Les Cahiers de l'ACFAS 97 . p 22 -30

Approche connexionniste au problème de l'extraction de connaissances terminologiques à partir de textes

Jean-Guy Meunier and Georges Nault
Laboratoire LANCI
Université du Québec à Montréal

Avril 1996

L'extraction de connaissances terminologiques sur des bases de données textuelles posent des problèmes. Comment maintenir un processus d'extraction opérationnel malgré la quantité massive de textes à traiter et qui sont en constante modifications. Nous proposons une approche connexionniste au problème. Cette approche permet d'appliquer les capacités classificatoires dynamiques des réseaux de neurones à un corpus textuel. Le domaine particulier d'application choisi pour cette expérience fut l'identification de réseaux lexicaux susceptibles d'interprétation sémantique par un terminologue. Le traitement connexionniste est intégré à des processus linéaires rapides de pré-traitement et post-traitement du texte. La chaîne de traitement ainsi produite allie dynamicité et rapidité pour fournir à l'usager terminologue un appui inestimable dans sa tâche d'extraction des connaissances.

Mots-clés : connexionnisme, connaissances terminologiques, textes, extraction de connaissance, etc.

Approche connexionniste au problème de l'extraction de connaissances terminologiques à partir de textes

Jean-Guy Meunier and Georges Nault
Laboratoire LANCI
Université du Québec à Montréal

Le problème

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. Rapidement, les tâches de dépistage, d'exploration et de récupération de l'information présente dans ces textes, c'est-à-dire des "connaissances", sont devenues extrêmement ardues, sinon impossibles. Un type particulier de ces connaissances que certains utilisateurs désirent extraire des textes est de nature terminologique. On veut connaître soit le vocabulaire propre du texte soit la signification de certains mots selon le contexte. Cependant, en raison de l'ampleur et de la dynamicité des corpus, cette extraction terminologique devient de moins en moins possible dans des temps raisonnables ou faisables avec des ressources restreintes.

Du point de vue méthodologique, cette question de l'extraction des connaissances dans les textes rencontre cependant des difficultés épistémologiques sérieuses. En effet, la situation est relativement inverse au cadre traditionnel de l'IA qui réclame pour le traitement de l'information que le système en appelle à des dépôts de connaissances préformées et 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 réussites. Les connaissances comporteraient des représentations d'objets, de propriétés, de relations d'événements et de situations propre à l'objet à traiter en l'occurrence le contenu informationnel du texte. En possession de ce savoir, un système expert pourrait alors réussir à "comprendre" le texte et donc en extraire les connaissances. De nombreuses recherches ont d'ailleurs montré la nécessité de 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.)

.

Pour un ordinateur, le problème de l'extraction des connaissances dans le cas d'un corpus textuel ne se pose pas en termes des connaissances qu'il doit posséder pour "comprendre " le texte mais en termes de ce qu'il lui faut faire pour les extraire du texte. Il est, en effet, délicat de donner à l'ordinateur, de manière a priori, les connaissances que le texte avait pour fonction de transmettre sauf peut-être pour celles qui sont de nature générales, encyclopédiques ou techniques. Dans le cadre classique de l'intelligence artificielle, il est possible de donner de telles connaissances mais pour autant qu'elles aient été acquises via des enquêtes cognitives (analyse de protocole) chez les experts ou puisées dans le répertoire encyclopédiques du savoir partagé.

Mais dans un horizon textuel, la connaissance se trouve dans le texte lui-même et doit en être extraite. Et les techniques qui ont donné des résultats intéressants 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 que des énoncés originaux qui n'ont pas encore été lus et dont le contenu tant lexical, sémantique et encyclopédique est inconnu du lecteur.

Extraction et classification.

La littérature technique relative au traitement de l'information textuelle a montré qu'il était devenu nécessaire d'explorer des outils d'extraction des connaissances dans des textes (data mining). Or, l'extraction de connaissances peut être vue de plusieurs manières. Dans notre perspective, elle n'est pas une " compréhension" du texte ni une paraphrase ni un rappel d'information mais une *identification qui classe des segments de textes* en regard d'une interprétation possible par un lecteur. Autrement dit, l'extraction des connaissances est définie comme résultant d'une opération de classification. Par ailleurs, cette classification ou extraction doit respecter les propriétés naturelles d'un texte. Il lui faut, en effet, respecter la dynamicité et la plasticité d'un texte c'est-à-dire il faut *classifier une information qui se modifie constamment tant dans sa quantité que dans son contenu.*

Pour les chercheurs dans le domaine de LATAO (lecture et analyse de textes assistés par ordinateur), 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 configuration interprétables.Outre les approches linguistiques et narratives, l'une des méthodes qui fut souvent explorée fut celle de l' analyse de données et de la reconnaissance de formes. Ces méthodes, moins fines certes que les approches linguistiques et conceptuelles permettent un premier traitement générale. 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. 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. (Salem et. Lebart 1992). Malgré le succès obtenu par ces méthodeson a dû constater que ces méthodes posaient 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'elles produisent ne sont pas sans problèmes théoriques. Elles posent des problèmes d'interprétation linguistique importants (Church, 1990). Ces associations 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. 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 approches émergentistes.

La recherche récente permet 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" ont trouvent ceux appelés "génétiques", (Holland 1973) markoviens (R. Kindermann and 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), de même que 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 une capacité interne de classification de l'information.

Cependant, tous ces modèles 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'ensuit que les résultats de classification obtenus sont valides pour autant qu'ils portent sur les données bien contrôlés où peu de modification sont possibles. Si, après la période d'apprentissage, pour quelque raison que ce soit, les systèmes sont 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 surclassification.

Or, le domaine dans lequel nous opérons, à savoir le texte, présente précisément ce type de problème. Chaque nouvelle page peut possiblement contenir des informations que le système peut ne jamais avoir rencontré et donc il ne peut se permettre de les classer dans ses prototypes qu'il a antérieurement construit. Il faut donc, outre la dynamicité de l'apprentissage, un système qui soit aussi plastique.

Le cadre théorique et l'hypothèse

Il n'existe pas, à notre connaissance, un grand nombre de modèles connexionniste pouvant confronter simultanément les problèmes de dynamicité et de plasticité de l'information. La dynamicité touche la capacité du système à traiter de manière adaptative les informations auquel il est confronté alors que la plasticité touche la capacité du système a traiter de l'information pour laquelle il n'avait été paramètrisé. Un des modèles qui avait ces visées est ART ONE (Grossberg

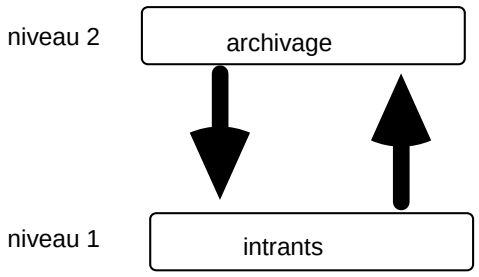
et Carpenter, 1988). En effet, dans sa définition originale le modèle ART ONE se veut un système classifieur auto-organisationnel, sans supervision pouvant opérer sur des stimuli non contrôlés et bruités.

> *" An Art system is capable of self stabilizing the self organization of their recognition codes in response to an arbitrarily complex environment of input pattern in a way that parsimoniously reconciles the requirements of plasticity, stability and complexity."* Carpenter et Grossberg*1988 : 254)*

Ce modèle a évolué a travers les années. Il a livré 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 adapatative des stimuli 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 stimuli. Au début ART-ONE ne pouvait traiter que des informations de nature binaires. Plus tard, ART 2, le modèle accepte de 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.

> *"An Art system can adaptively switch between a stable and plastic modes. It is capable of plasticity in order to learn about significant new events, yet it can also remain stable in response to irrelevants events, in order to make this distinction, an ART system is sensitive to novelty. It is capable,without a teacher, of distinguishing between familiar and unfamiliar events, as well as between expected and unexpected events." Grossberg 1988: 254*

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



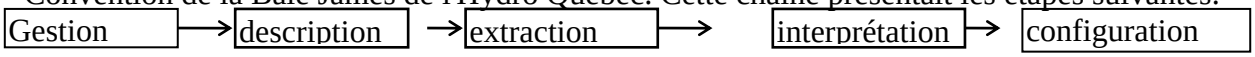
Le système reçoit en un premier niveau N1 des stimuli qui sont envoyés mais aussi modifiés (selon une distribution et un poids particulier) au deuxième niveau N2 qui est un niveau d'archivage.

Arrive donc au deuxième niveau un pattern différent de ce qui était à l'intrant. Ce pattern est alors conservé et servira de gabarit ou de prototype contre lequel les intrants seront éventuellement comparés. A ce moment, ce prototype sera retourné (selon les mêmes règles de sommation) au stimuli intrant.

En fait, le pattern en N2 servira d'hypothèse vis-à-vis lequel un intrant sera comparé. i.e. ce qui entre est-il vraiment différent ou semblable au pattern archivé? S'il est différent, un autre pattern sera essayé jusqu'à ce qu'une correspondance soit acceptable (selon des 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 une prototype en émergence et il servira éventuellement de nouveau gabarit aux autres intrants que le système rencontrera.
La résonance est définie comme la correspondance entre le pattern prototypal et le pattern intrant. Et au fur et à mesure que l'apprentissage se poursuit il y a consolidation de cette résonance. L'adaptativité survient par la modifications constante des interconnexions entre les niveaux.
Pour réaliser cette interaction le système doit être contrôlé par divers paramètres qui assurent la solidité du traitement.

Le traitement connexionniste du lexique d'un texte.

Le modèle Art a été utilisé dans une application d'extraction de connaissances terminologiques. La recherche servait a déterminer s'il était possible de structurer l'information lexicale dans un texte de manière à déterminer l'environnement sémantique d'un lexème particulier. Par exemple,quel est pour le mot CODE son ou ses usages sémantiques spécifiques? Autrement dit, les connaissances extraites consistent en des réseaux lexicaux susceptibles d'interprétation par un terminologue.
Notre chaîne de traitement était appliquée à un texte de quelques 900 pages. A savoir la Convention de la Baie James de l'Hydro Québec. Cette chaîne présentait les étapes suivantes:



Elle commence par une *gestion* du document, suit alors un *description* morphologique (lemmatisation) et matricielle du corpus. Vient ensuite une *extraction* classificatoire par réseaux de neurones ART ONE et, enfin, une *interprétation* des catégories par des tests inter-juges à des fins de validation. Une étape ultérieure à l'expérimentation s'ajoutera éventuellement pour opérer une mise (configuration) en forme hypertextuelle du plein texte à partir de la catégorisation effectuée par les réseaux de neurones. L'objectif global de la chaîne vise à développer un outil fonctionnel d'aide à l'analyse terminologique.

L'expérimentation

Dans la première étape de sa gestion,le texte est reçu et traité par un logiciel spécialisé dans l'analyse de contenu :SATO). A cette étape un filtrage sur le lexique du texte est appliqué. Via divers critères de discrimination SATO retire du texte certains mots accessoires (mots fonctionnels statistiquement non signifiants etc.) ou 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. (vg. aimerions-. AIMER) Ce processus se justifie en ce que les déclinaisons propres à la grammaire ou à la syntaxe d'une langue n'affecte en rien le contenu sémantique profond des termes. De la même façon, remplacer un mot décliné (soit dans sa forme verbale, adverbiale, adjective, pronominale ou autres) par exemple, sa forme nominale n'a aucun impact significatif sur le contenu sémantique principal de ce dernier Ces dimensions morphologique touchent surtout des aspects genres, aspects, temps etc.

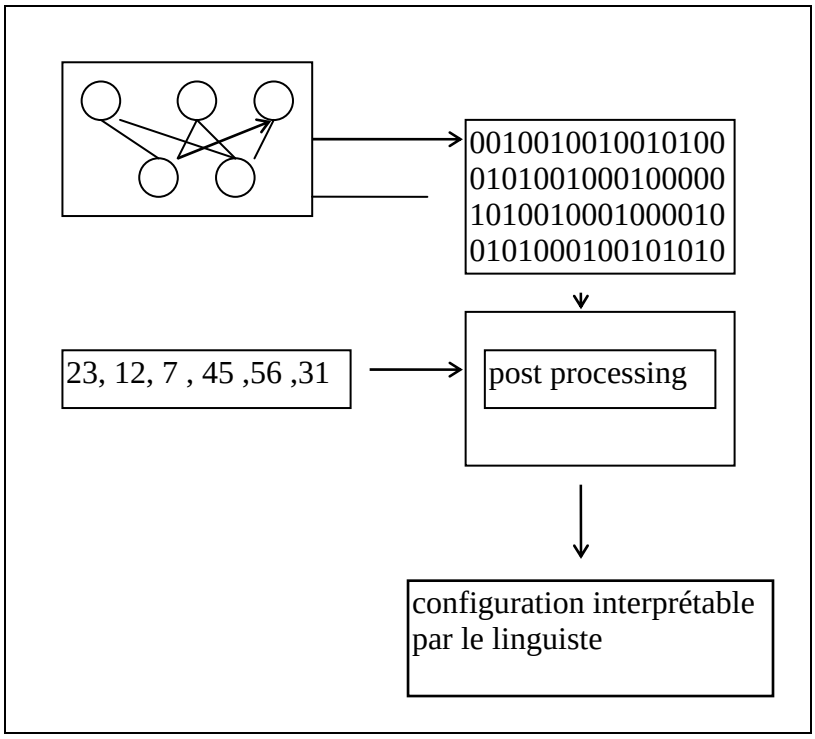
Par après une transformation est effectuée pour obtenir une représentation matricielle du texte. Cette transformation est opérée moitié par SATO et moitié par un logiciel développé explicitement à cette fin. SATO produit un fichier indiquant pour chaque lemme choisi sa fréquence dans chaque segment du texte. Un post-traitement est alors effectué pour construire une matrice dans un format acceptable pour les réseaux de neurones.

Selon le réseau utilisé (ART-ONE ou FUZZY-ART), la matrice générée pouvait être constituée exclusivement de données binaires (ART-ONE) ou de données non-binaires (FUZZY-ART). Dans le cas du réseau ART-ONE, 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.

La matrice générée représente la distribution des lemmes dans le texte. Pour chaque segment du texte (qu'il s'agisse de pages, de paragraphes, de chapitres, ou autres), la matrice contient une ligne dont chaque élément indique, pour un unif donné, la fréquence de son apparition dans le segment). Le choix du type de réseau de neurones utilisé repose sur plusieurs facteurs, principalement la nécessité d'opérer une classification dynamique, non-supervisée sur des intrants de longueurs fixes. Les réseaux de type ART se prêtent particulièrement à ces contraintes. Le second type de réseau semble donner des résultats supérieurs en ce que le processus de catégorisation produit moins de classes contenant une seul élément caractéristique (un seul lemme), ce qui du point de vue du terminologue n'est d'aucune utilité.

La matrice ainsi produite sert alors d'intrant au réseau de neurones [1] : les réseaux de neurones ART ONE utilisés pour notre expérimentation ont été développé sur une plate-forme de programmation(Crespo,Savaria, 1995) matricielle disponible sur le grand marché appelé MATLAB. Cette implantation informatique particulière avait été réalisée dans le cadre d'expérimentations sur le traitement massif de signaux. Des modifications mineures d'un certain nombre de caractéristiques ont permis à ce programme de générer des résultats interprétables du point de vue linguistique.

Le réseau neuronal, commence par générer une matrice de résultats qui représentent la classification trouvée.



Chaque ligne (ou vecteur) de cette matrice est constituée d'éléments binaires ordonnés.
Cette ligne indique pour chaque terme du lexique original si il fait, oui ou non, partie du prototype de la classe. Ainsi est crée 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 neurone 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.

Exemple :
Dans notre analyse, nous obtenions dans

La classe # 17 contient les segments 14,29,34,95,181
Tous ces segments sont caractérisées par les termes :
barrage, eau, réserve, digue, indiens, entente, rivière
Il existera évidemment plusieurs classes de ce type.

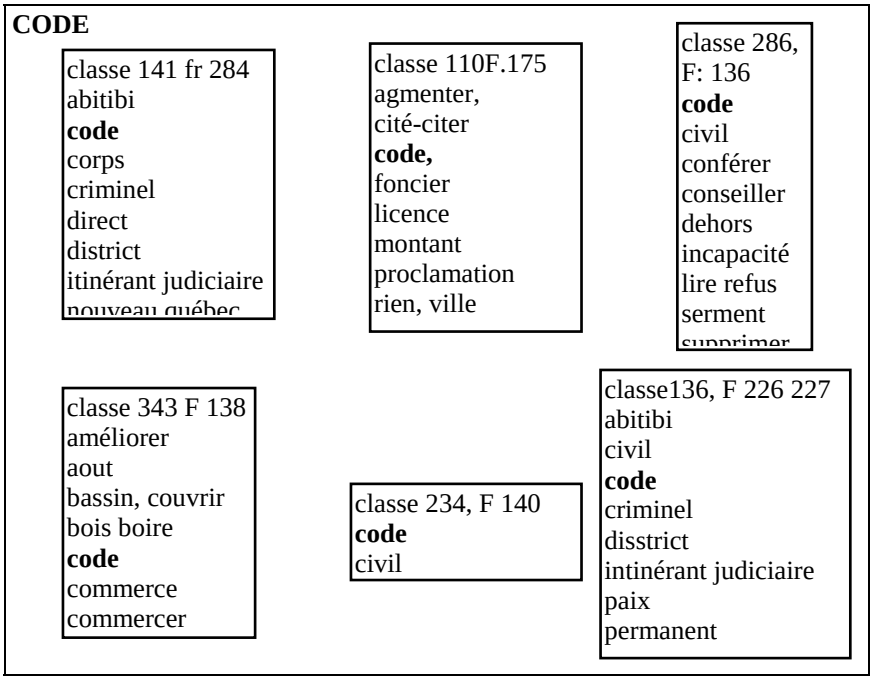
Par exemple:
la classe # 56 contient les segments 23, 34,55,76, 11 120
caractérisés par les termes :
attribution, citoyen, compétent, jugement,légistation
la classe # 57 contient les segments: 12, 39, 52,71,123, 346
caractérisés par les termes :

comté, Chibougamaou,ccôte, parcellaire, substrat.
la classe # 58 contient les segments: 17, 31, 58, 521, 579, 769, 832, 871

qui étaient tous caractérisés par les termes :
minerai, forêt, ressources, relocalisation, bénéfices, expropriation

Dans un texte, il arrive qu'un même terme se retrouve utlisé dans plusieurs contextes sémantiquement différents. A titre d'exemple, prenons le mot CODE. Celui-ci se retrouve effectivement dans plusieurs classes. Il était alors possible pour ce lexème de dessiner le graphe des classes dans lesquels il apparaît.

TABLEAU POUR LE MOT CODE.



On voit alors surgir la différence des réseaux sémantiques de ce terme, qui tantôt est utilisé dans le sens de *code civil*, dans le sens de *code de comportement*, code criminel,etc. Pour un terminologue, ce graphe est parlant. Il sert d'indice aux réseaux sémantique de ce termes et donc de ses acceptions dans ce 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 leur contenu sémantique ou conceptuel est associé.

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 constitué 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 d'établissement de bases terminologiques.

.

La forme actuelle des résultats ne se prêtent pas à l'interprétation directe par le terminologique. Avant de pouvoir servir concrètement, les données devront passer par une interface spéciale qui devrait dans un premier temps permettre l'approche catégorielle des lemmes et dans un second temps permettre la mise en forme hypertextuelle du plein texte à partir de la catégorisation effectuée par le réseau de neurones.

L'outil ainsi développé devrait permettre au terminologue d'accéder le plein texte suivant une stratégie de recherche guidée par le travail de catégorisation ou, advenant le cas, de construire les sens des lemmes par une recherche incrémentielle sur ces derniers.

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 (adaptivité) et sans faire appel à des connaissances pré-formées ou pré-digérées,celles-ci n'étant simplement pas disponible 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 comparable 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.

Recherches et développements futurs

Plusieurs variantes de l'expérimentation sont possibles et envisagées. On peut caractériser les classes autrement que par l'intersection entre les segments qui la constitue. Pourrait-on 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 non-indépendance 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.

Il va de soi qu'on pourrait aussi filtrer à la sortie les classes ne représentant pas une richesse d'information suffisante (selon un critère donné, ex. : pas assez d'unifs dans cette classe).

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 chantier.

.

Bibliographie

Burr, D. J. (1987). "Experiments with a connectionist 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

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. :

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 Exteraction 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). "Acquirig 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,

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. , F. Daoust. , Rolland, S. ,. (1976,). "Sato: A system for Automatic Content Analysis of Text,". *Computer and the Humanities vol. XX. ,* p 45. 64

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. Modelling 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.

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.

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 Bibliothéque de France Station". 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.

note

Plusieurs chercheurs explorent actuellement l'utilisation du modèle neuronal dans le domaine du texte soit a des fins de rappel d'information (Salton et al. 1994;Veroniss J. 1990, Williams, M. (1990), soit a des fins d'analyse sémantique (Lin et al. 1991) ou l'analyse discursive,(Tapiero 1993, Kosko 1991Burr,1987). Notre propre recherche porte sur l'acquisition des connaissances terminologiques sur des corpus en constante mutation.