L'INGENERIE TEXTUELLE ET COGNITIVE DES SYSTEMES "EXPERTS" EN DROIT

par: Louis-Claude PAQUIN
chercheur
Centre d'Analyse de Textes par Ordinateur
Université du Québec à Montréal

Je ne suis ni juriste ni informaticien. De cette position privilégiée je compte proposer quelques réflexions à propos des multiples utilisations possibles de la technologie des systèmes à base de connaissances dans un contexte juridique. Je poursuis avec Luc Dupuy des recherches en ingénierie cognitive et textuelle. A ce titre, une grande partie de nos activités consistent en un transfert de technologie vers l'administration publique[1]. Plutôt que le modèle du faire-faire par des tiers (consultants), nous offrons le modèle du faire soi-même avec le pari que seul un système d'information produit de l'intérieur reflète véritablement la "culture" locale de l'organisation.

Les environnements informatiques que nous avons développés à cet effet, SATO[2] et D_expert[3] obligent en quelque sorte leurs utilisateurs à jouer le rôle de développeurs d'applications sur mesure pour leurs besoins sans qu'ils n'aient à apprendre la programmation. Tout le savoir-faire quant à la tâche à accomplir, tant textuelle dans le cas de SATO que cognitive dans le cas du D_expert, est la responsabilité de l'utilisateur. L'encadrement que nous offrons aux équipes internes à l'organisation qui ont le mandat avec nos outils logiciels de mettre au point un système d'information est surtout d'ordre méthodologique.

Les réflexions proposées concernent:

- l'aspect logiciel des générateurs de systèmes à base de connaissances,
- l'aspect textuel de la connaissance,
- l'ingénierie textuelle,
- la dimension textuelle de l'expertise,
- les activités de l'ingénierie textuelle,
- l'effet de référence dans les textes,
- la construction d'un dictionnaire de concepts,
- l'introduction d'extraits de textes pour supporter une expertise,

---

[1] Notamment au gouvernement du Québec à l'Environnement, au Revenu, au Secrétariat du conseil du Trésor et à la Commission des mormes du travail

[2] François Daoust, Centre d'analyse de textes par ordinateur (UQAM) version SATO 3.5

[3] Louis-Claude Paquin, Centre d'analyse de textes par ordinateur (UQAM) version D_expert 2.01

- l'architecture d'un système textuel et cognitif,
- la modélisation de la lecture des textes plutôt que de leur contenu.
- le caractère synthétique du raisonnement obtenu par les systèmes.

Les nouveaux types de moteur d'inférences qui tiennent la route sont de moins en mois fréquents. Il semble donc que l'aspect algorithmique de cette technologie est maîtrisé. Il n'en va pas de même pour l'aspect logiciel; il n'y a pas, à ma connaissance, de générateur de systèmes à base de connaissance qui serait, dans sa catégorie, l'équivalent des produits Borland pour les compilateurs. Pour s'attaquer à des problèmes le moindrement d'envergure, ne serait-ce que pour une large couverture à la solution d'un problème simple, il faut pouvoir dépasser le millier de règles d'inférences et même plusieurs fois. Vu leur nombre, les règles ne pourront pas toutes être écrites par l'équipe de développement. Les paquets de règles deviendront alors des marchandises, comme les logiciels, avec des droits d'auteurs (espérons-le) des guaranties et bien sûr des limitations à ces dernières. Pour cela, il faut des standards d'expression des règles d'inférences.

L'administration publique est régie, se réfère et manipule (production, analyse, gestion, etc.) des données textuelles. C'est ainsi que la l'inventaire des sources de la connaissance passe obligatoirement par la loi, les règlements, leur interprétation, la jurisprudence et enfin sur des entrevues avec les experts du domaine. Cependant, jusqu'à maintenant, sauf de rares exceptions, les textes fondamentaux ne sont pas pris en charge par l'informatique après leur production. Le bénéfice qu'il y aurait à réinjecter dans l'organisation la connaissance qu'elle dépose dans ses textes est pourtant clair. Le premier besoin serait de pratiquer un accès sélectif au contenu du corpus qui sorte du sentier battu de l'index ou encore qui en tienne lieu.

Cette ingénierie textuelle recouvre entre autres des activités de stockage, d'indexation conceptuelle et d'exploitation d'immenses bases de données en format libre. Le format des données textuelles est dit libre parce qu'il est sans contrainte de longueur et à géométrie variable (le paragraphe, la page ou le chapitre). En regard du volume des données textuelles à indexer, pratiquer cette opération manuellement entraîne des délais inacceptables dans la disponibilité des nouveaux documents. Par ailleurs, les algorithmes d'indexation automatique restent à raffiner; une approche de type système expert permettrait aux indexeurs humain de transférer leurs heuristiques et de procéder par tentative à l'élaboration d'un modèle. Ce système à base de connaissances devrait être en mesure d'effectuer des opérations textuelles:

filtrage tronqué sur les chaînes de caractères, tri, catégorisation en contexte, vérification des positions avant et après, etc.

L'ingénierie textuelle nous sert aussi au dépistage de la structure cognitive, c'est-à-dire l'ensemble des concepts et leurs caractéristiques qui sont mis en oeuvre dans le corpus. L'utilisation intégrale des textes pour l'acquisition de la connaissance présente toutefois des difficultés. Les mots ne réfèrent pas toujours au monde extérieur, mais peuvent servir à une re-catégorisation de ceux-ci pour constituer des paradigmes, c'est-à-dire des classes d'équivalences contextuelles, ou pour constituer des méréonomies. Il s'agit de hiérarchies régissant des complexes de relations entre un tout et ses parties, entre les parties de parties, etc. La modification continuelle du cadre référentiel du texte amène souvent le lecteur à produire des inférences.

Les méthodes strictement cognitivistes laissent dans l'ombre la dimension textuelle de l'expertise. La recherche de concepts qui seraient formulés en clair et de façon complète dans les textes est habituellement fort décevante. Les définitions sont la plupart du temps partielles, contexualisées ou relatives à d'autres concepts, donc inutilisables directement parce que locales. Les concepts ne sont pas déposés dans les textes, ils sont en partie fabriqués par les textes et en partie par leur lecteur. L'accès aux concepts dans les textes n'est que rarement direct, il plutôt est médiatisé par les termes. Les termes sont les mots qui, dans un contexte donné, désignent une instance particulière du concept.

Dans les textes, l'effet de référence est largement tributaire des formes nominales. Cet effet de référence n'est presque jamais le fait du terme seul, il est habituellement consolidé, spécifié, qualifié, élaboré par d'autres références {épithète, complément du nom, proposition relative}. Certaines formes nominales, les termes, exercent une fonction de régie sur d'autres mots qui en sont les caractéristiques; par ex.: "congé sans solde pour fonder une entreprise". Les termes dépistés dans les textes permettent de construire les concepts. Un passage systématique et rigoureux des termes aux concepts nécessite un cadre ou modèle de «concept» qui tienne autant compte du point de départ, les contextes, que du point d'arrivée, les concepts ces descriptions formelles qui servent à construire les règles d'inférences.

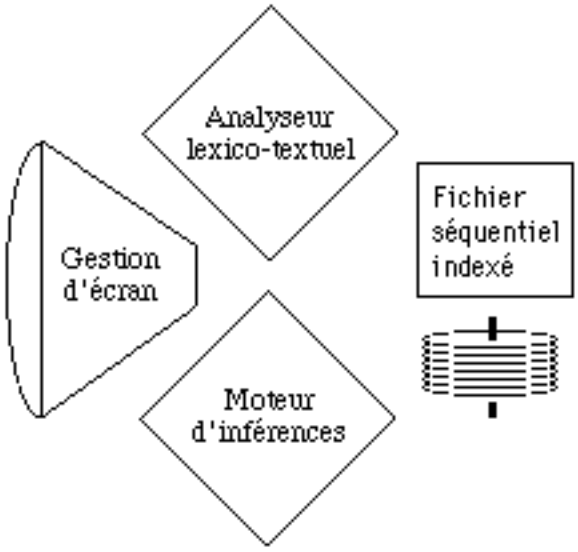
Il n'est pas, à ma connaissance, de définition formelle unifiée de l'unitée de base: le concept. Celle-ci s'avère pourtant nécessaire pour effectuer des opérations autres que l'inférence et la généralisation explicite (par héritage), soit l'association, la différenciation, l'évaluation, la prise en compte des

contradictions, etc. Les concepts ne sont certainement pas des prédicats qui régissent des arguments, ils sont plutôt des arguments valués c'est-à-dire des formes schématiques qui "encapsulent" toutes leurs consolidations possibles en terme de caractéristiques. Le modèle des concepts comme combinaison de "primitives" décrivant la réalité s'avère le plus intéressant parce qu'il est autant efficace dans le cadre logique des systèmes experts que dans le cadre morpho-syntaxique des textes. Les objets valués conviennent autant à la rédaction des règles d'inférences qu'à la description du groupe nominal. Ce modèle est cependant réducteur parce qu'il ne permet pas la prise en compte aisée du phénomène textuel de la re-catégorisation, soit lorsqu'un terme est mis à la place d'un autre, plus général, plus spécifique.

Des techniques peuvent avec profit être empruntées à l'analyse de textes par ordinateur pour constituer un dictionnaire de concept. Il s'agit d'isoler, en raison de leur récurrence, les principaux termes et, pour chacun, de relever l'ensemble des contextes d'occurrence. Cet ensemble de contextes est par la suite classé à partir des traits sémantiques communs ou différents que portent les déterminants. Pour reconstituer ces configurations conceptuelles, des patrons morpho-syntaxiques sont utilisés, ce qui guarantit un dépistage indépendant des problématiques à l'oeuvre dans les textes. Le recours aux experts intervient après coup, pour valider et réduire le matériau cognitif recueilli.

Des extraits de textes peuvent être dépistés par un patron de fouille formulé autour de la lexicalisation de concepts afin de se constituer un sous-texte exhaustif qui permette la formulation de règles d'inférences. Ainsi, par exemple, tous les partagraphes où il est en même temps question de «reprise de possession du logement» et de la «bonne foi du propriétaire». Des extraits de textes peuvent aussi être directement intégrés dans le système à base de connaissance pour enrichir les couches communicationnelles qui entourent le noyau d'expertise, juridique dans le cas qui nous occupe. Il s'agit d'une structure d'interprétation de l'inférence. Ce rôle est particulièrement bien joué par les systèmes d'hypertexte qui permettent un accès associatif à l'information qui correspond en gros à la notion de «voir aussi» dans les index.

L'Atelier Cognitif et TExtuel (ACTE), auquel nous travaillons depuis plus d'un an, intègre dans un même environnement informatique accessible à tous (MacPlus 1mo. ou PC 640k. sous DOS) un moteur d'inférences (D_expert) couplé à un fichier séquentiel indexé (SIGIRD) et à un moteur textuel (SATO). Cet atelier logiciel permettra d'accomplir les fonctions précédemment décrites.



La modélisation des articles de loi ou règlements en règles d'inférences s'avère souvent insatisfaisante. La "textualité" de la loi est aussi importante que son contenu. De plus, toute base de règles ne peut aspirer à un statut supérieur à celui d'une interprétation parmi d'autres du texte de loi. Une architecture comme celle ACTE permet de revoir complètement notre stratégie de construction de systèmes experts juridiques ou administratifs.

Le texte demeure texte mais acquiert le statut d'une seconde base de fait parce qu'il est pleinement accessible: il peut être fouillé par en prémisse de règles d'inférences et peut être catégorisé par des règles déclenchées via le langage d'interrogation de SATO et sa structure de propriétés au lexique et au texte. Ces fonctionnalités nous permettent de modéliser, non plus les textes, mais leur lecture. C'est ainsi que l'ingénierie de la connaissance se déplace de la loi vers le juriste-lecteur.

Il ne s'agit pas pour autant de préconiser une approche strictement linguistique au problème de l'acquisition de la connaissance dans les textes mais plutôt de constituer un modèle de lecture qui comporte une bonne part d'inférences. On constate que le lecteur- expert établit sans cesse des liens entre ce qui est entrain de lire et ce qu'il a précédemment lu. Ce mécanisme lui permet d'établir des trajectoires de lecture particulières. Quels que soient ses objectifs de lecture qui entraînent un filtrage particulier du texte, les trajectoires se dessinent par renforcement des tendances manifestées par les indices relevés.

L'implantation d'un modèle de lecture dans un système dont le contrôle est assuré par un moteur d'inférences, nous libère des contraintes de la

programmation fonctionnelle. Pour modéliser la lecture, au moyen d'automates par exemple, on doit préalablement mettre au point un modèle qui prévoit nécessairement et dans le moindre détail le déroulement de la description. Un tel modèle est déterministe. Le moteur d'inférences gère un modèle dirigé par les faits. Ainsi, la description d'un phénomène n'est entreprise que si assez d'indices sont réunis.

Il reste à définir la place des systèmes à base de connaissance dans l'environnement informationnel pré-existant. Il reste aussi à redéfinir les activités des utilisateurs, à en évaluer l'impact sur leur rendement. Ces systèmes viennent participer à une solution plus globale au problème du débordement de la capacité de traitement de l'utilisateur en lui fournissant de l'information qualifiée. Il serait probablement improductif et éventuellement contre-indiqué d'envisager l'implantation de ces systèmes pour remplacer les experts humains.

A titre d'explication, une analogie entre le raisonnement synthétique des systèmes à base de connaissance et les produits phramaceutiques qui sont issus de synthèse en laboratoire de quelconques éléments naturels qui présentent les propriétés thérapeutiques recherchées. Celles-ci se trouvent extraites de leur milieu, reproduites et éventuellement maximisées, tant et si bien que le produit de synthèse s'avère un redoutable poison. Il importe de nous garder de faire de même avec le principe du "modus ponens" qui n'est qu'un élément parmi d'autres qui interviennent lors d'une prise de décision. D'autres aspects comme l'évaluation en fonction d'objectifs contradictoires, l'intuition, etc. s'avèrent tout aussi importants bien qu'ils échappent à nos tentatives actuelles de modélisation.