**L'informatique et les disciplines qui y sont rattachées adorent les « mots magiques ». Ils nous rappellent que le développement de cette industrie est rythmé par ses avancées technologiques que sont censés désigner ces mots magiques. Le terme « sémantique » est de ceux-là.**

# Il y a sémantique et... sémantique

par Alain Beauvieux

Nous assistons à une frénésie de «quelque chose de » sémantique : l'analyse bien sûr mais aussi les moteurs, la cartographie et plus récemment des interfaces, sans doute ... parce qu'elles comprennent les intentions de l'utilisateur !

Or, si cette dernière remarque peut faire sourire, le sujet est extrêmement sérieux. Tout d'abord, remarquons qu'il n'est pas vraiment nouveau. En linguistique, bien évidemment mais aussi en informatique.

**Sémantique : le terme a été repris il y a près d'un demi-siècle par les pionniers d'une nouvelle discipline, le traitement automatique des langues naturelles (T.A.L.N.) qui, appliquant les méthodes d'analyse de la linguistique traditionnelle, distinguait quatre niveaux : lexical, syntaxique, sémantique et pragmatique.**

## Question d'écoles

Dès les années 1970, ce sujet fut l'objet de controverses. Pour certains, ce débat s'est bien évidemment déplacé sur un terrain plus philosophique, convaincu qu'avec les modèles mathématiques appropriés un ordinateur pourrait rivaliser avec l'intelligence humaine. Ils s'opposaient à ceux qui persistaient à croire qu'un ordinateur n'était qu'une machine, offrant certes un potentiel très important, mais qui ne rivaliserait jamais, au moins dans l'état actuel de nos connaissances mathématiques, avec l'intelligence humaine. Je fais partie de cette deuxième école et mes vingt dernières années professionnelles n'ont fait que confirmer mon choix. Notons que cette divergence se retrouvait plus généralement en Intelligence artificielle.

Ces deux écoles ont conduit à des directions de recherche très différentes et abouties à deux technologies et donc à des produits totalement différents dans leur finalité. Constatons que la première est plutôt d'origine européenne et la

seconde américaine. Cette compréhension est essentielle pour s'y retrouver derrière la surenchère actuelle autour du mot «sémantique » et donc comprendre ce que l'on peut attendre de tel ou tel produit et surtout à quelles conditions. Ceci est particulièrement vrai dans le domaine de l'Intelligence économique où la matière première est essentiellement textuelle et où le T.A.L.N. a tout son sens. L'objet de cet article est précisément d'offrir un éclairage sur l'analyse sémantique appliquée à ce domaine.

## Une première école : décrire les langues naturelles

La première école, souvent européenne, décrit. C'est un principe incontournable qui se retrouvait dans les systèmes experts où la première opération à réaliser était de constituer la base de règles censée modéliser le domaine d'expertise tel que le maitrisait les spécialistes humains. Cette base était une « description a priori» de ce savoir.

Dans le domaine du T.A.L.N., ces modèles descriptifs correspondent aux quatre niveaux d'analyse décrits (voir encadré), l'objectif étant de les faire reproduire par un ordinateur. Le niveau lexical est pour l'essentiel maîtrisé. Les dictionnaires électroniques, héritant de la technologie des S.G.B.D., sont couramment employés dans les programmes informatiques à commencer par le traitement de textes qui a servi à rédiger cet article.

De la même façon, il existe plusieurs modèles de grammaire décrivant une langue ; l'analyse syntaxique automatisée a fait d'immenses progrès. Toutefois, elle ne peut être complètement maîtrisée que par une compréhension de la phrase voire du contexte. La phrase « La belle ferme le voile » reste un beau cas d'école pour l'analyste afin de déterminer le sujet et le verbe suivant que l'on a un point de vue agricole ou romantique ! Une analyse syntaxique ne suffit pas. Des résultats intéressants sont obtenus et il arrive parfois à mon correcteur grammatical de trouver des erreurs et de proposer des corrections justifiées. Parfois et certainement pas toujours.

L'analyse sémantique reste un domaine pour lequel il n'y a pas vraiment de standards ou normes. Il n'existe pas de modèle sémantique universel et général d'une langue exploitable par un ordinateur et nous sommes très loin d'en avoir un. Il n'existe que des modèles partiels répondant à des applications bien spécifiques : extraction de données (ou plus exactement de structures linguistiques), enrichissement de requêtes par génération de termes associés, etc. Réseaux sémantiques, graphes conceptuels, cartouches linguistiques, etc. sont des exemples de formalismes utilisés afin de décrire la sémantique d'un domaine particulier.

Ces modèles doivent la plupart du temps être particularisés aux besoins spécifiques d'un client. Sur un plan opérationnel et donc économique, la mise en œuvre de ces solutions se fait de la façon suivante: d'une part, par la fourniture de composants logiciels assimilables à des produits et d'autre part, par la réalisation de prestations de services visant à particulariser les composants linguistiques. Pour avoir commercialisé ce type de produits il y a quelques années, la facture finale se décomposait la plupart du temps en un tiers de « logiciels » et deux tiers de prestations de services. Cette remarque est essentielle car elle ne permet pas de développer une démarche industrielle qui repose toujours sur la mutualisation des coûts et donc la réplication de l'objet conçu et développé, ce qui n'est pas le cas ici compte tenu de la part essentielle des prestations de services.

Une deuxième conséquence est que ces solutions ne peuvent traiter que ce qu'elles connaissent. Ce sont des applications qui ne font que reproduire ce qui a été décrit dans leurs modèles. Chercher à découvrir ou à mettre en évidence de nouveaux concepts avec ce type de technologie est illusoire. De ce point de vue, on peut légitimement s'interroger sur l'opportunité d'utiliser ces solutions pour des applications de veille et plus généralement l'Intelligence économique dans la mesure où par définition un veilleur ne sait pas ce qu'il va trouver (sauf à penser que ses concurrents le préviennent à l'avance pour lui permettre de mettre à jour ses modèles linguistiques !).

De plus, l'évolution c'est-à-dire la maintenance de ces composants est significative notamment si le domaine modélisé change, ce qui est souvent le cas.

# Niveaux d'analyse

**Tout collégien a pratiqué l'analyse lexicale et l'analyse syntaxique (grammaticale) qui suivaient généralement l'épreuve de la dictée lors de ses cours de français.**

**Les deux autres niveaux peuvent se comprendre par des exemples. « La fenêtre mange des radis » est juste grammaticalement mais n'a aucun sens. L'analyse sémantique permet de le déterminer. « Le pont plia sous le poids de l'oiseau et du camion» pose un autre problème : si un pont peut plier sous le poids de quelque chose ou de quelqu'un, il est peu probable que cela soit sous le poids d'un oiseau comme me le laissent penser mes connaissances sur ce qu'est la résistance d'un pont et le poids d'un oiseau. Ce dernier niveau est bien du ressort d'une analyse pragmatique pour déterminer qui du camion et de l'oiseau fait plier le pont.**

**Le T.A.L.N. s'est développé en parallèle de l'Intelligence artificielle avec un âge d'or entre 1980 et 1990, à une époque où tout le monde a compris que les ordinateurs n'étaient pas que de simples calculateurs mais qu'ils pouvaient aussi réaliser des opérations que l'on considérait jusqu'alors réservées à l'homme et à son intelligence.**

Les applications reposant sur un modèle purement linguistique (première école) ne font que reproduire ce qui a été décrit dans leurs modèles. Chercher à découvrir ou à mettre en évidence de nouveaux concepts avec ce type de technologie est illusoire. De ce point de vue, on peut légitimement s'interroger sur l'opportunité d'utiliser ces solutions pour des applications de veille et plus généralement l'Intelligence économique dans la mesure où par définition un veilleur ne sait pas ce qu'il va trouver (sauf à penser que ses concurrents le préviennent à l'avance pour lui permettre de mettre à jour ses modèles linguistiques !).

## Une deuxième école utiliser l'ordinateur pour ce qu'il sait faire

Et la deuxième école ? Si la première décrit, la seconde calcule. Et les axes de recherche ont été totalement différents. En particulier, il n'y pas eu la volonté d'automatiser les quatre niveaux d'analyse (voir encadré). La promesse est totalement différente et correspond à l'application d'algorithmes issus de la statistique ou créés de façon ad hoc pour réaliser des opérations d'analyse des données textuelles qui vont faciliter leur exploitation ultérieure.

Cette algorithmique est dépendante de l'application considérée : un logiciel développé pour un spécialiste de l'Intelligence économique n'exploitera pas les données textuelles sous-jacentes de la même façon que le fera un produit de traduction. En revanche, elle ne dépend pas du client considéré et ne nécessite pas de prestations de services particulières.

Cette voie a été empruntée dès les années 70 par les chercheurs spécialistes de reconnaissance vocale. Les technologies actuelles notamment implémentées dans les produits Dragon NaturallySpeaking (1) reposent sur des composants statistiques (2) construits au préalable sur de grands corpus de textes auxquels sont couplés des dictionnaires. Il n'y a pas d'analyse syntaxique, sémantique… mais simplement un calcul probabiliste qui conduira le système à proposer par exemple la suite « je vais aller » avec un infinitif et non un participe passé pour le troisième mot parce que cette suite aura été observée dans les corpus traités. A aucun moment, il n'y aura eu exploitation de connaissances linguistiques particulières, règles grammaticales ou autres. Remarquons aussi que la même technologie est utilisée pour toutes les langues – avec plus ou moins de bonheur compte tenu du taux d'homophonie propre à chaque langue – et que la puissance de calculs des ordinateurs, en croissance constante, a permis au fil du temps de disposer de composants statistiques généraux applicables à de nombreux domaines.

Cette démarche a conduit à des recherches particulières pour des applications d'Intelligence économique en partant des besoins spécifiques à ces disciplines. Nous l'évoquions précédemment, une des spécificités essentielles de l'activité de veille et donc d'Intelligence économique est que l'expert sait rarement ce qu'il va trouver (autrement, il serait plus judicieux de parler de recherche documentaire). Par conséquent, les éditeurs de logiciels, dont la société qui m'emploie, ont plutôt travaillé sur la mise au point d'algorithmes visant à établir des corrélations entre termes et à en étudier l'évolution dans le temps afin de détecter des informations de rupture et de mettre en évidence des tendances plus ou moins significatives. De la même façon, il est intéressant pour un analyste de connaître pour un corpus de textes quelles sont les « idées principales » notamment les expressions les plus significatives et de les mettre en évidence. Il est aussi possible par des algorithmes particuliers de détecter des noms de personnes, de lieux, d'organisations, … toute structure linguistique ayant une construction régulière.

Est-ce de la sémantique ? Sur un plan strictement linguistique, tel que nous l'avons rappelé au départ, certainement pas. En revanche, la réponse est positive si « analyse sémantique, » signifie analyser un contenu afin de dégager le sens qu'il véhicule pour le veilleur. Et ce type d'outils produit des analyses sémantiques de qualité quotidiennement exploitées par des experts de très haut niveau. Il me semble que c'est plutôt cet objectif qui est recherché, et en tout état de cause, celui qui intéresse les professionnels concernés.

Un des intérêts majeurs de la deuxième approche est qu'elle n'impose pas d'adaptations préalables à un client particulier. Sur un plan économique, le même produit est diffusé à tous et nous sommes dans une logique d'éditions de logiciels où la prestation de services correspond à l'installation du produit et aux éventuelles formations, et rien d'autre. Dans le cas présent, la facture correspondra plutôt à 80% à des licences de logiciels et 20% à des prestations de services avec une garantie de pérennité sans commune mesure.

## Notes

(1) http://www.nuance.fr/naturallyspeaking/

(2) Le terme « modèle de langage » est souvent employé par les spécialistes ; il a volontairement été omis afin de ne pas introduire de confusion par rapport à l'emploi du terme « modélisation » au sens « description » fait dans cet article.

# Et la performance ?

Il est évident qu'un système ayant fait l'objet d'une modélisation très poussée sur un domaine bien spécifique aura toujours une performance nettement supérieure à celle d'un système généraliste. Par performance, nous entendons capacité à analyser avec justesse des données textuelles par exemple pour détecter une structure particulière. Il y a peu de temps, un acteur du domaine présentait une application dans le domaine des ressources humaines, sur le traitement des petites annonces de cadres, dont la performance était à l'évidence de très bon niveau.

D'une part, celle-ci était obtenue après une mise au point de composants linguistiques réalisée en commun avec le client, travail non négligeable. D'autre part, il s'agissait d'un domaine sémantique très particulier (« une petite annonce ») avec une rédaction des textes et donc un lexique et une grammaire tout aussi particuliers.

Il n'en demeure pas moins que ces technologies réclament une puissance de calculs importante qui, dans le cas d'applications d'Intelligence économique, peut s'avérer rédhibitoire. En effet, ces applications nécessitent le traitement de très gros volumes de données qui sont progressivement analysées afin de valider leur niveau de pertinence par rapport à la thématique suivie. Ainsi, il est fréquent d'avoir des flux de plusieurs milliers de documents provenant d'automates de collecte. Or, comme nous avons pu le constater dans plusieurs contextes opérationnels, une telle volumétrie demande des temps de calculs qui se comptent en heures pour des analyseurs linguistiques, et sont donc incompatibles avec les besoins d'alertes propres à la veille.

## Une nouvelle voie

Une approche médiane se dessine visant à combiner les deux approches. En complément de calculs purement statistiques, des composants linguistiques sont exploités. Dans une utilisation standard, ceux-ci ne nécessitent pas d'adaptations particulières à un client. L'approche industrielle décrite précédemment est donc conservée.

Toutefois, si un client le souhaite ou si un contexte opérationnel particulier le nécessite, il est toujours possible d'augmenter le niveau de particularisation de ces composants via des prestations de services.

Pour reprendre la terminologie américaine fréquemment utilisée dans le domaine de l'édition de logiciels, la première approche correspond à du « sur-mesure » (tailored made), la deuxième à du « sur étagère » (off the shelves), l'approche médiane étant qualifiée de « one of the kind ».

# Qui sera le gagnant ?

L'histoire industrielle des deux derniers siècles a déjà répondu. D'un coté, des produits industriels facilement réplicables, amenant à une mutualisation des coûts de développements, gigantesques dans l'industrie du logiciel, donc à une économie d'échelle. De l'autre, des solutions adaptées au cas par cas, dont le prix intègre un coût humain lié à la complexité de la solution d'un client, où les économies d'échelle ne se situent que dans l'espoir de pouvoir réutiliser pour l'un ce qui a été fait pour l'autre.

Il est probable que les deux approches cohabiteront mais que la première aura une dimension économique sans commune mesure avec la seconde. Le lecteur scientifique sera peut-être perturbé par mes « comparaisons de factures » et mes propos de comptable exprimés tout au long de cet article. Je pense qu'en dernière analyse ceux-ci seront déterminants.

Je constate par ailleurs que l'Intelligence artificielle a suivi la même voie. La vieille quête de « l'ordinateur aussi intelligent que l'homme » est terminée mais des algorithmes issus de ces travaux de recherche sont en exploitation courante dans de nombreux logiciels. De ce point de vue, l'approche que nous avons qualifiée de médiane a sans doute un bel avenir dans la mesure où elle s'inscrit dans une démarche strictement industrielle.

Alain Beauvieux
Docteur en informatique,
Université Pierre & Marie Curie
Président d'AMI Software