"la belle ferme le voile"

Oui, la sémantique est aujourd'hui mature

par Bernard Normier

L'article intitulé « Il y a sémantique et ... sémantique » paru dans ces mêmes colonnes (Cf. Veille Magazine N° 116) pose bien des questions. En effet le mot « sémantique » tend à être galvaudé et on ne sait plus bien ce qu'il recouvre.

débat

L'article fait clairement la distinction entre deux écoles historiques ; celle basée sur la représentation des connaissances (celle qui décrit) et la seconde basée sur des méthodes statistiques (celle qui calcule). Et, sur cette distinction, nous sommes bien d'accord.

A LA RENCONTRE DE DEUX MONDES

Pour ma part, la société que je dirige, ou tout du moins son équipe R&D, appartient historiquement à la première école. Et comme la plupart des spécialistes du TAL (Traitement Automatique des Langues), nous sommes adeptes des deux mondes et combinons à la fois la représentation des connaissances et statistiques pour le plus grand bénéfice des utilisateurs.

Mais au-delà de ce constat, il me semble important d'apporter un certain nombre de précisions, qui tiennent en 6 points.

1• Tout d'abord, tordons définitivement le cou à « la belle ferme le voile ». Ce jeu de linguiste, qui consiste à démontrer que certaines phrases sont intrinsèquement ambigües et donc ne peuvent pas être traitées par une machine est tout-à-fait marginal.

2• Sur la question des normes et des standards et en dépit des efforts du W3C, la situation est il est vrai loin d'être stabilisée. Quoi qu'il en soit, ce n'est pas parce qu'il y a peu de normes que l'on doit s'empêcher d'utiliser ces techniques, et c'est au contraire leur généralisation qui va permettre l'émergence des normes.

Ne prendre en compte que des techniques éprouvées et des modèles normalisés serait une démarche tuant dans l'œuf toute innovation, alors que notre domaine est un de ceux où la recherche est la plus active et prometteuse.

3• Selon l'article cité, le montant de la facture finale d'une application sémantique et l'absence de produits logiciels basé sur des technologies linguistiques réutilisables serait prohibitif, du fait d'un très (trop) important effort d'adaptation à chaque contexte applicatif. Si l'affirmation était sans doute vraie il y a 10 ou 15 ans, ce n'est plus le cas aujourd'hui.

Justement, la rupture technologique apparue ces dernières années dans le domaine du traitement automatique des langues et amenée par plusieurs acteurs du domaine, a permis la mutualisation nécessaire à toute démarche industrielle. Cela passe

notamment par la construction de bases de connaissances, qu'on appellera selon les écoles ontologies, référentiels sémantiques, réseaux sémantiques, dictionnaires sémantiques … qui permettent un degré de généralisation et d'abstraction suffisant pour que les connaissances linguistiques et sémantiques amassées soient réutilisables d'une application à l'autre. Il en résulte alors un cercle vertueux, qui permet de réduire considérablement les coûts.

4. Oui, les méthodes statistiques et linguistiques sont complémentaires, mais elles ne servent pas à la même chose. Pour prendre un exemple volontairement (très) simplificateur, une méthode statistique basée sur des co-occurrences pourra, après analyse d'une grande masse de textes, « découvrir » un « descripteur » comme « augmentation de capital ».

Dans une approche plus descriptive, on saura dans une base de connaissances, que le terme « augmentation de capital » est formé d'un nom, d'une préposition et d'un nom, que « augmentation » est la forme nominale de « augmenter », que « renforcer » est un verbe qui dans certains contextes a le sens de « augmenter » et que donc une phrase comme « la société a décider de renforcer prochainement son capital » peut être indexée par « augmentation de capital ».

Les méthodes statistiques auront beaucoup de mal à atteindre le même résultat. En revanche, elles seront un excellent outil pour aider à fabriquer ces bases de connaissances.

5. En fait, le coût d'une application sémantique dépend de 3 dimensions : la taille de l'univers de l'application, la complexité des corpus considérés et la finesse de l'analyse sémantique souhaitée. Il est important pour tout utilisateur qui s'intéresse aux moteurs sémantiques de savoir se situer sur ces trois axes, comme l'illustre les 3 exemples d'applications opérationnelles ci-dessous :

Dans le domaine des ressources humaines, un logiciel d'analyse de CV ou d'offres d'emploi s'appuiera sur une analyse linguistique et sémantique très poussée. Cependant, comme le type de documents (un CV) et l'univers de référence (la description de personnes, de postes, dans une logique de RH) sont très limités, on peut développer une application parfaitement duplicable et proposer au marché une offre en SaaS pour quelques centaines d'euros par mois.

Dans d'autres applications, la taille de l'univers est beaucoup plus grande, mais le type de documents traités reste homogène, et par la nature des informations à extraire reste limitée : ainsi dans le domaine de la veille en Propriété Intellectuelle, on peut assez facilement modéliser l'information qu'on doit trouver dans des brevets (les inventeurs, les avantages de l'invention, les technologies, les « revendications indépendantes » …) et construire des systèmes d'analyse performants pour ce type de tâche. La solution sera alors parfaitement duplicable et industrialisable dans son contexte propre.

Dans d'autres cas enfin, la complexité du corpus et la taille de l'univers de référence seront trop grands et on devra réduire le niveau de modélisation visé, ou combiner une approche automatique avec une expertise humaine pour obtenir des résultats. C'est par exemple le cas dans le domaine du marketing, et plus précisément de la « e-Réputation ». Dans ce domaine, l'univers de référence est souvent très vaste et multilingue, et la nature des textes très hétérogènes par leur provenance (blogs, twitts, sites de presse, forums …), leur niveau de langue (journalistique, professionnel, argotique …) ou leurs formats.

Pour autant, l'analyse sémantique basée sur l'analyse linguistique rendra de grands services, pour peu que l'on dispose de bases de connaissances linguistiques générales, et que l'on définisse clairement au départ ce que l'on cherche : des opinions sur des marques, des sociétés, des personnes, des avis consommateurs …

6. Enfin, l'argument selon lequel il n'est pas besoin d'utiliser des connaissances propres à une langue ou au domaine, dans un système automatique, à plus forte raison dans un système de veille, au motif qu'un veilleur ne sait pas ce qu'il va découvrir, est à notre avis un prétexte que mettent en avant ceux qui n'ont pas ces connaissances à leur disposition. C'est comme si, pour constituer une équipe de veille, on choisissait de préférence de parfaits étrangers à la langue et au domaine de l'entreprise dans l'espoir que ce dispositif coûte moins cher. C'est une conception pour le moins surprenante à la fois du métier de la veille et de la notion de retour sur investissement.

Ne pas savoir ce qu'on va découvrir n'empêche pas d'avoir une idée de ce que l'on cherche !

Bernard Normier
Président de Lingway
www.lingway.com

Président de Lingway, Bernard Normier est titulaire d'un doctorat en Informatique de Paris IX. Fondateur d'ERLI en 1977, société pionnière du Traitement Automatique du Langage Naturel (TALN), devenue ensuite LexiQuest, il a participé à de nombreux projets couvrant la plupart des technologies et domaines d'applications du TALN : indexation et recherche, bases structurées ou non structurées, traduction et multilinguisme, dictionnaires et parsers, etc.

Il est l'auteur aux Editions ADBS d'un ouvrage de référence intitulé "L'apport des technologies linguistiques au traitement et à la valorisation de l'information textuelle".