FÉVRIER 2006

numéro 11

Actualité de l'Institut de Recherche en Informatique de Toulouse

noir SUR blanc

Le Traitement Automatique du Langage Naturel

Jacques Durand

L'informatique a partagé un grand nombre de concepts avec la linguistique dès ses débuts, ainsi par exemple la notion de grammaire formelle qui a été un outil essentiel aussi bien pour la mise en place des analyses syntaxiques des langues naturelles que pour la définition et la compilation des langages de programmation.

Les champs d'interaction entre les deux disciplines n'ont fait que progresser ces dernières années dans des thématiques très variées, comme la formalisation du dialogue, les systèmes de communications dégradées, ou la recherche d'informations dans de très grandes masses de données textuelles.

À l'IRIT, le domaine interdisciplinaire du traitement automatique des langues (TAL), est un thème majeur pour le laboratoire, aussi bien pour ses enjeux scientifiques que pour les possibilités qu'il offre en matière de collaborations fortes avec d'autres instituts de recherche de la région comme l'Institut des Sciences du Cerveau de Toulouse ou l'ERSS (Equipe de Recherche en Syntaxe et Sémantique).

Dans ce numéro, Patrick Saint-Dizier nous expose quelques défis importants du TLN (Traitement du Langage Naturel) et notre invité Jacques Durand (directeur de l'ERSS) nous fait partager les enjeux scientifiques de l'interdisciplinarité en TLN tels qu'ils apparaissent en linguistique.

Les rapports entre la linguistique et l'informatique sont voués à continuer à se développer et l'IRIT a des atouts majeurs à jouer dans les relations entre ces deux disciplines.

Luis Fariñas del Cerro

IRIT CNRS INPT UPS UTI
Institut de Recherche en Informatique de Toulouse

L'IRIT est associé au CNRS, à l'INPT, à l'UPS et à l'UT1

118 Route de Narbonne
31062 Toulouse cedex 9
tél. 05 61 55 67 65
fax 05 61 55 62 58
info@irit.fr - www.irit.fr

Conformément à sa politique de valorisation, l'IRIT participe avec le LAAS-CNRS et l'ONERA à un nouveau laboratoire commun avec la société AIRBUS France qui porte pour nom AIRSYS. La signature de la convention liant les partenaires interviendra dans le courant du mois d'avril 2006.

Airbus France a pour vocation de concevoir et valider des systèmes aéronautiques innovants. Pour cela, il doit disposer des méthodes et techniques les plus avancées dans ce domaine. L'IRIT au travers de plusieurs équipes de recherche prendra sa part dans cette collaboration aux côtés de partenaires qu'il a l'habitude de côtoyer dans le cadre de la structure fédérative FéRIA.

Les activités menées dans le cadre de la coopération AIRSYS seront des activités de recherche et de technologie qui concerneront notamment : la veille scientifique et technique, le développement de méthodes et outils, l'identification et l'évaluation de nouvelles technologies, l'étude de nouveaux concepts d'architectures et la réalisation de démonstrations ou démonstrateurs échelle laboratoire.

Directeur de la publication
Luis Fariñas del Cerro

Directeur adjoint de la publication
Jean-Luc Soubie

Secrétariat de rédaction
Véronique Debats, Katalyn Sangla

Comité de rédaction
Régine André-Obrecht, Vincent Charvillat, Olivier Gasquet, Jean-Pierre Jessel, Mustapha Mojahid, Gérard Padiou, Pascal Sainrat, Patrick Sallé, Jean-Luc Soubie, Jacques Virbel

Maquette Allard & Création

Contact de la rédaction
05 61 55 65 10 – nsb@irit.fr

ISSN 1620-2937

Le traitement automatique du langage naturel : enjeux et problématiques

Le traitement automatique du langage naturel, par les liens profonds qu'il entretient avec les sciences du langage, l'informatique, la logique, les sciences cognitives, a vu, depuis une décennie, une prolifération considérable de problématiques, de techniques et de champs applicatifs.

Le traitement automatique du langage naturel (TALN), en tant que problématique scientifique et technique à part entière, mais aussi de par les liens profonds qu'il entretient avec les sciences du langage, l'informatique, la logique, les sciences cognitives, pour ne citer que les plus importants, a vu, depuis une décennie, une prolifération considérable de problématiques, de techniques et de champs applicatifs. La communauté scientifique du TALN est passée d'un groupe d'artisans à une large communauté.

Nous abordons ici l'évolution du langage naturel telle que nous l'avons vécue sur 25 ans, et situons quelques enjeux intellectuels actuels. Nous avons assisté à de nombreuses mutations technologiques bien entendu, mais aussi sociologiques (l'arrivée d'Internet, par exemple) et culturelles. Ces mutations se sont déroulées avec beaucoup d'hésitations, d'aller et retours et de contradictions, probablement en raison de la complexité et de la pluridisciplinarité très marquée de cette discipline.

La communauté du TALN a un peu de mal à se définir. Les acteurs du TALN sont en très grande majorité des informaticiens attirés par le traitement des langues. En France, on y trouve une proportion beaucoup plus réduite de linguistes (alors que de mon point de vue, ce devrait être l'inverse, aussi bien en linguistique descriptive que formelle), de collègues de l'intelligence artificielle, et encore moins des intervenants des disciplines théoriques de l'informatique (langages, grammaires et automates, lambda-calcul, théorie des types, etc.).

À l'heure actuelle, on observe deux tendances qui vont en s'éloignant quelque peu l'une de l'autre. D'une part, une tendance applicative, sans que ceci soit péjoratif, préoccupée d'expérimentations, pour développer des travaux à finalité pré-industrielle, avec un niveau de qualité défini

a priori. C'est, par exemple, le cas de nombreux travaux en recherche d'information. L'autre tendance, plus ancienne mais minoritaire, bien qu'en progression à nouveau, est plus proche de la linguistique et de l'intelligence artificielle. Elle vise aussi à développer des éléments applicatifs, mais par le truchement d'une analyse plus fine et plus globale des comportements linguistiques, avec un niveau explicatif et d'abstraction adéquat. Entre ces extrêmes, bien entendu, on trouve des systèmes hybrides, qui tentent d'organiser ces points de vue.

Les champs d'application

L'un des tout premiers domaines d'application du TALN a été la traduction automatique. C'était un défi immense, qui a monopolisé un nombre conséquent de chercheurs et de lexicographes. Plusieurs approches ont été testées : traduction mot à mot, peu performantes même sur des langues proches, traduction par transfert (réécriture d'arbres syntaxiques d'une langue à l'autre), qui a donné des résultats d'une certaine qualité, mais qui s'est révélé d'un coût élevé au niveau des ressources, traduction autour d'un pivot sémantique *a priori* indépendant de la langue, très intéressant scientifiquement, mais complexe au niveau de la sémantique. D'autres approches ont été tentées : traduction sur une base statistique avec apprentissage sur corpus multilingues alignés, systèmes d'aide à la traduction, etc. À ce jour, plusieurs systèmes de traduction automatique sont disponibles, ils produisent des résultats acceptables, mais qui demandent malgré tout une relecture et des corrections à réaliser par un traducteur professionnel.

Un autre grand champ applicatif gravite autour de la recherche d'informations, des systèmes question-réponse et du dialogue. Les applications en sont innombrables : extraction de connaissances, veille technologique, résumé, etc. Contrairement à la traduction automatique qui doit donner d'emblée des résultats parfaits, ces domaines peuvent être structurés en étapes, chacune apportant un résultat qui peut être évalué, et qui s'inscrit dans une progression vers un objectif idéalisé. La plupart des techniques utilisées ont une base statistique, cependant, un besoin de raisonnement commence à apparaître, par exemple en question réponse, qui va faire évoluer ces travaux vers une base symbolique. Un autre aspect qui va dans ce sens est l'introduction de connaissances pour améliorer la recherche de l'information : connaissances ontologiques, connaissances du domaine, connaissances grammaticales sur des points particuliers.

D'autres champs applicatifs sont à souligner : le dialogue interactif, les systèmes de communication dégradée, l'analyse de compte-rendus, l'aide à la rédaction (orthographe aussi bien que contenus), les systèmes tutoriaux pour apprendre une langue, etc. Les supports évoluent aussi : de l'ordinateur classique, ces supports vont vers l'oral, les supports mobiles, les bornes interactives, etc. Le langage est aussi parfois associé à la parole, l'image ou le geste, pour plus de clarté, d'efficacité et d'attrait.

Quelques défis du TALN

Le TALN se trouve devant un paradoxe étonnant. S'il a su développer des technologies avancées (par exemple en matière de stratégies d'analyse ou de développement de ressources lexicales), il doit faire face à plusieurs problèmes essentiels qu'il n'a pas su résoudre et qui sont un frein considérable à son développement.

Citons par exemple les restrictions de sélection : il s'agit de préciser, par exemple, pour un verbe, le type des différents compléments

qu'il peut prendre. S'il est simple de dire que dévorer a un sujet *a priori* animé et un objet de type comestible, ce que stipulent les lexiques habituels, il est aisé de trouver d'autres cas acceptables dont il faudra analyser la signification (dévorer des romans). Les usages réguliers des prédicats peuvent être caractérisés par les nœuds d'ontologies, cependant, de telles ontologies restent à construire et à valider sur une grande échelle, et les restrictions restent aussi à définir, manuellement ou par apprentissage à partir de corpus annotés.

Citons aussi le problème des références qu'elle soient pronominales (références à des entités), temporelles ou spatiales. Les textes et dialogues abondent en références, les humains savent instinctivement les résoudre, une machine a beaucoup plus de difficultés. Imaginons la difficulté de comprendre le lien entre Jean et « le malheureux » dans : Jean est tombé dans l'eau, le malheureux grelottait de froid.

Dans un autre ordre d'idées, notant que le développement de ressources linguistiques, essentiellement lexicales, représente environ 70% du coût total d'un projet, il paraît naturel de tenter de regrouper et de partager des ressources développées ici et là pour une langue donnée. Malheureusement, ce n'est pas si simple. Autant on peut partager des listes de mots étiquetés de leur catégorie, autant le partage de données un tant soit peu plus élaborées est problématique. Ces données ont été, en effet, souvent conçues dans un cadre théorique précis, ou bien elles ont fait l'objet de décisions de catégorisation propres à un groupe de personnes, ou bien de découpages selon certaines granularités. La mise en commun est certes possible, mais les ré-utilisations sont très rares. Des projets tentent de normaliser les descriptions, mais en dehors d'éléments extrêmement simples, il n'en sort que peu de résultats. Ce n'est pas une raison pour ne pas persévérer.

Le TALN et les autres disciplines

Le TALN a des liens soutenus avec de nombreuses disciplines. Il n'est pas une sorte de valorisation ou d'application de ces autres disciplines. Il y a toujours une fécondation croisée.

Par exemple, le TALN exige des descriptions linguistiques complètes par rapport à un objectif : on ne peut se contenter de fragments d'un phénomène comme on le fait en linguistique lorsque l'on traite d'un point particulier. Il y a exigence de synthèse, d'homogénéité et de « complétude ». Dans chaque cas, le TALN doit opérer des reformulations souvent simplificatrices pour répondre à des exigences de robustesse et de faisabilité.

Le TALN est aussi un immense champ d'expérimentation pour l'Intelligence Artificielle, par l'observation de comportements naturels, l'IA apportant en retour un savoir faire et des bases théoriques solides. On peut en dire autant de tout l'informatique.

En conclusion, le TALN, malgré environ 40 ans d'existence est encore une discipline jeune, à la croisée de plusieurs disciplines, avec des enjeux gigantesques. En France, il a su se doter d'une conférence annuelle (TALN), d'une association et d'une revue (ATALA) de niveau international. Les défis, on l'a vu, restent très complexes. La communauté industrielle étant déficiente celle-ci ne joue pas son rôle : à notre communauté scientifique d'aller de l'avant.

Patrick Saint-Dizier

L'ERSS, un partenaire privilégié de l'IRIT pour le traitement du langage naturel

entretien avec
Jacques Durand

L'Équipe de Recherche en Syntaxe et Sémantique (ERSS) a la particularité de compter à la fois des linguistes et des informaticiens, comment est organisée cette pluridisciplinarité ?

À vrai dire, en dehors d'un spécialiste d'informatique, les collègues qu'on pourrait qualifier d'informaticiens sont de véritables linguistes. Mon sentiment est qu'ils représentent un nouveau type de linguiste capable de renouveler les données et de vérifier des hypothèses en utilisant les moyens que nous offre l'informatique. Je dirais donc que, en interne, nous travaillons plutôt dans l'interdisciplinarité que dans la pluridisciplinarité. L'interdisciplinarité est un fait acquis en linguistique puisque l'on décrit souvent nos recherches comme faisant partie des sciences du langage. J'ai quelques réserves sur la dénomination « sciences du langage » mais elle a l'avantage de souligner la diversité des angles d'attaque et de rappeler que les faits langagiers n'appartiennent pas de façon exclusive au linguiste professionnel.

Mais alors pourquoi exprimer des réserves sur la notion de sciences du langage ?

Cette dénomination est en adéquation avec la diversité des approches et je ne me prive pas de l'utiliser car il peut y avoir un véritable abîme entre, par exemple, le travail du phonéticien expérimental et une sémantique du langage basée sur le lambda-calcul. Cependant, si on estime que la tâche d'une théorie du langage et des langues est au final une modélisation du rapport son-sens, il peut être dangereux de poser dès le départ l'hétérogénéité des domaines. Je me réjouis par exemple du fait que des sémanticiens comme N. Asher, qui a récemment rejoint l'IRIT, se penchent de près sur les phénomènes prosodiques (rythme, accentuation, intonation) pour mieux comprendre la structuration des énoncés. Dans ce cas, deux domaines apparemment hétérogènes se rejoignent et se fertilisent mutuellement.

À travers votre réponse, vous évoquez indirectement la collaboration entre l'IRIT et l'ERSS puisque N. Asher travaille de façon très étroite avec des membres de votre unité de recherche. Pourriez-vous dire quelques mots sur l'évolution de cette collaboration en donnant peut-être quelques éléments d'histoire ?

À ma connaissance, les relations entre nos deux laboratoires ont commencé dès la création de l'ERSS sous la coordination de Andrée Borillo au début des années 80 et l'installation à l'IRIT d'une équipe autour de Mario Borillo. Pendant plusieurs années, ces deux équipes, de petite taille, se sont renforcées mutuellement, tant du point de vue pratique, par un soutien logistique et l'hébergement de doctorants, que scientifique, par des collaborations et des aides réciproques sur des sujets ayant trait à la sémantique formelle et à la linguistique. Pendant tout ce temps, les relations suivies entre les deux laboratoires ont été renforcées par la venue à Toulouse de chercheurs étrangers accueillis sur des postes « rouges » du CNRS à l'IRIT ou à l'ERSS. La création de PRESCOT, dans les années 90, a conforté ce rapprochement sur des problématiques communes au sein des programmes de recherche mis sur pied dans le cadre de cette structure interuniversitaire toulousaine. En particulier, plusieurs programmes-phares sur la sémantique du temps et de l'espace ont été conduits conjointement pendant quelques années. Cette collaboration, suite à divers départs, s'était un peu distendue. La nomination toute récente de N. Asher à l'IRIT et le retour à l'ERSS de M. Aurnague courant 2006 me rendent très optimiste sur l'avenir de la sémantique à Toulouse mais il faut souligner que ce n'est pas le seul axe collaboratif entre nos unités.

Avant de passer aux autres domaines, pourriez-vous néanmoins préciser la nature de cet axe collaboratif ?

Pour faire court, disons que notre opération « Sémantique et discours », coordonnée par A. Le Draoulec et J. Busquets (Bordeaux), compte désormais dans ses rangs un groupe de spécialistes (dont M. Bras, F. Cornish, M.-P. Péry-Woodley, J. Rebeyrolle) qui peuvent mener avec les collègues de l'IRIT, comme P. Muller, une réflexion de fond sur la sémantique de l'espace et du temps, le discours et le dialogue. Le travail effectué à l'ERSS en collaboration avec l'IRIT sur l'annotation temporelle des textes, les connecteurs temporels et

[suite page 12]

Nicholas Asher rejoint l'IRIT

Je suis nouveau directeur de recherche CNRS en Cognisciences. Je travaille essentiellement dans la formalisation de la sémantique et la pragmatique du langage naturel. Je m'intéresse aussi à la logique formelle, et j'ai publié sur la logique non monotone, la logique modale, la logique des croyances et de l'intention, la théorie des jeux, le vague et les paradoxes de vérité.

Depuis une quinzaine d'années je me suis concentré sur le problème de l'interprétation du discours. Ma démarche a été de développer la sémantique dynamique, où le contenu d'une phrase n'est pas comme dans la sémantique intentionnelle de Montague traditionelle un ensemble simple de mondes mais plutôt une relation entre des paires de mondes et des fonctions d'assignation, afin qu'elle puisse refléter les effets de la structure discursive d'un texte.

Chaque lecteur se rend compte que dans un texte cohérent chaque clause joue un rôle rhétorique ou discursif dans le texte : soit il explique quelque chose qui a déjà été dit, soit il continue un récit narratif, soit il entre plus en détail sur un sujet déjà mentionné dans le texte. Mes collaborateurs et moi avons montré que ces rôles rhétoriques et la structure discursive en général, ont des effets primordiaux sur le contenu du discours, particulièrement sur l'interprétation des expressions anaphoriques comme les pronoms ou les descriptions définies, mais aussi sur l'interprétation des temps verbaux, les modalités, les quantificateurs et la sémantique lexicale (le sens des mots) en général.

Mes projets actuels se divisent en trois. Le premier est de décortiquer les relations complexes entre les sens des mots et leur sensibilité au contexte discursif. Par exemple, beaucoup de chercheurs en sémantique lexicale ont remarqué que le sens d'un mot peut changer suivant le contexte de la prédication. Par exemple si je dis « je viens de commencer une cigarette » tout le monde comprendra qu'il s'agit (normalement) de fumer la cigarette. Mais on observe que souvent ce changement de sens n'est pas local au prédicat et son argument. Si je dis, « Julie a commencé par la cuisine » il n'est pas très clair de quoi l'on parle, sauf si on le met dans un contexte discursif particulier : « hier, Julie a nettoyé sa maison. Elle a commencé par la cuisine ». Dans ce contexte on comprend parfaitement de quoi il s'agit : elle a commencé par nettoyer la cuisine. L'analyse de cette sensibilité au contexte du sens m'a amené à développer une théorie qui combine une approche formelle de l'interprétation du discours avec une théorie formelle de la prédication et des types sémantiques.

Le deuxième projet est de développer des expériences computationelles pour vérifier certaines hypothèses fait dans ma théorie du discours, la SDRT (Segmented Discourse Representation Theory). Je continue à diriger un projet NSF (National Science Foundation) sur le développement d'un analyseur du discours qui utilise des méthodes statistiques (Maximum Entropy). L'hypothèse à vérifier est que la structure du discours peut dramatiquement aider le repérage automatique des antécédents d'une expression anaphorique. À l'IRIT, j'ai l'intention d'étendre ce projet pour voir si on peut améliorer le traitement automatique de la structure temporelle d'un texte. Le troisième projet est de continuer à approfondir l'étude du dialogue dans la SDRT, particulièrement les liens entre la prosodie, le contenu sémantique et les intentions inférables des interlocuteurs.

Je suis très content d'être à l'IRIT au sein de l'équipe LILaC. L'IRIT dispose d'une équipe très forte et internationalement reconnue en logique appliquée à la modélisation du raisonnement et de l'action. Et en particulier, je compte collaborer avec les chercheurs sur la logique des intentions pour approfondir le modèle SDRT des intentions dans le dialogue. De plus, il y a déjà à l'IRIT et à l'Université du Mirail dans l'UMR ERSS une expertise sur le discours et en particulier sur la SDRT, respectivement Laure Vieu et Philippe Muller à l'IRIT et Myriam Bras et Anne le Draoulec à l'ERSS.

J'estime que mes collaborations scientifiques à l'IRIT, à Toulouse, en France et enfin en Europe seront très fructueuses et plaisantes.

Nicholas Asher

Le traitement du langage naturel à l'IRIT

Des équipes

CSC Conception de Systèmes Coopératifs

DIAMANT Dialogue, InterAction, Multimodalité, Accessibilité, Nouvelles Technologies

ILPL Informatique Linguistique et Programmation Logique

LILaC Logique, Interaction, Langue et Calcul

SAMOVA Structuration, Analyse et MOdélisation de documents Vidéo et Audio

SIG Systèmes d'Informations Généralisés

L'IRIT a une activité de longue date en TALN (Traitement Automatique du Langage Naturel) avec des thèmes singuliers autour de la sémantique, de l'extraction de connaissances et de la recherche d'informations, du discours, du dialogue et du traitement de la parole. Ces recherches associent les traitements et modèles de la langue avec ceux du raisonnement, de la représentation des connaissances, des mathématiques et des sciences cognitives. L'IRIT s'est aussi inséré dans un maillage pluridisciplinaire riche où l'informatique constitue à la fois une fin en soi et un outil performant, en particulier au service des sciences humaines et du grand public.

Dialogue

De nombreux travaux à l'IRIT s'attachent à la modélisation de la communication langagière homme-machine, ou homme-homme. Au carrefour de la sémantique et de la pragmatique, la modélisation des stratégies de dialogue, des intentions et croyances en contexte sont au centre des préoccupations des chercheurs.
Les différentes équipes se sont récemment coordonnées pour expérimenter ces approches sur la notion de dialogue « correctif » (dialogue en situation d'erreur).
www.irit.fr/projets/DIALOGUE.html

Parole

Le langage est également étudié sous sa forme orale, essentiellement dans une perspective d'indexation automatique de documents audio et vidéo. Le traitement automatique de la parole et des sons s'appuie sur des chaînes de traitements stochastiques pour la segmentation, une modélisation différenciée (associant des connaissances de niveau supérieur à chaque classe de sons) et une étude de la prosodie.

Traitement du texte, analyse de corpus et modélisation de connaissances

Le traitement du langage s'accompagne de la constitution de corpus de données diverses en quantité croissante, ce qui génère de nouvelles approches tout en posant des problèmes d'annotation des informations pertinentes dans de tels corpus. Plusieurs équipes de l'IRIT sont impliquées dans des efforts d'analyse et de normalisation pour le traitement des textes, par la définition de méthodes et de logiciels d'analyse

syntaxique, lexicale ou sémantique. Certains de ces traitements débouchent sur l'identification d'éléments de connaissances (concepts et relations sémantiques), sur la construction de terminologies et d'ontologies. D'autres visent l'étude des dialogues oraux, retranscrits ou en langue des signes (annotations prosodiques, des actes de langage, des gestes, etc.).

Recherche d'informations

La recherche d'informations sur le web se manifeste sous des formes toujours plus variées, dont la plupart font l'objet de recherches à l'IRIT : systèmes de questions-réponses, moteurs de recherche généraux, veille technologique, exploration de collections pour des utilisateurs ciblés, etc. La recherche de documents pertinents en adéquation avec des besoins en information précis s'appuie sur des techniques et des ressources liées au langage naturel pour l'extraction d'information : indexation conceptuelle à l'aide de bases de données lexicales, de lexiques multi-lingues ou d'ontologies, expression de requêtes ou de préférences à partir de hiérarchies de concepts, ... La plate-forme RFIEC (www.irit.fr/RFIEC), nœud du réseau national PLEXIR (www.irit.fr/PLEXIR), rassemble plusieurs logiciels, ressources et expérimentations qui rendent compte de la collaboration des équipes SIG, CSC, LILaC et SMAC de l'IRIT avec l'ERSS autour de ces différentes problématiques.

Temps et espace

Les concepts de temps et d'espace sont cruciaux dans la cognition et la communication humaine.
De nombreux travaux sur la représentation de ces informations à l'IRIT ont servi pour des travaux plus généraux sur la structure de textes et de dialogues, ou bien sur des tâches d'extraction de ces informations.
Les concepts spatiaux jouent aussi un rôle important dans des domaines connexes où l'utilisation du langage est aussi un élément à prendre en compte : l'architecture, la synthèse d'images (dans des modeleurs déclaratifs), et l'analyse de conversation en langues des signes.

Le projet WebCoop

À la requête «je cherche un gîte au bord de la mer en Midi-Pyrénées», vous n'aimez qu'un système vous réponde « zéro réponse trouvée », mais vous souhaiteriez plutôt qu'il vous explique « la région Midi-Pyrénées n'est pas au bord de la mer. Je peux proposer des gîtes à la campagne dans cette région, ou au bord de la mer dans une autre région». C'est ce que réalise WebCoop, un système de question-réponse coopératif réalisé par F. Bénamara (équipe ILPL). Il permet de traiter des questions précises de manière intelligente, en fournissant des réponses descriptives ou générales, toujours justifiées ou argumentées. Un modèle des connaissances et du lexique du domaine étudié, ainsi que des stratégies de raisonnement sur la manière de gérer les critères formulés dans une question sont les principes originaux à la base de ce logiciel.

Ressources : BDLex et MHATLex

Depuis 1983, des ressources lexicales générales du français sont développées à l'IRIT en vue de les utiliser pour le traitement automatique de la parole et des textes (*). Deux ressources lexicales phonologiques ont été réalisées : BDLex, dont les premières versions ont été développées dans le cadre du GDR-PRC Communication-Homme-Machine, puis MHATLex, mieux adaptée au traitement automatique de la prononciation dans sa variabilité. BDLex et MHATLex sont en partie distribuées par ELRA/ELDA.

(www.irit.fr/~Martine.Decalmes/IHMPT/PAROL/LEX/)

le Séminaire IRIT...

Dans le cadre de son séminaire, l'IRIT propose pour l'année 2006

Cycle Transformation de modèles

La transformation de modèles est une des activités qui devient prépondérante dans les développements de logiciel : génération de code, optimisation de code, génération de documents, composition d'aspects, rétro-conception, ...

Un langage de transformation de modèles est donc une composante principale d'un environnement de développement. Le modèle est l'objet de base de ces langages. Il requiert la définition de nouveaux opérateurs de construction, navigation, composition, décomposition, comparaison, évaluation, ... La sémantique de ces opérateurs et les langages associés, e.g., langages de transformation de graphe, de réécriture, constituent un axe de recherche actuel.

L'objectif de ce cycle de séminaires est de dresser un panorama des différentes approches nationales et internationales des langages de transformation de modèles.

Le séminaire est ouvert à tous. Pour recevoir le programme demander à être inscrit sur la liste de diffusion électronique. **05 61 55 65 10** / **info@irit.fr** / **www.irit.fr/MANIFS/manifs.html**

Les séminaires passés

octobre 2005 > février 2006

Belief merging and judgment aggregation
par Gabriella Pigozzi (Department of Computer Science at King's College, London)

VSI, Vision-Simulated Imaging
par Brian A. Barsky (University of California, Berkeley / Laboratoire d'Informatique Fondamentale de Lille (LIFL)

Mille ressources, un (pi-)calcul pour (essayer de) toutes les contrôler
par David Teller (Informatics, University of Sussex)

Création de vocabulaires visuels pour la reconnaissance d'objets
par Frédéric Jurie (INRIA Rhône-Alpes, projet LEAR (LEArning and Recognition in vision))

Cycle Transormation de modèles 1 : Présentation
par Louis Féraud (équipe MACAO, IRIT)

Cycle Transormation de modèles 1 : L'ingénierie des langages dédiés : une vision de l'évolution des approches de modélisation
par Jean Bézivin (Projet INRIA ATLAS, Université de Nantes)

Interactive Unawareness
par Martin Meier (Instituto de Análisis Económico (CSIC), Barcelona)

Cycle Transormation de modèles 2 :Transformation de documents par filtrage
par Pierre-Étienne Moreau (LORIA, Villers-lès-Nancy)

agenda séminaires

2 mars 2006

■ **Cycle Transformation de modèles 3 - Graph Transformation: Tutorial Introduction and Application to Model Transformation**
Gabrielle Taentzer, Technische Universität, Berlin

9 mars 2006

■ **Integrating Biomedical Data Sources: Current Approaches and Future Challenges**
Sharifullah Khan, PostDoc équipe PYRAMIDE, IRIT, Toulouse

6 avril 2006

■ **Cycle Transformation de modèles 4 - Transformation de modèles du type graphes en représentation de connaissances**
Michel Chein, LIMM, Montpellier

...À venir...

■ **Cycle Transformation de modèles - Fondements des transformations de modèles**
Reiko Heckel, University of Leicester

...À venir...

■ **Cycle Transformation de modèles - Transmorpher**
Jérôme Euzenat, INRIA, Rhône-Alpes

...À venir...

■ **Cycle Transformation de modèles - TOPCASED**
Jérôme Delatour, ESEO, Angers

Les manifestations passées

juillet 2005 > février 2006

JETOU 2005 : Rôle et place des corpus en linguistique
UT1, Toulouse

GLOBE'05
Copenhague, Danemark

RJC PAROLES 2005
IRIT, www.irit.fr/RJC2005/

IHM 2005
Maison de la Recherche, UT1, www.irit.fr/ihm2005/

Journée Sécurité Informatique
IRIT, Toulouse

WIT 2005 : Second International Workshop on Isomorphisms of Types
IRIT, www.irit.fr/recherches/TYPES/ZENO/WIT2005/

SEM'05 : Symposium on the Exploration and Modelling of Meaning
Biarritz, www.univ-tlse2.fr/erss/sem05/

1er séminaire VSST
ENIC Télécom, Lille

Interactive Unawareness
par Martin Meier, IRIT, Toulouse

GIPI
IRIT, Toulouse, www.gipi.org/web/index.htm

Réunion Thématique «L'ÉCRITURE»
Hôpital Purpan, Toulouse

agenda congrès

23 - 24 mars 2006

- **Journées FAC'2006, Formalisation des Activités Concurrentes**
  Groupe SVF / FéRIA, IRIT, Toulouse
  www.feria.cnrs.fr/FAC

11 - 12 mai 2006

- **MMUA'06, MultiModal User Authentication**
  IRIT, Toulouse
  http://mmua.cs.ucsb.edu/

5 - 16 juin 2006

- **EJCP'06, École Jeunes Chercheurs en Programmation**
  IRIT, Toulouse et Luchon
  www.irit.fr/ejcp2006

26 - 30 juin 2006

- **SdC 2006, Semaine de la Connaissance**
  Faculté des Sciences, Nantes
  www.sdc2006.org

4 - 8 septembre 2006

- **GLOBE'06, Grid and Peer-to-Peer Computing Impacts on Large Scale Heterogeneous Distributed Database Systems**
  Krakow, Poland
  www.irit.fr/globe2006

19 - 20 octobre 2006

- **LFA 2006, Rencontres francophones sur la Logique Floue et ses Applications**
  IRIT, Toulouse
  www.irit.fr/LFA06/

26 - 27 octobre 2006

- **WACA'02, Deuxième Workshop francophone sur les Agents Conversationnels Animés**
  IRIT, Toulouse
  www.enseeiht.fr/lima/ia/MEMBRES/adam/waca/

Les Passerelles de l'IRIT

Le GIPI, club d'innovation pour l'industrie, qui regroupe de nombreuses entreprises de Midi-Pyrénées, a proposé à l'IRIT l'organisation d'une visite-conférence le 16 février 2006. Cette association dynamique de responsables de petites et moyennes entreprises et de structures de transfert et incubateurs, a souhaité aborder le thème des outils pour la coopération recherche-industrie.

La soirée a débuté par l'accueil des participants par Luis Fariñas del Cerro, directeur de l'IRIT, qui a dressé un panorama des thématiques de recherche du laboratoire et présenté le potentiel de collaboration du laboratoire au travers de sa politique de valorisation.

La visite du laboratoire a permis aux participants d'apprécier la visualisation d'images sur grand écran, la capture de mouvement et l'utilisation de la projection sur écran hémisphérique pour les applications de réalité virtuelle.

La conférence a été présentée par Claude Detrez, responsable du partenariat à la délégation régionale du CNRS. Celui-ci a demandé à deux créateurs d'entreprises innovantes de faire part de leur expérience dans le domaine de la collaboration avec la recherche. Parmi eux, un ancien chercheur de l'IRIT, Laurent Karsenty qui a créé Intuilab, star-up dans le domaine de l'interaction homme-machine. Il a rapporté les difficultés rencontrées pour assurer la pérennité d'une petite entreprise de ce type. Richard Bru a présenté l'expérience de NOveltis, qui, dans le domaine de l'imagerie satellitaire, se tient à la pointe de la technologie grâce à ses relations avec les laboratoires de recherche, ce qui lui assure une notoriété suffisante pour garantir son développement.

Claude Detrez a fait le panorama des nombreuses possibilités de coopération et évoqué la future loi, qui devrait apporter des nouveautés dans le domaine. Il a fait part de l'état d'esprit ouvert de l'institution vis-à-vis de ces partenariats, ce qui n'exclut pas une certaine vigilance sur les questions de propriété intellectuelle et industrielle.

L'ERSS, un partenaire privilégié de l'IRIT pour le traitement du langage naturel

[suite de la page 6]

les adverbes cadratifs me paraît tout à fait prometteur. Le fait que certains de mes collègues travaillent dans le cadre de la SDRT est évidemment un grand avantage. À condition que nos lecteurs connaissent tous l'acronyme « SDRT » bien sûr.
La « Segmented Discourse Representational Theory » a été développée par N. Asher dans les années 90 comme une extension de la DRT (« Discourse Representation Theory » de H. Kamp), dans laquelle chaque proposition a une ou plusieurs fonctions rhétoriques au sein du discours. Une quinzaine d'années de recherche ont montré l'avantage théorique de combiner une notion riche de structure de discours avec la sémantique dynamique. Dans ce cadre, et dans d'autres apparentés exploités à l'IRIT, on arrive à mieux traiter des phénomènes comme la structure temporelle et spatio-temporelle des textes, l'anaphore, la présupposition, certaines ambiguïtés lexicales, la quantification, le calcul des implicatures et le positionnement des agents dans un dialogue. Il n'est pas difficile de percevoir l'importance de ces recherches dans les domaines fondamental et applicatif.

Quels sont les autres champs de collaboration entre l'IRIT et l'ERSS ?

Je ne suis pas sûr de pouvoir établir un tableau exhaustif. Pour commencer, des liens historiques nombreux sont tissés par l'opération « Sémantique et Corpus » de l'ERSS, coordonnée par A. Condamines, avec nos collègues de l'IRIT. En effet, les travaux autour de la constitution de ressources termino-ontologiques à partir de corpus ont démarré dans le laboratoire mixte ARAMIIHS (Action, recherche et application Matra/IRIT en interface homme-système) dans lequel N. Aussenac et A. Condamines se sont trouvées en contrat post-doctorat (1992). Leur collaboration n'a pas cessé depuis ; elle a pris la forme de nombreux projets interdisciplinaires qui ont permis de faire progresser la terminologie textuelle, tant du point de vue de l'ingénierie des connaissances que de celui de la linguistique.

Par ailleurs, avec la création en 2005 de l'opération TAL à l'ERSS, de nouvelles convergences sont apparues qui portent sur la recherche d'information. Je songe par exemple à des collègues de l'ERSS comme D. Bourigault, C. Fabre, N. Hathout, F. Sajous et L. Tanguy qui coordonne l'opération TAL chez nous. Dans ce contexte, l'analyseur Syntex développé par D. Bourigault avec la collaboration de C. Fabre se révèle extrêmement précieux pour de nombreuses recherches fondamentales et appliquées. En effet, Syntex permet d'extraire d'un corpus une liste de mots et de syntagmes structurés par des relations de dépendance syntaxique. Il fait partie des ressources de la plate-forme RFIEC développée à l'IRIT avec une participation forte de nos chercheurs.

Un autre exemple de collaboration est celui de B. Gaume qui travaille avec plusieurs de nos chercheurs sur la structure du lexique ou encore le cas de J. Mothe (IRIT/équipe SIG) et L. Tanguy (ERSS) qui coordonnent un projet CNRS/TCAN baptisé « Adaptation d'une chaîne de Recherche d'Information à l'Expression des besoins sur la base de traitements Linguistiques » (2004-2006). J. Mothe et L. Tanguy y emploient des techniques de profilage afin d'adapter le traitement des requêtes en fonctions de leurs caractéristiques linguistiques. Ce projet a concrétisé et renforcé un ensemble de travaux communs sur la place des méthodes linguistiques en recherche d'information. La création d'un séminaire régulier ERSS-IRIT est d'ailleurs un signe fort des relations entre nos deux laboratoires.

Il semble clair que la collaboration entre l'IRIT et l'ERSS est plus forte que jamais. Y-a-t-il cependant des domaines où des projets communs pourraient être lancés ?

L'ERSS comporte sept opérations qui balayent la linguistique de la phonologie au discours, en incluant la dialectologie. Mon sentiment est que toutes nos opérations pourraient bénéficier d'une collaboration avec l'IRIT. Nous avons des choses à apprendre sur le traitement de grandes masses de données et de systèmes complexes. En particulier, il me semble que le grand défi de la recherche à venir sera de combiner les analyses symboliques discrètes que certains caractérisent comme frégéo-chomskyennes avec des analyses quantitatives ou dynamiques (dites parfois subsymboliques). Les compétences de l'IRIT me paraissent incontournables si on veut aborder ces questions de façon sérieuse.

Jacques Durand
Directeur de l'ERSS
UMR 5610 (CNRS/Université Toulouse Le Mirail (UTM)/ Université Bordeaux 3)