Un modèle hybride pour le *textual data mining* : un mariage de raison entre le numérique et le linguistique

Ismaïl Biskri et Sylvain Delisle

Département de mathématiques et d'informatique
Université du Québec à Trois-Rivières, Québec, Canada
{Ismail_Biskri, Sylvain_Delisle}@uqtr.uquebec.ca

Résumé

Une des recherches de pointe menée actuellement en informatique 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. Pour résoudre ce problème, plusieurs stratégies sont possibles : les unes relèvent des mathématiques et les autres de l'informatique linguistique. Nous présentons dans cet article un modèle hybride, à la fois robuste et fin, qui s'inspire des modèles neuronaux et de l'analyse linguistique informatique.

1. Introduction

Un nombre croissant d'institutions accumulent très rapidement des quantités de documents qui ne sont souvent classés ou catégorisés que très sommairement. Les tâches de dépistage, d'exploration et de récupération de l'information présente dans ces textes, c'est-à-dire des "connaissances", deviennent alors extrêmement ardues, sinon impossibles. 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 pose 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 puisqu'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.

Le traitement d'un texte par ordinateur en appelle souvent à des dépôts de connaissances préconstruites acquises via des enquêtes cognitives (analyses de protocole) auprès des experts ou puisées dans le répertoire encyclopédique du savoir partagé. Ces connaissances sont alors utilisées comme gabarit dans le dépistage et la reconnaissance. Plusieurs des systèmes qui opèrent dans ce domaine doivent être dotés de mécanismes (moteur d'inférence, maintien de cohérence, 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 au contenu informationnel du texte à traiter. En possession de ce savoir, un tel système informatique de type expert pourrait alors réussir à "comprendre" le texte et donc à en extraire les connaissances. De nombreuses recherches ont d'ailleurs montré la nécessité d'avoir des connaissances de multiples niveaux (syntaxiques, psycholinguistiques, lexicales, sémantiques, encyclopédiques, etc.) : entre autres, Jacobs et Zernik (1988), Moulin et Rousseau (1990), Regoczei et Hirst (1989), Zarri (1990).

Du point de vue de la 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 : cela constitue une première difficulté importante. Dans le cadre de la 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 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.

La deuxième difficulté importante 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 en profondeur des phrases en deçà de quelques secondes par phrase. On peut imaginer le temps requis pour traiter des milliers de pages—sans compter que, souvent, seule une relativement faible proportion d'une telle masse est véritablement utile par rapport à des besoins spécifiques. La situation de la 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 grandes masses 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 (reconnaissance visuelle, sonar, géomatique, etc.). Ce n'est qu'une fois qu'il a cerné le lieu potentiel de 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, dans une deuxième étape, de sélectionner les parties du texte sur lesquelles il veut extraire des connaissances d'une manière plus fine et ce au moyen de méthodes linguistiques. C'est le mariage de raison entre ces deux visions qui constitue le fondement du modèle hybride que nous proposons ici.

2. Stratégies et outils pour le *textual data mining*

La littérature technique relative au traitement de l'information textuelle a montré qu'il était possible de concevoir des outils d'extraction des connaissances dans des textes (*textual 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 un ensemble de critères d'équivalence. Pour les chercheurs dans le domaine de la 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 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—certaines recherches les ont d'ailleurs nettement privilégiées (Church *et al.* 1989, Lebart et Salem 1988, Salton 1988). 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 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, comme dans le cas des 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 et Hanks 1990). Les associations des mots dans les classes ne sont pas toujours facilement interprétables. Pourtant, malgré leurs limites, ces approches ont été jugées des plus utiles pour l'extraction des connaissances. 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. 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, markoviens (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, les modèles thermodynamiques, et les modèles basés tantôt sur la compétition, tantôt sur la rétropropagation, mais surtout sur des règles complexes d'activation et d'apprentissage (Kohonen 1982). 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. 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 modifications sont possibles.

3. Le modèle hybride

Dans sa réalité informatique, le modèle hybride que nous proposons consiste en deux systèmes correspondant à autant d'étapes fondamentales : le système numérique et le système linguistique. Le premier est plus faiblement dépendant de la langue utilisée que le second.

3.1. Le système numérique

Un filtrage numérique grossier du corpus est d'abord effectué. Il permet de classifier et de structurer le corpus en des classes de termes qui serviront d'indices de régularités d'associations lexicales que l'ingénieur de la connaissance ou le linguiste—que nous désignerons par utilisateur—peut utiliser comme tremplin pour approfondir les étapes ultérieures d'interprétation, de construction de réseaux sémantiques, et finalement d'élaboration de ses fiches terminologiques, par exemple. Une plate-forme réalisée au LANCI[1], en l'occurrence la plate-forme CONTERM (programmée en Visual Basic), permet d'exécuter une chaîne de traitement qui réalise un tel filtrage. La chaîne présente les étapes suivantes : elle commence par la préparation du lexique, suit alors une transformation matricielle du corpus puis, finalement, une extraction classificatoire par réseaux de neurones ART[2].

Ainsi, dans un premier temps, le texte est reçu et traité par des modules d'analyse de la plate-forme CONTERM qui constitue un atelier utilisant 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) 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 du texte qu'on se propose de construire, soit parce que leur présence nuirait au processus interprétatif qui suit la tâche de catégorisation. Vient ensuite une description morphologique minimale de type lemmatisation. Une transformation est ainsi opérée pour obtenir une représentation matricielle du texte. Cette transformation est effectuée par d'autres modules de CONTERM explicitement dédiés à cette fin. On produit ainsi un fichier indiquant pour tout lemme choisi sa fréquence dans chaque segment du texte. Enfin un post-traitement suit de façon à construire une matrice dans un format acceptable par le réseau de neurone ART.

Finalement, 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 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 par les termes qui sont présents également dans tous les segments du texte qui ont été placés 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.

3.2. Le système linguistique

Le système linguistique a pour rôle d'effectuer un traitement linguistique en profondeur des segments sélectionnés lors du filtrage qu'a permis le système numérique. L'utilisateur sélectionne donc des segments dont il veut une analyse plus fine et pour lesquels il cherche à extraire une représentation des connaissances plus structurée. Il peut décider, par exemple, de focaliser son attention sur un verbe donné d'un segment particulier et en construire son réseau sémantique. L'analyse linguistique effectuée par ce système permet d'organiser les phrases

---

[1] Laboratoire d'ANalyse Cognitive de l'Information de l'Université du Québec à Montréal.

[2] Le réseau de neurones FUZZYART utilisé pour l'expérimentation de CONTERM a été développé sur une plate-forme de programmation matricielle disponible sur le grand marché appelé MATLAB.

dans lesquelles apparaît le verbe choisi par l'utilisateur sous forme de structures prédicatives *Prédicat (argument$_1$ argument$_2$ ... argument$_n$).* Ainsi, pour une sélection de phrases, l'utilisateur peut engendrer une liste d'expressions prédicatives.

À titre d'exemple, s'il y avait dans cette liste des structures prédicatives ayant des arguments en commun, comme dans Prédicat1 (argument1 argument2) et Prédicat2 (argument3 argument1), cela permettrait à l'utilisateur d'identifier la relation sémantique entre les arguments 2 et 3 par rapport à l'argument 1. D'autres investigations sont aussi intéressantes pour des prédicats identiques comportant des listes d'arguments différentes. Le système linguistique que nous avons utilisé se nomme TANKA (Text ANalysis for Knowledge Acquisition : Delisle 1994, Barker *et al*. 1998) ; il permet tout cela en plus de la conservation dans une base de données des structures prédicatives. TANKA est implémenté en Quintus Prolog et en C sur des stations de travail Sun. Il est indépendant de tout domaine particulier et ses deux principaux modules sont DIPETT et HAIKU : le premier est un analyseur syntaxique en constituants alors que l'analyseur sémantique HAIKU traite les arbres de DIPETT pour déterminer les relations sémantiques qui lient un verbe à ses arguments (Delisle *et al*. 1996) et un nom (tête) à ses modificateurs (Barker et Szpakowicz 1998).

L'analyse syntaxique nous permet d'accéder aux constituants syntaxiques dont nous avons besoin pour l'analyse sémantique. DIPETT (Domain-Independent Parser for English Technical Texts) effectue une analyse syntaxique détaillée des phrases en entrée (Delisle 1994, Delisle *et al.* 1998). En l'absence de connaissances sémantiques *a priori*, une analyse syntaxique détaillée et indépendante du domaine est le seul guide fiable vers le sens des énoncés d'un corpus. La partie centrale de l'analyse sémantique est l'analyse casuelle semi-automatique et interactive de HAIKU qui utilise un système de cas général, indépendant de tout domaine particulier (Barker *et al*. 1997). À la section 5, nous présenterons certains détails supplémentaires en rapport avec les sorties de notre système linguistique.

3.3. Illustration des rôles spécifiques des systèmes numérique et linguistique

C'est grâce au système numérique décrit à la section 3.1 qu'il est possible d'obtenir une classification en trois groupes de l'usage du mot *ferme* dans un certain corpus. Ces trois groupes sont :

Groupe 1 (adjectif) : `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 (nom) : `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 (verbe) : `Le vote ferme la discussion. La décision du gouvernement ferme les options. Le président de l'assemblée ferme le vote.`

De même, c'est grâce au système linguistique de la section 3.2 que l'utilisateur peut identifier le rôle sémantique de chaque argument d'un certain verbe, par exemple, en construisant le réseau sémantique qui lui est associé. Considérons les phrases du groupe 3 ci-dessus : elles utilisent toutes le terme *ferme* en tant que verbe. Les structures prédicatives associées à ces phrases et obtenues par une analyse linguistique comme celle présentée en 3.2 seraient alors les suivantes :

```
ferme (le vote) (la discussion)
ferme (la décision du gouvernement) (les options)
ferme (le président de l'assemblée) (le vote).
```

4. La méthodologie associée au modèle hybride

Au modèle hybride que nous proposons dans cet article, nous associons une méthodologie d'extraction de connaissances dans les textes qui s'organise en quatre étapes majeures :

1) **Classification (séparation) du corpus initial en ses différents domaines effectué à l'aide de CONTERM** : Les parties textuelles abordant des sujets relativement différents dans un texte hétérogène peuvent être facilement séparées les unes des autres à travers une classification par un traitement automatique avec CONTERM. En effet, il est certain que des rédacteurs humains, pour aborder des sujets différents, vont utiliser un vocabulaire spécifique représentatif du sujet abordé. Les contraintes 'esthétiques' langagières, dirons nous, pourraient avoir un impact négatif sur la séparation des parties textuelles en question. En effet, un vocabulaire 'fonctionnel', qui est généralement présent de façon régulière dans toutes les parties d'un texte et qui n'a aucune pertinence spécifique avec un domaine de connaissance particulier pourraient corrompre les résultats d'une classification neuronale à la CONTERM. Toutefois, avec les étapes de nettoyage du lexique du corpus (section 3.1), le vocabulaire fonctionnel est éliminé dans sa grande majorité. Il est certain dès lors que la classification ne tiendra compte que du vocabulaire spécifique aux différents grands thèmes abordés dans le texte. Bien entendu, cette classification n'est pas une interprétation automatique. C'est à l'utilisateur de faire l'interprétation sachant qu'à cette étape-ci la classification que nous obtenons est uniquement une séparation des grandes parties du corpus traitant de thèmes différents. Enfin, il faut savoir que chaque grande partie est représentée par une ou plusieurs classes de mots. C'est le regroupement manuel de ces classes de mots qui va permettre à l'utilisateur de préciser selon son interprétation les grands thèmes.

2) **Identification des thèmes (classes) dans un des textes ou un des domaines obtenu à l'étape #1** : Dès lors que l'utilisateur a réussi à reconnaître les grandes parties très générales, il cherche les sous-thèmes caractérisant chaque partie. Ceci lui permet d'identifier des thèmes plus pertinents et d'aller plus en profondeur dans le détail de l'interprétation. L'identification de thèmes moins généraux mène à identifier des sous-parties du corpus plus spécifiques qu'à l'étape précédente. Ces sous-parties seront choisies ou non à l'étape suivante par l'utilisateur de sorte qu'il en puisse extraire des connaissances au moyen d'outils linguistiques. Comme pour la première étape, la décision dans l'interprétation et le choix du thème d'une classe par l'utilisateur est une opération manuelle. Nous rappelons que nous n'avons pas l'ambition ici d'une interprétation automatique : ce serait d'ailleurs contraire aux principes de la LATAO. Seulement, l'usager aura des regroupements de mots (des classes) qui apparaissent régulièrement ensemble dans le texte. Il est indéniable dès lors que ces classes constituent un indice qui permet à l'usager de prendre sa décision quant aux choix du thème associé à la classe.

3) **Exploration des classes identifiées à l'étape #2 en fonction des besoins en extraction de connaissances** : Cette étape est manuelle mais il est possible de la rendre automatique. Ceci est une de nos prochaines intégrations envisagées. Toutefois, au stade actuel du développement, l'utilisateur ayant associé à chaque classe un thème particulier à l'étape précédente aura à choisir les thèmes qui l'intéressent et par conséquent les parties du texte (ou, segments) relatives à ces thèmes. Dans notre méthodologie, un thème caractérise une classe et une classe contient les parties du texte ayant des similitudes. Donc, choisir un thème particulier signifie décider quelles vont être les parties du texte à analyser plus en détail par des outils linguistiques.

4) **Analyse détaillée des phrases sélectionnées dans les segments jugés intéressants par l'utilisateur** : effectué à l'aide du système linguistique, tel que nous l'illustrons ci-dessous.

5. Le modèle hybride confronté aux phénomènes météo et à la mécanique des petits moteurs : une première expérimentation

Nous allons maintenant illustrer l'application de notre modèle et de la méthodologie associée que nous venons de décrire en relatant les résultats d'une expérience effectuée récemment qui avait pour double objectif de peaufiner notre modèle et de le tester sur un véritable corpus. Ce corpus était composé de deux textes indépendants écrits en langue anglaise : un livre pour jeunes sur certains phénomènes météo contenant environ 600 phrases (Larrick 1961) et un livre sur la mécanique des petits moteurs contenant environ 1600 phrases (Atkinson 1990). Ainsi, contrairement à Toussaint *et al*. (1998), par exemple, nous abordons la problématique

du traitement de corpus formés de textes non homogènes et provenant de différents domaines. Les deux textes ont été sauvegardés l'un à la suite de l'autre dans un unique fichier. À l'étape #1 (nous référons ici aux étapes identifiées à la section 4), la segmentation du corpus a donné 243 segments; les segments ayant en moyenne une longueur de 10 mots. Les segments #1 à #67 constituent le premier texte et les segments #68 à #243 constituent le deuxième texte : nous soulignons que CONTERM a pu parfaitement séparer les deux textes. Avec ces paramètres nous obtenions un nombre de classes égal à 83. Notons toutefois que CONTERM a séparé les deux textes du fait qu'ils sont dissemblables. Si les deux textes étaient plus proches dans leur contenu, CONTERM aurait donné une autre classification pouvant mener à d'autres conclusions (par exemple que les deux textes ont la même thématique). Nous insistons particulièrement sur un point précis au risque de nous redire : CONTERM ne donne que des indices lexicaux qui permettent à l'utilisateur de formuler des conclusions ou des interprétations.

Ensuite, à supposer que nous ayons été spécifiquement intéressés à la météo, l'étape 2 nous a amené à un examen des classes associées à ce domaine pour identifier une classe portant sur les nuages et la pluie *(clouds, moisture, rain)*, puis une autre sur la terre *(earth)*, une autre sur les orages *(storm)*, une autre sur le ciel *(sky)*, etc. L'étape 3 nous a permis de choisir parmi ces classes celles qui sont pertinentes par rapport à nos besoins en information, disons les nuages et la pluie, car nous cherchions à construire un réseau sémantique (ou graphe conceptuel) du concept *clouds*, et de sélectionner les phrases (des classes retenues) qui méritaient une analyse linguistique plus fine. Aux fins d'exemple, supposons qu'un des segments retenus à l'étape #3 ait été le suivant (extrait authentique de Larrick (1961)) : "`These are cumulus clouds. On a sunny afternoon, these clouds are likely to be white. The tops may shine in the sunlight. We might call them fair-weather clouds. But sometimes cumulus clouds turn into black storm clouds.`". Finalement, à l'étape 4, nous avons procédé à l'analyse des phrases retenues à l'étape #3 à l'aide du système linguistique. Nous allons maintenant montrer les grandes lignes du traitement linguistique à travers l'analyse d'une des phrases du segment retenu ici, soit la phrase "`But sometimes cumulus clouds turn into black storm clouds.`".

L'analyse linguistique débute par une analyse syntaxique détaillée (nous passons outre les détails de l'analyse lexico-morphologique préalable) qui est produite automatiquement par DIPETT et nous donne un unique arbre d'analyse qui est correct. Nous reproduisons ici une forme simplifiée de cet arbre qui nous permet d'identifier facilement les constituants de la phrase choisie :

```
( but:conj-coord
  ( sometimes:adv
       ( ( ( cumulus:noun ) clouds:noun )
         turn:verb
         ( into:prep ( ( black:adj ( ( storm:noun ) clouds:noun ) ) ) ) ) ) )
```

Puis, à l'aide de l'analyseur sémantique semi-automatique HAIKU, l'utilisateur identifie les relations sémantiques entre un verbe (prédicat) et ses arguments, de même qu'entre un nom (tête) et ses modificateurs. Ces résultats sont continuellement sauvegardés par HAIKU dans une base de données et sont utilisés dans les traitements subséquents afin que le système apprenne à assister l'usager de plus en plus "intelligemment" pour les prochains énoncés. Par la même occasion, cette base de données constitue aussi un modèle de plus en plus étoffé des concepts analysés par HAIKU. Poursuivons maintenant avec notre exemple : voici une version simplifiée (commentaires en italiques et entrées de l'utilisateur encadrées en foncé) de la première partie du traitement avec HAIKU pour le verbe et ses arguments; il s'agit d'une analyse casuelle.

{HAIKU indique comment il a découpé la phrase à partir de l'arbre syntaxique}

```
CURRENT SUBJECT: "cumulus clouds"            CURRENT VERB   : turn
CURRENT COMPL  : "into black storm clouds"   CURRENT MODIFS : sometimes
```

{Le patron syntaxique (CMP) correspondant au découpage ci-dessus est formé d'un groupe sujet, d'une phrase prépositionnelle introduite par 'into' et par un adverbe}

```
CMP (syntactic pattern) found by HAIKU: psubj-into-adv
```

{HAIKU informe l'utilisateur qu'après consultation de sa base de données et de l'application d'un algorithme de recherche d'un "meilleur voisin", il n'a pas de "brillante" suggestion à proposer quant au patron sémantique (CP) qui pourrait être associé au patron syntaxique (CMP) courant.}

```
S3: entirely new CMP with known verb; no obvious candidate CPs to propose
```

{Il informe l'utilisateur de ce qu'il connait par rapport à la préposition 'into'. Cas présentés ici : tto="time to", lto="location to", dir="direction"}

```
  here is what the cmDict knows about this Case marker: INTO
  {format: Case(COmmonness,Example); CO=1:common, CO=2:uncommon}
    tto(1,We worked INTO THE NIGHT.)
    lto(1,Jump INTO THE POOL.)
    dir(1,Turn the boat INTO THE WIND.)
```

{Il demande maintenant à l'utilisateur d'identifier le patron sémantique (CP) qui doit être assigné à l'interprétation de la phrase courante. Notons que le patron entré par l'usager, expr-eff-freq (en foncé), est la seule entrée qu'a eu à effectuer l'utilisateur jusqu'ici. Cas présentés ici : expr= "experiencer", eff="effect", freq="frequency".}

```
  please enter the new CP (e.g. agt-obj-tat),
    or enter 'h' to see the current input string and CMP,
       or CR to abort ["new CP"/h/CR]? expr-eff-freq
```

{HAIKU s'assure que les cas sémantique ont bien été assignés aux bons arguments du verbe de la phrase courante.}

```
CMP & CP will be paired as follows: psubj/expr into/eff adv/freq  ; Correct [n/Y]?
Y
```

{HAIKU nous informe de toutes les mises à jour qu'il a effectué sur sa base de données, composées des différents "dictionnaires" ci-dessous.}

```
  UPDATING mDict for turn ...
   ⇨  turn  [psubj-into-adv:1,psubj-to:1]
            [[adv,[freq:1]],[into,[eff:1]],[psubj,[expr:2]],[to,[eff:1]]]
  UPDATING cmpDict for psubj-into-adv ...
   ⇨  psubj-into-adv  [[expr-eff-freq:1,
            '[but,sometimes,cumulus,clouds,turn,into,black,storm,clouds,.]']]
  UPDATING cpDict for expr-eff-freq ...
   ⇨  expr-eff-freq  [turn]
  UPDATING ccvpIndex ...
   ⇨  ccvpIndex:  psubj-into-adv  expr-eff-freq  turn102
  VERB PREDICATE STRUCTURE (verb_pred_struct) FOR turn ...
   ⇨  [psubj-into-adv/expr-eff-freq/np-into-adv]
```

Le système met aussi à jour l'arbre d'analyse syntaxique en lui ajoutant l'information casuelle qui vient d'être identifiée. La structure syntaxico-sémantique est la suivante (le lecteur pourra constater qu'il est trivial de traduire cette structure à l'aide de graphes conceptuels) :

```
   ( but:conj-coord
        ( [freq: sometimes:adv]
             ( [expr: ( ( cumulus:noun ) clouds:noun )]
             turn:verb
             ( [eff: into:prep ( ( black:adj ( ( storm:noun ) clouds:noun ]
```

Finalement, la dernière partie du traitement effectué par HAIKU est l'analyse des groupes nominaux (appelés *noun compounds* en anglais). La phrase courante contient deux groupes nominaux : *cumulus clouds* et *black storm clouds*.

```
For the phrase 'cumulus cloud'
there is a relationship between (cumulus) and (cumulus_cloud).
```

{Un "NMR label", différent des cas utilisés pour les verbes, sert à identifier le rôle sémantique d'un modificateur par rapport à un nom tête. Ici, prop=property.}

```
> Please enter a valid NMR label ('a' to abort): prop
Property (prop): cumulus_cloud is cumulus
```

{HAIKU propose une première interprétation qui est erronnée : 'black' n'est pas un modificateur de 'storm' mais de 'clouds'. L'utilisateur répond 'n' (non) et HAIKU corrige sa suggestion.}

```
For the phrase 'black storm cloud'
is that 'black storm' [Y/n]? n

      _____________
     |             |
   black       _________
              |         |
            storm     cloud

For the phrase 'storm cloud'
there is a relationship between (storm) and (storm_cloud).
```

{Ici, resu=result.}

```
> Please enter a valid NMR label ('a' to abort): resu
Result (resu): storm is a result of storm_cloud

For the phrase 'black storm_cloud', do you accept the assignment:
Property (prop): black_storm_cloud is black [n/a/<nmr>/Y] Y
```

{Et, finalement, HAIKU nous livre les résultats du modèle des groupes nominaux de la phrase analysée.}

```
NOUN MODIFIER RELATIONSHIPS
  Property (prop): cumulus_cloud is cumulus
  Property (prop): black_storm_cloud is black
  Result (resu): storm is a result of storm_cloud
  storm_cloud isa cloud
```

Encore une fois, on peut constater combien il est simple d'augmenter la structure syntaxico-sémantique de traits supplémentaires à partir de ces résultats. Grâce au traitement de DIPETT et HAIKU, l'utilisateur a associé à sa phrase un arbre syntaxique, puis une structure syntaxico-sémantique avec information casuelle et traits sémantiques des groupes nominaux composés.

6. Conclusion

Nous venons de présenter un modèle d'extraction des connaissances pour ingénieurs de la connaissance et linguistes. 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'application, dans l'ordre, d'une méthode numérique avant de faire intervenir une méthode linguistique est stratégique et résulte du compromis nécessaire pour faire cohabiter ces deux approches de manière complémentaire et productive. En effet, la méthode numérique est plus à même de "débroussailler" un gros texte et de permettre à l'utilisateur de soumettre des segments choisis à l'analyseur linguistique plus fin.

Ce type d'approche permet de répondre à une des critiques importantes faites 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 analyses linguistiques détaillée sont trop fines et donc trop lentes sur un corpus ample. Mais bien placées elles ne travaillent que sur des sous-corpus sur lesquels a déjà été effectué un premier travail de désambiguisation. En fait cette approche oblige à effectuer la désambiguisation, ou du moins une partie de celle-ci, 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 linguistique fine 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. De plus, la configuration des résultats telle que permise par les expressions prédicatives permet à l'utilisateur d'avoir une organisation plus limpide sur le plan ergonomique et cognitif.

Nous travaillons également sur deux autres modèles hybrides numérique/linguistique associant réseaux de neurones et le modèle d'exploration contextuelle d'une part (Jouis *et al.* 1997) et réseaux de neurones et Grammaires Catégorielles d'autre part (Biskri et Meunier 1998). Le premier système permet de détecter des liens sémantiques de types cinématiques ou dynamiques (mouvements d'objets, changements d'états d'objets, relations de causalité, recherche des contextes définitoires d'un terme, etc.). Cela permettra d'élargir les interprétations par l'utilisateur des niveaux descriptifs statiques du domaine vers des descriptions évolutives. Le deuxième système permet de construire, pour la langue française, des structures applicatives de type prédicat arguments qui permettent de structurer un texte à des fins d'extraction de connaissances sémantiques fonctionnelles.

Références

Atkinson, H.F. (1990), *Mechanics of Small Engines*, New York: Gregg Division, McGraw-Hill.

Barker, K., Copeck T., Delisle S., Szpakowicz S. (1997), "Systematic Construction of a Versatile Case System", *Journal of Natural Language Engineering* 3 (4), 279-315.

Barker, K., Delisle, S., Szpakowicz, S. (1998), "Test-Driving TANKA: Evaluating a Semi-Automatic System of Text Analysis for Knowledge Acquisition", *Proceedings of the 12th Canadian Artificial Intelligence Conference — CAI-98*, Vancouver (BC, Canada), 60-71.

Barker, K., Szpakowicz, S. (1998), "Semi-Automatic Recognition of Noun Modifier Relationships", *Proceedings of COLING-ACL'98 Conference*, Montréal (Québec, Canada), 96-102.

Biskri, I., Meunier, J.G. (1998), "Vers un modèle hybride pour le traitement de l'information lexicale dans les bases de données", *Actes du Colloque JADT-98*, Nice (France).

Bouchaffra, D., Meunier, J.G. (1993), "Theory and Algorithms for analysing the consistent Region in Probabilistic Logic", *International Journal of Computers & Mathematics with Applications*, Vol. 25, No. 3, edit. Ervin Y. Rodin, Published by Pergamon Press.

Church, K., Gale, W., Hanks, P., Hindle, D., 1989. "Word Associations and Typical Predicate-Argument Relations", *Proceedings of the 1st International Workhop on Parsing technologies*, Carnegie Mellon University.

Church, K.W., Hanks, P. (1990), "Word Association Norms, Mutual Information, 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.D. Thesis, Department of Computer Science, University of Ottawa.

Delisle, S., Barker K., Copeck T., Szpakowicz S. (1996), "Interactive Semantic Analysis of Technical Texts", *Computational Intelligence*, 12(2), 273-306.

Delisle, S., Létourneau, S., Matwin, S. (1998), "Experiments with Learning Parsing Heuristics", *Proceedings of COLING-ACL'98 Conference*, Montréal (Québec, Canada), 307-314.

Jacobs P., Zernik, U. (1988), "Acquiring Lexical Knowledge from Text: A Case Study", *Proceedings of the Seventh National Conference on Artificial Intelligence (AAAI-88)*, Saint Paul (Minnesota, USA), 739-744.

Jouis, C., Biskri, I., Descles, J.P., Le Priol, F, Meunier, J.G., Mustafa, W., Nault G. (1997), "Vers l'intégration d'une approche sémantique linguistique et d'une approche numérique pour un outil d'aide à la construction de bases terminologiques", *Actes du colloque JST97*, Avignon (France).

Kohonen, T. (1982), "Clustering, Taxonomy and Topological Maps of Patterns", *Proceedings of the 6th IEEE International Conference on Pattern Recognition*, 114-122.

Larrick, N. (1961), *Junior Science Book of Rain, Hail, Sleet & Snow*. Champaign: Garrard Publishing Company.

Lebart, L., Salem, A. (1988), *Analyse statistique des données textuelles*, Paris: Dunod.

Meunier, J.G., Biskri, I., Nault, G., Nyongwa, M. (1997), "Exploration de classifieurs connexionnistes pour l'analyse terminologique", *Actes du colloque RIAO97*, Montréal (Québec, Canada).

Moulin, B., Rousseau, D. (1990), "Un outil pour l'acquisition des connaissances à partir de textes prescriptifs", ICO, Québec 3 (2), 108-120.

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).

Toussaint, Y., Namer, F., Daille, B., Jacquemin, C., Royauté, J., Hathout, N. (1998), "Une approche linguistique et statistique pour l'analyse de l'information en corpus", *Actes de la Conférence TALN-98*, Paris (France), 182-191.

Zarri, G.P. (1990), "Représentation des connaissances pour effectuer des traitements inférentiels complexes sur des documents en langage naturel", Office de la langue française (Ed), *Les industries de la langue : Perspectives 1990*, Gouvernement du Québec.