr 1987 Veronis 1990, Balpe et Lelu 1996, etc) Et c'est dans cette perspective que la présente recherche a été effectuée. Nous en présentons ici les résultats préliminaires à l'occasion d'une application d'analyse terminologique.

4. La méthode

Dans sa forme concrète et expérimentale, la recherche a consisté à explorer un modèle connexionniste pour extraire de l'information de type terminologique sur des fichiers textuels. La réalisation de cette recherche comporte les étape suivantes :

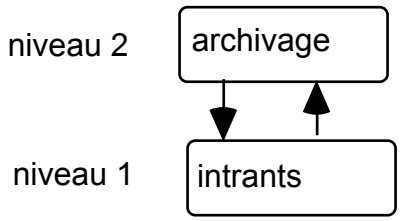
1- une modélisation d'un traitement connexionniste du texte,
2- une expérimentation d'une chaîne de traitement sur des textes.

3.1. Modélisation d'un traitement connexionniste de texte.

Il n'existe pas, à notre connaissance, un grand nombre de modèles connexionnistes pouvant confronter simultanément les problèmes de dynamicité et de plasticité de l'information. La dynamicité est la capacité du système à traiter de manière adaptative les informations qu'il reçoit. Quant à la plasticité il s'agit de la capacité du système à traiter de l'information pour laquelle il n'avait pas été paramétrisé. Un des modèles qui avait ces visées est ART 1 (Grossberg et Carpenter, 1988). En effet, dans sa définition originale le modèle ART 1 se veut un système classifieur auto-organisationnel, sans supervision, pouvant opérer sur des stimulis non contrôlés et bruités.

Ce modèle a évolué à travers les années. Il a donné naissance à de nombreuses variantes de la règle de transmission et de la règle d'apprentissage. Il a introduit des facteurs de multiplication dans l'activation et des facteurs de dégradation dans l'encodage de l'information. L'une de ses prétentions importantes est qu'il est en mesure de traiter de manière adaptative des stimulis qui sont changeants (plasticité) c'est-à-dire qui ne font pas partie d'un corpus contrôlé d'avance. L'objectif ultime de ce modèle est de créer une grande stabilisation dans la représentation des patrons de stimulis. Au début ART 1 ne pouvait traiter que des informations de nature binaires. Plus tard, le modèle ART 2 accepte des informations dont les valeurs ne sont plus discrètes ou binaires. Enfin dans ART 3 le modèle a consolidé ses stratégies et offre un traitement plus fiable.

L'idée centrale du modèle ART est celle d'un système d'interaction entre deux niveaux qui entrent en résonance mutuelle.



Le système reçoit dans un premier niveau 1 des stimulis qui sont envoyés, mais aussi modifiés (selon une distribution et un poids particulier) au deuxième niveau 2, qui est un niveau d'archivage.

Arrive donc au deuxième niveau un pattern différent de ce qui était à l'intrant. Il y a alors comparaison par un processus dit de résonance. Si le nouveau pattern n'a aucune ressemblance avec les anciens, il sera alors conservé et servira de gabarit ou de prototype avec lequel les intrants nouveaux seront éventuellement comparés. En fait, le pattern au niveau 2 servira de modèle de comparaison avec les nouveaux intrants. S'il diffère de ces derniers, un autre pattern sera essayé jusqu'à ce qu'une correspondance soit acceptable (selon certains paramètres) et s'il y a correspondance acceptable l'intrant sera alors classé avec le prototype. Mais s'il n'est pas acceptable, le nouveau pattern sera considéré comme un prototype en émergence et il servira éventuellement de nouveau gabarit aux autres intrants que le système rencontrera.

La correspondance entre le pattern prototypal et le pattern intrant est la résonance. Au fur et à mesure que l'apprentissage se poursuit, il y a consolidation de cette résonance. L'adaptabilité survient par la modification constante des interconnexions entre les niveaux. Pour réaliser cette interaction le système doit être contrôlé par divers paramètres qui assurent la solidité du traitement.

4.2. L'application du modèle ART

La deuxième étape de la recherche est une expérimentation de ce modèle sur des textes et sur une tâche spécifique. Un texte regorge de connaissances de plusieurs types qui peuvent en être extraits. On ne lit pas un texte uniquement pour savoir qui a fait quoi (connaissances du monde). On le lit pour connaître des faits, certes, mais aussi des actions, des valeurs, des jugements etc. (Meunier, 1996). Une des connaissances qu'on cherche est aussi de nature métalinguistique. i.e. qui porte sur la nature de la langue même du texte, tel par exemple,

son style, sa terminologie, son lexique, etc. Dans la présente recherche, nous avons choisi d'explorer l'extraction des connaissances métalinguistiques de type terminologique et sémantique. Pour ce faire nous avons développé une *chaîne de traitement* modulaire qui tente d'intégrer les étapes du travail analytique du terminologue ou du documentaliste devant travailler à l'identification des champs lexicaux d'un terme particulier. Par exemple, la châine de traitement pourrait assister le terminologue a identifier rapidement , que dans un texte, le mot CODE peut avoir des champs sémantiques différents où ce mot prend le sens de CODE civil, CODE de la route, CODE informatique, CODE de la construction, etc. Cette chaîne de traitement a été appliquée à deux textes spécifiques qui nous servent ici d'illustration: le premier de 900 pages, la Convention de la Baie James de l'Hydro Québec ; le second, la revue Spirale (Belgique) de quelques 180 pages.

L'expérimentation sur les textes

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. Cette opération consiste à remplacer chaque mot par son équivalent canonique. (e.g. aimerions --> AIMER). Ce processus se justifie par le fait que les déclinaisons propres à la grammaire ou à la syntaxe d'une langue n'affectent en rien le contenu sémantique réel des termes. De la même façon, remplacer un mot décliné (soit dans

sa forme verbale, adverbiale, adjectivale, pronominale ou autres) par sa forme nominale n'a aucun impact significatif sur le contenu sémantique principal de ce dernier. Ces dimensions morphologiques touchent surtout des modalités tels : le genre, l'aspect, le temps, etc.

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 les réseaux de neurones. Dans la présente expérimentation, la précédente matrice est alors soumise aux classifieurs ART . Selon le réseau utilisé (ART 1 ou FUZZY-ART), la matrice générée peut être constituée exclusivement de données binaires (ART 1) ou de données non-binaires (FUZZY-ART). Dans le cas du réseau ART 1, les données subissent alors une réduction, puisque la fréquence d'apparition des lemmes est alors remplacée par une simple indication de présence ou d'absence.

Les réseaux de type ART se prêtent particulièrement à ces contraintes. Le Fuzzy Art semble donner des résultats supérieurs car le processus de catégorisation produit moins de classes contenant un seul élément caractéristique (un seul lemme), ce qui du point de vue du terminologue n'est d'aucune utilité.

Les résultats de la classification neuronale.

Appliqué à la matrice précédente , ART génère des classes de SEGMENT qui présentent entre eux une certaine similarité lexicale. Autrement dit , chaque classe de segments constitue pour ART un "prototype" qui sont donc caractérisées,par les termes qui sont présents également dans tous les segments du texte. Une fois trouvé ces classes de segments présenant une similarité lexicale, on en extrait, pour chacune, le lexique .c'est -à-dire , on trouve pour les termes qui la caractérise.

Ensuite , on choisit un terme particulier , et on étudie les classes dans lesquels il apparaît. C'est ainsi par exemple que dans l'analyse sur la revue SPIRALE, nous avons obtenu la classification suivante pour les termes *Rapport* et *Tête*

Le terme rapport

apparaît dans les classes 28 35,39 40 54,

la classe 28: les segments 71-73

contient : choix connaissance, document , façon fiction personnage, rapport, savoir et travail

la classe 35 , les segments 89-92 qui contiennent les mots

autres connaissances doute formes image processus production et rapport

la classe 39: 100 101 qui contiennent les mots

autres élèves enseignant ensemble genre, jeunesse rapport roman

la classe 40 : 102 103 qui contiennent les mots

écrit, écriture élémentaires, jeunesse , monde problème rapport réel, scolaire, situation

la classe 54: 59 64 qui contiennent les mots

auteurs, autres, discours, jeunesse, lecture, mode, question, rapport, rôle, temps, vie

Sur la distribution de ce terme dans les classes de segments trouvées , on peut tenter l'interprétation sémantique suivante On peut dire que le terme *rapport* est utilisé dans deux ensembles de contextes relativement différents

Un premier pointe vers le concept de *rapport* comme *document* où est déposé de l'information (classe 28 dans les segments 71-73

Un deuxième pointe vers le concept des *liens entre des individus et autres chose* (39, 40, 54)

Enfin la classe 35 n'est pas clairement situable dans l'un ou l'autre sens précédent.

Cependant, on voit bien que dans le ce texte, ces deux significations sont les deux seules possibles. Pour un terminologue , le terme rapport dans la Revue Spirale # 15 n'est pas employé dans le sens suivants :d'une proportion c'est-à-dire d'un rapport logique, d'un rapport financier,d'une maison de rapport. d'une communication,ou encore d'une perspective, etc.

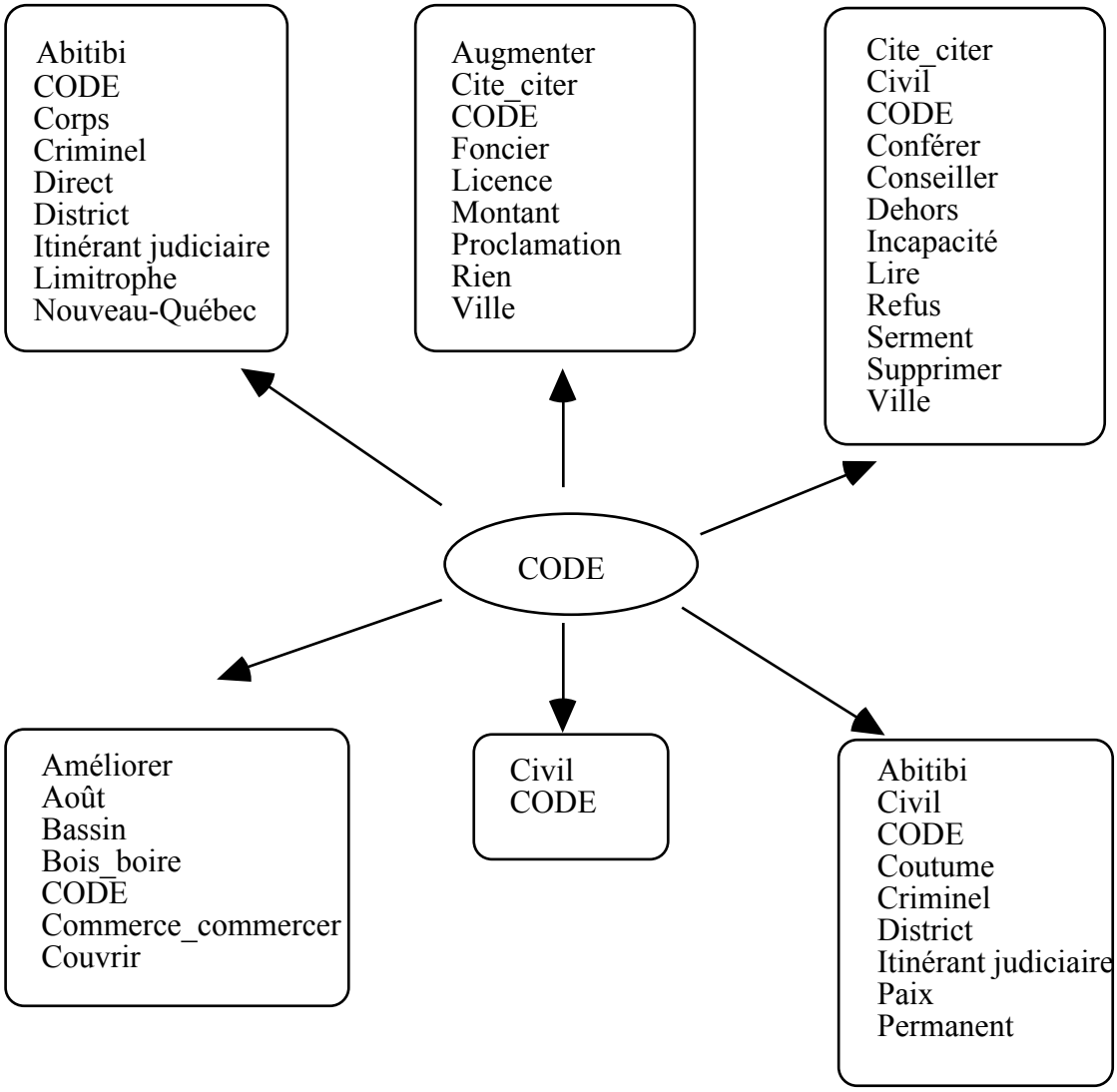
Une analyse similaire peut être tentée sur le terme Tête.
Nous retrouvons le terme Tête dans les classes dans la classe 20 et la classe 58.
La classe 20 : les segments 53 et 107.
façon, moment, place, rôle, savoir, tête.
La classe 58 : le segment 93
aide, bon, comprendre, connaissance, École, ensemble, histoires, jour, lecteur, loisirs, narratif, père, perspective, schéma, situation, souris, suite, tête, textuels, titre, traitements.

Ces deux classes montrent que le terme Tête est utilisé dans deux sens différents. Pour la classe 20 le sens du terme tête est proche de la signification “ leader ” or dans la classe 58 la signification est plus proche de l'interprétation “ intelligence ”.

Le *post-traitement*

La présentation linéaire de ces résultats rend l'interprétation difficile a réaliser. On peut imaginer une meilleure convivialité de la configuration des résultats. En effet, une légère transformation peur leur donner une lecture plus facile. A titre d'exemple nous présentons ci après la configuration d'un terme plus riche d'acception sémantique que nous avons trouvé dans le texte de l'Hydro Québec: La convention de la Baie James (870 pp) Dans ce texte,

le mot CODE se retrouve dans plusieurs classes. Il est alors possible pour ce lexème de dessiner le graphe des classes dans lesquels il apparaît.



On voit ainsi apparaître la différence des réseaux sémantiques de ce terme, qui tantôt est utilisé dans le sens de code civil, tantôt dans le sens de code de comportement, tantôt dans le sens du code criminel, etc. Pour un terminologue, ce graphe est intéressant. Il sert

d'indice au réseau sémantique de ce terme, et donc de ses acceptions dans le discours particulier. La thèse associative classique explique ce fait en postulant que si deux termes se retrouvent ensemble dans un même contexte, c'est que leurs contenus sémantiques ou conceptuels sont associés.

5. Interprétation et validation des résultats

Les résultats produits sont alors prêts à être validés. Une méthode de validation inter-juges a été utilisée pour évaluer la qualité et la pertinence de la classification opérée par le réseau de neurones. La méthode consiste essentiellement à comparer les évaluations faites et les appréciations effectuées par les différentes personnes impliquées dans l'évaluation des résultats. Le travail a consisté essentiellement en une analyse des classes produites pour déterminer leurs pertinences d'un point de vue terminologique. De façon générale, l'analyse a montré que plus de 80 % des classes constituées seraient pertinentes et utilisables dans le processus de création des fiches terminologiques. Mais un travail de validation sur un corpus banc d'essai est maintenant nécessaire.

6. Conclusion

L'objectif de notre recherche était de permettre l'extraction de connaissances terminologiques à partir du texte plein. Cette extraction devait, de plus, pouvoir se faire sur un corpus en évolution constante (plasticité) et les catégorisations effectuées se devaient de rester pertinentes et utilisables. Le processus de catégorisation quel qu'il soit se devait d'opérer sans supervision aucune (adaptabilité) et sans faire appel à des connaissances préformées ou prédigérées, celles-ci n'étant simplement pas disponibles dans le cas qui nous concerne.

Une méthode connexionniste comme solution au problème de l'extraction terminologique sur des textes entiers a été expérimentée avec des résultats très encourageants. L'approche

choisie montre un intérêt certain et des avantages indéniables. Par exemple, le gain de temps estimé par rapport au travail manuel requis pour effectuer un travail terminologique équivalent est considérable. De plus, la précision et la richesse des suggestions faites par le système ne sont en aucune mesure comparables avec ce qu'on obtient avec les méthodes actuelles.

L'approche choisie s'inscrit parfaitement dans l'optique d'une solution opérationnelle aux problèmes flagrants et réels qui minent l'industrie de la langue et toute industrie qui implique la manipulation et la classification de masses importantes de documents.

Plusieurs variantes de l'expérimentation sont possibles et envisagées. On pourra opérer les classes autrement que par l'intersection entre les segments. Par exemple tirer des informations pertinentes de l'union des unifs présents dans les segments regroupés. On pourrait sûrement obtenir des résultats encore plus probants en tenant compte de la dépendance entre les variables (ou unifs) pris en considération. Un pré-processing sémantique (non-approfondie pour conserver l'avantage de temps) pourrait amener des améliorations considérables. Plusieurs variantes des filtres de pré-traitement utilisés sont à l'étude. Le filtrage à la sortie des classes ne représentant pas une richesse d'information suffisante (selon un critère donné, ex. : pas assez d'unifs dans cette classe) améliorerait les résultats.

Plusieurs problèmes restent à résoudre (grandeurs fixes des intrants, dégradation du temps d'apprentissage avec le nombre des intrants, interprétation, etc.).

Un module d'interprétation des résultats (avec interface graphique) est en cours de développement.

6- Bibliographie

Balpe, J. P., Lelu, A., Papy, F., & I, S. (1996). *Techniques avancées pour l'hypertexte.* Paris : Hermes.

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.

Church, K., Gale, W., Hanks, P., & Hindle, D. (1989). "Word Associations and Typical Predicate-Argument Relations". *International Workhop on Parsing technologies,*Carnegie Mellon University, Aug. 28-31*,*

Church, K. W., & Hanks, P. (1990)."Word association norms, mutual informaton, and lexicography". *Computational Linguistics 16* , 22-29.

Delisle, S. (1994). Text Processing without a priori domain knowledge: semi automatic linguistic analysis for incremental knowledge acquisition. PH Thesis, Ottawa University. :

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.),

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.

Lelu, A. (1995). "Hypertextes: la voie de l'analyse des données". In L. Bolasco..S L ,A.Salem (Ed.), *Anilisi statistica dei dati testuali vol2. (pp.* 85-96). Rome: CISU.

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.

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.

Veronis, J. , Ide, N. M. , & Harie, S. (1990). "Utilisation de grands réseaux de neurones comme modèles de représentations sémantiques". Neuronimes,

Virbel, J. (1987)."L'apport de conniasances linguistiques à l'interprétation des structures textuelles". *Structure des documents, Bigre++Globule 53* , 77-97.

Virbel, J. E., F. Pascual, E. (1992). *La lecture assisfée par ordinateur,Raport de recherche.* Toulouse: Laboratoire IRIT.

Virbel, J. (1993). "Reading and Managing Texts on the Bibliothèque de France Stations". In P. 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.