INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Projet AÏDA

Modélisation et Apprentissage pour l'Interprétation de Données et l'Aide à la décision

Rennes

THÈME 3A

Table des matières

1 Composition de l'équipe

Responsable scientifique

Jacques Nicolas [CR Inria]

Assistante de projet

Maryse Auffray [AA Inria]

Personnel Inria

Yves Moinard [CR Inria]

René Quiniou [CR Inria]

François Coste [CR Inria (depuis octobre 2000)]

Personnel Université de Rennes 1 et autres établissements d'enseignement

Catherine Belleannée [maître de conférences]

Marie-Odile Cordier [professeur, détachement Inra puis délégation INRIA à partir de nov. 2000]

Israël-César Lerman [professeur]

Véronique Masson [maître de conférences]

Dominique Py [maître de conférences, IUFM de Rennes]

Sophie Robin [maître de conférences]

Laurence Rozé [maître de conférences, Insa de Rennes]

Pascale Sébillot [maître de conférences, délégation CNRS]

Basavaneppa Tallur [maître de conférences]

Raoul Vorc'h [maître de conférences]

Chercheurs doctorants

Vincent Claveau [bourse MENRT (à partir d'octobre 2000)]

Daniel Fredouille [bourse MENRT (depuis octobre 1999)]

Irène Grosclaude [bourse MENRT]

Christine Largouët [AERC Ensar]

Konan Lemée [bourse MENRT (terminée en septembre 2000)]

Emmanuel Mayer [contrat - Ater université de Rennes 1 depuis octobre 1999)]

Yannick Pencolé [bourse MENRT]

Ronan Pichon [bourse MENRT (terminée en avril 2000)]

Romuald Texier [bourse CIFRE-Société IDP]

Collaborateurs extérieurs

Philippe Besnard [DR CNRS, IRIT, Toulouse]

Yoann Mescam [service civil - objecteur de conscience depuis le 15 novembre 1999]

Raymond Rolland [maître de conférences, IRMAR, université de Rennes 1]

2 Présentation et objectifs généraux

2.1 Présentation générale et objectifs

Notre problématique générale est de fournir une assistance intelligente à un opérateur confronté à l'analyse de données complexes et de taille importante. Il s'agit d'extraire de ces données les éléments qui permettent à l'opérateur d'agir au mieux. Ceci suppose au minimum la mise au point d'un modèle explicatif des données traitées et souvent celle d'un modèle de l'utilisateur lui-même, afin de réaliser l'interface nécessaire à cette assistance.

Par assistance intelligente, nous entendons donc le développement de capacités automatiques de modélisation, de reconnaissance de situations intéressantes et d'élaboration de recommandations d'actions adaptées et explicables.

Nous nous situons dans une perspective intelligence artificielle. Le but est de rendre l'utilisateur autonome face à l'analyse de ses données, c'est-à-dire de ne pas requérir la présence d'un tiers (spécialiste) pour l'interprétation des résultats fournis. Respecter cet objectif suppose de fournir des résultats facilement interprétables et donc de travailler sur des modèles qui restent compréhensibles par cet utilisateur.

Ce thème correspond à des besoins bien identifiés en terme d'utilisateurs : opérateur chargé de la surveillance d'un système, scientifique cherchant à découvrir des relations intéressantes dans une masse de données, utilisateur sélectionnant des documents dans une base documentaire.

Les *thèmes scientifiques* sur lesquels se focalisent le projet concernent tous des capacités fondamentales pour l'interprétation de données : il s'agit de synthèse, de généralisation ou d'abstraction. Ces capacités sont de nature essentiellement abductive (pouvoir ajouter des hypothèses pertinentes à un ensemble de connaissances pour tenir un raisonnement) ou inductive (pouvoir induire des règles à partir de connaissances de même nature), c'est-à-dire que le problème central est celui de la sélection dans un ensemble donné (généralement infini) d'une ou de plusieurs hypothèses pertinentes pouvant expliquer au mieux un ensemble d'observations. De façon plus précise, le projet s'articule en deux composantes :

- **Modélisation** de systèmes (physiques, biologiques) ou de données complexes (langage naturel), en vue du diagnostic ou plus généralement de l'extraction de l'information pertinente. On s'intéresse à des modèles symboliques, par opposition aux modèles mathématiques numériques utilisés en automatique.

- **Apprentissage** pour l'acquisition ou la mise au point de ces modèles (essentiellement programmation logique inductive, inférence grammaticale et analyse de données). Là encore, il s'agit d'apprentissage symbolique, par opposition à des techniques d'apprentissage par renforcement.

Les *thèmes d'application* sur lesquels se focalisent le projet sont les suivants :

- **Aide à la surveillance de systèmes physiques**
  Un système physique évolue dans le temps, soit du fait de sa dynamique propre, soit sous l'effet d'actions ou d'événements extérieurs. La surveillance d'un tel système consiste à analyser les observations issues de capteurs, à en inférer l'état courant du système afin de détecter un éventuel dysfonctionnement, à caractériser ce dysfonctionnement en localisant le ou les composants défectueux, et éventuellement à préconiser l'action (ou la suite d'actions) qui semble la plus appropriée au maintien ou au rétablissement des fonctionnalités du système. Nous nous limitons aux systèmes de surveillance dans lesquels un opérateur est impliqué ; il s'agit donc plus précisément d'*aide* à la surveillance d'un système.

- **Aide à l'interprétation de séquences**
  Nous considérons ici deux types très différents de séquences naturelles : les textes (documents) et les séquences biologiques (ADN, ARN, protéines), vues comme des textes sur un alphabet généralement réduit. Dans les deux cas, on s'intéresse prioritairement à l'analyse de contenu. Le but est d'extraire la connaissance incluse dans les textes, en passant par une phase d'indexation automatique. Celle-ci consiste à traduire le contenu

de ces textes en une structure de données facilitant la recherche lors du traitement des requêtes qui lui sont adressées. Le filtrage d'éléments pertinents nécessite de plus l'emploi d'outils d'analyse syntaxique et/ou statistique.

3 Fondements scientifiques

3.1 Aide à la surveillance de systèmes physiques

Mots clés : surveillance, diagnostic, modèle de fonctionnement, modèle de panne, simulation, reconnaissance de scénario, graphe causal temporel, acquisition de scénario.

Glossaire :

alarme indicateur discret émis par un système de surveillance à partir d'événements et censé provoquer une réaction humaine ou automatique.

scénario (ou chronique) ensemble d'événements ponctuels et de contraintes temporelles sur ces événements caractéristiques d'une situation.

reconnaissance de scénario système permettant, à partir d'un ensemble de scénarios décrivant des situations (la base de scénarios), d'analyser au vol une séquence d'observations datées et de reconnaître les situations.

Résumé :

Les principales approches de l'intelligence artificielle au problème de la surveillance (et supervision) de systèmes sont basées sur un modèle de fonctionnement ou des dysfonctionnements au cœur du système de surveillance. Nous décrivons essentiellement le domaine de la modélisation de systèmes (ou d'activités complexes) évoluant dans le temps en vue de leur surveillance en ligne. Pour plus de détails et pour les références, consulter par exemple [BC96,GRO97,GRO98].

Le problème de la supervision par gestion d'alarmes est au cœur de nos travaux. Un opérateur chargé de la surveillance reçoit des événements (les alarmes) datés et émis par les composants eux-mêmes en réaction à des événements extérieurs. Les observations recueillies sur le système sont des informations discrètes, correspondant à un événement ponctuel ou à une propriété associée à un intervalle de temps. Les principales difficultés pour analyser ce flux d'alarmes sont alors les suivantes :

– la profusion des alarmes reçues : le superviseur peut recevoir jusqu'à plusieurs centaines de messages par seconde, dont certains sont non significatifs.

[BC96] M. Basseville, M.-O. Cordier, «Surveillance et diagnostic de systèmes dynamiques : approches complémentaires du traitement de signal et de l'intelligence artificielle», *rapport de recherche n° 1004*, IRISA, Mars 1996.

[GRO97] GROUPE ALARME, *Surveillance et interprétation d'alarmes en milieu industriel*, Actes des journées PRC-IA, Éditions Hermès, Grenoble, 1997, p. 9–30, S. Cauvin, M.-O. Cordier, C. Dousson, G. Defrandre, P. Laborie, F. Lévy, J. Montmain, M. Porcheron, I. Servet, L. Travé-Massuyès.

[GRO98] GROUPE ALARME, « Monitoring and alarm interpretation in industrial environments », *AI Communications 11, 3-4*, 1998, p. 139–173, S. Cauvin, M.-O. Cordier, C. Dousson, P. Laborie, F. Lévy, J. Montmain, M. Porcheron, I. Servet, L. Travé-Massuyès.

- l'imbrication des alarmes reçues : les ordres dans lesquels sont émises et reçues les alarmes peuvent être différents. De plus, les séquences d'alarmes résultant de pannes concourantes peuvent s'imbriquer. Les délais de propagation et, éventuellement, les voies d'acheminement doivent ainsi être pris en compte, aussi bien pour rétablir l'ordre des événements que pour décider à partir de quand on peut supposer avoir reçu la totalité des messages pertinents.

- leur redondance : certaines alarmes sont de simples conséquences d'autres. C'est en particulier le cas dans le phénomène connu sous le nom d'avalanche d'alarmes.

- perte et masquage : certaines alarmes émises peuvent être perdues ou masquées au superviseur par suite du dysfonctionnement d'un composant intermédiaire chargé de leur transmission. L'absence d'une alarme doit être prise en compte et peut fournir une indication intéressante sur l'état du système.

On peut distinguer deux cas posant des problèmes un peu différents. Les alarmes de conduite sont destinées à être traitées *en ligne* par l'opérateur de conduite. Le but de la surveillance est alors l'aide à la conduite, et l'analyse doit être faite en temps réel. L'opérateur a un objectif d'optimisation à court terme : il s'agit en général de rester au plus près d'un régime idéal, en tenant compte de la variabilité des entrées et de l'évolution naturelle des processus. En revanche, les dérives structurelles du système (usure des pièces, modifications lentes des propriétés de ses composants, etc.) ne sont pas prises en compte en tant que telles et sont corrigées par un réglage de paramètres.

Ce traitement *réactif* s'oppose au traitement *en profondeur* des alarmes de maintenance. On procède, dans ce cas, à une analyse *hors ligne* plus fouillée de l'historique du système, en cherchant à prévoir les incidents, à planifier les opérations d'entretien pour limiter au maximum les défaillances et les interruptions de service.

Dans le cadre de l'aide à la surveillance, nous cherchons non seulement à détecter les dysfonctionnements mais à les caractériser en localisant les composants responsables (diagnostic). Nous utilisons les approches dites à base de modèles pour lesquelles on suppose disponibles des modèles de fonctionnement ou de dysfonctionnement des systèmes surveillés.

L'exploitation en ligne des modèles est rarement envisageable car trop complexe vis-à-vis des contraintes temps réel, ceci en particulier en raison de la dimension temporelle que ces modélisations prennent en compte (automates communicants temporels ; graphes causaux temporels). Une approche consiste à transformer ces modèles hors ligne en en extrayant les éléments utiles au diagnostic.
Deux méthodes sont étudiées :

- Dans la première, le modèle est utilisé en simulation afin d'acquérir pour chaque panne significative les séquences d'observations correspondantes et constituer ainsi une base significative d'apprentissage. Les simulations associent à chaque situation de pannes ce que l'on appelle un scénario, c'est-à-dire un ensemble d'observables et un ensemble de contraintes temporelles qu'ils doivent respecter. Une des techniques permettant la supervision de systèmes dynamiques est alors la reconnaissance à la volée de ces scénarios. Son principe consiste en un suivi, en fonction des messages reçus, d'un ensemble de

scénarios potentiels jusqu'à une reconnaissance complète d'un ou plusieurs d'entre eux. L'apport d'une base de scénarios est, dans ce cas, nécessaire au bon fonctionnement de la supervision. Cette base doit contenir l'ensemble des scénarios de pannes possibles. Or son obtention n'est pas toujours aisée. Elle doit, par ailleurs, être actualisée au fur et à mesure de l'évolution, physique ou structurelle, du système sous surveillance. Une expertise humaine régulière s'avère coûteuse, raison pour laquelle il est préférable de s'orienter vers une méthode d'acquisition automatique de scénarios. Les séquences étiquetées sont ensuite généralisées afin d'obtenir un ensemble de scénarios discriminants. Un système de reconnaissance de scénarios est alors utilisé en ligne pour la surveillance du système.

– Dans la seconde approche, l'automate qui sert de modèle est transformé hors ligne en un automate adapté au diagnostic, appelé «diagnostiqueur». Ses transitions s'effectuent uniquement à partir des événements observables et ses états contiennent de l'information sur les pannes rencontrées par le système. Diagnostiquer le système consiste à parcourir le diagnostiqueur au fur et à mesure de l'arrivée d'événements observables.

Dans les deux cas, le point important est la réduction de la complexité. Dans la première approche, le point clé est d'extraire les informations discriminantes suffisantes pour identifier les dysfonctionnements. L'apprentissage automatique peut s'effectuer par différentes techniques. L'utilisation de la programmation logique inductive avec contraintes semble à cet égard représenter une voie de recherche intéressante. Dans le second cas, l'idée est celle de généricité et de répartition. Profitant de la structure du système (dans notre cas, la structure arborescente), le modèle générique est une représentation économique et suffisante permettant d'éviter de construire le modèle global et se contentant du modèle d'une branche.

3.2 Apprentissage automatique

Mots clés : inférence grammaticale, analyse de données, classification automatique, programmation logique inductive.

Glossaire :

PTA Prefix Tree Acceptor : il s'agit du plus petit automate fini déterministe reconnaissant l'ensemble des préfixes d'un ensemble de mots donné.

programme logique ensemble fini de clauses définies.

clauses définies disjonction de littéraux contenant un seul littéral positif, un littéral étant soit une formule atomique, soit la négation d'une formule atomique.

variable au sens «analyse des données» : il s'agit d'un attribut, d'un élément d'un système descriptif

ensemble des modalités domaine, ensemble des valeurs possibles pour une variable.

Résumé : *On décrit ici les techniques étudiées dans le projet, visant à acquérir des modèles et à les mettre au point de manière automatique à partir d'un ensemble d'observations. Cette automatisation pose des problèmes de filtrage, de structuration des observations, puis de spécification du «saut inductif», c'est-à-dire de la manière dont vont être définis puis calculés les modèles acceptables au vu des observations.*

Le projet s'appuie pour cela sur les travaux issus de l'apprentissage, de la classification et de l'analyse des données. Plus précisément, nous nous intéressons à un apprentissage de type structurel, c'est-à-dire où il s'agit de faire émerger des relations entre données parmi lesquelles les dépendances ne sont pas connues. Les techniques associées ressortent de l'inférence grammaticale ou de la programmation logique inductive suivant que les structures visées sont des grammaires ou des programmes.

3.2.1 Inférence grammaticale et programmation logique inductive

On appelle inférence grammaticale l'apprentissage automatique d'un modèle de langage à partir d'un échantillon fini des phrases du langage que la grammaire accepte (instances positives) et éventuellement d'un échantillon fini de phrases n'appartenant pas à ce langage (instances négatives). Les phrases correspondent dans les applications à un ensemble d'observations sur l'état ou le comportement du système et peuvent être aussi bien des séquences biologiques, des séquences d'alarmes ou des suites d'actions.

Spécifier complètement un problème d'inférence grammaticale suppose de

- définir la classe des langages acceptés ;
- définir la représentation des langages sur laquelle on travaille (grammaires formelles, automates, expressions) ;
- définir une relation d'ordre (relation de généralité) sur ces représentations, compatible avec l'inclusion sur les langages ;
- définir les conditions de présentation des phrases d'apprentissage («oracle» répondant aux questions de l'algorithme, présentation en bloc des instances ou incrémentale) ;
- définir un critère d'acceptation des solutions en fonction des instances, qui raffine la simple acceptation des instances positives et le rejet des instances négatives dans les langages associés aux solutions ;
- enfin, spécifier une stratégie d'exploration de l'espace des représentations choisi.

Nous nous intéressons plus particulièrement aux travaux tendant à renforcer l'applicabilité pratique des techniques d'inférence. Notre objectif est de démontrer que, moyennant un certain nombre de recherches, les résultats de l'inférence grammaticale sont transférables à l'analyse de corpus réels. De façon annexe se pose le problème de la constitution de benchmarks permettant la comparaison et l'évaluation des algorithmes produits.

Nous nous restreignons au cas où la classe acceptée est la classe des langages rationnels et où on travaille sur une représentation par automates finis. Il existe une relation d'ordre de généralité naturelle sur les automates induite par la fusion d'états dans un automate : toute fusion d'états dans un automate mène à un automate (appelé automate dérivé) reconnaissant un langage plus général ou équivalent au langage reconnu initialement. Si de plus on prend comme critère d'acceptation la complétude structurelle (c-à-d. toutes les transitions et états

d'acceptation d'un automate sont exercés), on montre que l'espace de recherche de toutes les solutions est un treillis. Celui-ci peut être construit à partir d'un automate canonique reconnaissant uniquement les instances positives. Les éléments du treillis sont dérivés de cet élément nul (l'automate canonique) par une fonction correspondant à la fusion de ses états. L'élément universel du treillis est l'automate universel, reconnaissant n'importe quelle suite de caractères. On peut restreindre encore l'espace de recherche si l'on s'intéresse uniquement aux automates déterministes. Dans ce cas, on remplace l'automate canonique par le PTA. L'apprentissage se ramène alors fondamentalement à un problème d'énumération dans un (grand) ensemble partiellement ordonné.

Les travaux que nous développons cherchent à étendre l'applicabilité des méthodes d'inférence sur les deux points suivants, en relation avec la liste que nous avons définie précédemment :

- mode de présentation des instances : passer d'un apprentissage «à données fixées», c'est-à-dire où l'on dispose initialement de toutes les instances, à un apprentissage incrémental, où les instances peuvent être disponibles en plusieurs étapes, suppose la résolution d'un certain nombre de problèmes difficiles si on ne souhaite pas recommencer l'apprentissage à partir de zéro à chaque nouvelle instance présentée.

- stratégie d'exploration : que le critère soit explicite ou non, la plupart des méthodes se contentent de fournir une seule solution, correspondant à un minimum local. Il s'agit d'une limitation importante par rapport aux applications : la plupart du temps, l'automate ayant une vertu explicative, on souhaite une caractérisation de l'ensemble des solutions possibles (combien y en a-t-il, en quoi diffèrent-elles ?). Ceci suppose de s'attacher à l'étude de stratégies complètes.
  Un second point concerne le sens de la recherche dans l'espace des grammaires ou automates : la plupart des méthodes procèdent par fusion d'états ou de non-terminaux, suivant en cela une progression par généralité croissante. Le critère de généralité maximale étant cependant souvent retenu, il est intéressant d'étudier à l'inverse l'inférence par «fission», autrement dit par spécialisation croissante d'un reconnaisseur universel. On espère ainsi aboutir aux solutions en un nombre réduit d'étapes.

La programmation logique inductive (PLI) consiste à inférer un programme logique P (par exemple, dans le langage Prolog) à partir de la donnée de faits complètement instanciés F qui doivent être vérifiés dans le programme cible et éventuellement d'un noyau de programme T qui modélise des informations déjà connues, qui peuvent faciliter l'apprentissage. Sur un plan logique, on souhaite vérifier la relation $T, P \models F$. Les prédicats pouvant intervenir dans les clauses de P sont généralement fixés, de même que l'ensemble des termes admissibles. Par rapport aux techniques d'inférence grammaticale présentées précédemment, on s'intéresse un peu au problème structurel qui consiste à trouver l'ensemble des relations intervenant dans les clauses du programme, et beaucoup au problème de la généralisation des termes intervenant dans les relations. Nous nous intéressons particulièrement aux techniques d'induction sur des clauses contraintes où les variables sont soumises à un système de contraintes [SR96]. L'étude

[SR96] M. Sebag, C. Rouveirol, « Induction de clauses contraintes », *in : Reconnaissance des formes et*

des relations entre inférence grammaticale et programmation logique est pertinente mais reste un domaine vierge. Les résultats escomptés sont des apports croisés dans ces approches et une meilleure maîtrise de leurs domaines d'application respectifs. Un autre intérêt est de pouvoir étudier le problème de l'inférence grammaticale dans un contexte logique, où l'induction est ramenée à un problème d'unification.

3.2.2 Classification

La classification est l'étape la plus en amont d'un processus d'analyse, étape considérant les données de manière globale, qui va faciliter des analyses postérieures plus fines, en regroupant ou au contraire en discriminant des ensembles de données brutes. L'enjeu et l'objectif est donc celui de la réduction la plus importante de la complexité qui permette cependant de filtrer au mieux l'information significative. Le contexte général où se situent nos travaux est celui d'une interaction entre d'une part, une approche de classification non métrique, combinatoire et statistique et, d'autre part, un ensemble de problèmes algorithmiques fondamentaux qui se présentent dans l'analyse de données complexes issues de l'observation, de la connaissance ou de modèles.

L'aspect classification comprend aussi bien la classification non supervisée par Analyse de la Vraisemblance du Lien (AVL) que celle supervisée qui relève de la discrimination par arbres de décision. D'autres méthodes d'analyse combinatoire des données peuvent également intervenir.
La classification est un outil fondamental pour spécifier une algorithmique de résolution approchée ; inversement, l'algorithmique intervient de façon essentielle dans la résolution de nos problèmes combinatoires de classification.

Les thèmes scientifiques que nous développons concernent les points suivants :

- réduction de la complexité d'un système descriptif (e.g. classification pour l'inférence de connaissances lexicales à partir de corpus de textes, voir la section suivante) ;
- élaboration de coefficients d'associations (e.g. pour la classification de parcelles agricoles, à partir d'images en vue de la surveillance d'une zone) ;
- comparaison de classifications sur des données complexes (e.g. pour la comparaison de différentes classifications de parcelles agricoles obtenues avec différents paramètres).

3.3 Recherche d'information dans un ensemble de documents, construction de lexiques

Mots clés : recherche d'information, terminologie, séquence binominale, sémantique, acquisition d'informations lexicales en corpus.

Glossaire :
composé, séquence binominale, structure binominale complexe dans nos travaux, association de deux noms de la forme *Nom Nom* en anglais et *Nom à/de (déterminant) Nom*

intelligence artificielle (RFIA'96), p. 706–716, Rennes, 1996.

en français. Ces noms peuvent être simples ou obtenus par adjonction d'un suffixe à un verbe (constituant déverbal).

interprétation ou calcul sémantique d'un composé détermination de la relation qu'entretiennent les constituants d'un composé.

prédicat, arguments un prédicat désigne un opérateur mettant en relation des arguments. Dans la phrase, le verbe joue en général le rôle de prédicat, les compléments étant ses arguments. La liste des arguments d'un prédicat forme sa structure argumentale.

terme symbole conventionnel qui désigne de façon univoque une notion à l'intérieur d'un domaine de connaissances.

Résumé : *Nous nous intéressons à la modélisation du contenu des textes via la modélisation de la sémantique de ses éléments descripteurs en indexation automatique. Notre but est de fournir des méthodes linguistiques permettant d'augmenter les possibilités d'apparier une requête et les textes de la base documentaire. Nous proposons d'une part un modèle hors domaine dont la fonction est de calculer le sens des séquences complexes*[1]*, qui constituent l'essentiel des termes des domaines techniques, en rétablissant leurs structures prédicatives sous-jacentes, et nous acquérons d'autre part les informations lexicales nécessaires à ce calcul (et à son extension en domaine) de manière automatique sur corpus. Ce module présente les idées centrales des différents thèmes que nous abordons, la recherche d'information, l'analyse de séquences complexes (extraction et interprétation) et l'acquisition d'informations lexicales en corpus.*

3.3.1 Recherche d'information - Indexation automatique

La recherche d'information (recherche documentaire) consiste, à partir d'un ensemble de textes et d'une requête d'un utilisateur, à proposer à ce dernier les textes adéquats. Il convient donc d'identifier les notions importantes d'un texte et de mesurer la proximité entre une requête et les textes de la base en déterminant celles qu'ils partagent. Les travaux de ce domaine passent généralement par une phase d'indexation automatique[SM83]. La qualité des systèmes de recherche d'information dépend de ce fait largement des techniques employées pour traduire le contenu des textes dans un langage d'indexation et pour réaliser l'appariement entre les textes indexés de la base consultée et la requête. Leur performance est mesurée à l'aide du *rappel*, proportion de réponses retrouvées parmi celles à produire, et de la *précision*, proportion de réponses pertinentes retrouvées parmi celles produites.

On oppose en général deux types d'indexation : l'indexation par index atomiques (indexation simple), qui assimile les indicateurs de contenu aux mots simples du texte (objectif premier : le rappel) mais conduit à des index peu discriminants et ambigus, et l'indexation par index complexes (indexation syntagmatique), qui manipule des groupes de mots (objectif premier : la précision) et aboutit donc à des index plus spécifiques et plus dispersés. En fait, les

1. Nous utilisons ici indifféremment composé, séquence complexe, ou séquence binominale dans le cas de deux noms.

[SM83] G. Salton, M. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill Computer Science Series, 1983.

résultats des systèmes ayant choisi l'une ou l'autre option ne permettent pas de trancher de manière définitive entre ces deux techniques et une voie moyenne semble raisonnable. Une façon d'aboutir à ce résultat consiste à privilégier une indexation syntagmatique sémantiquement riche, afin d'augmenter les possibilités d'appariement entre une requête et les textes de la base documentaire.

3.3.2 Analyse des séquences complexes

L'analyse des séquences complexes, en particulier binominales, est un enjeu fondamental dans de nombreuses applications du traitement automatique du langage naturel (TALN).

Une première phase de cette analyse, qui a fait l'objet de nombreux travaux, concerne l'extraction automatique de ces séquences qui constituent une grande proportion des termes, surtout dans les domaines scientifiques. Le repérage des séquences candidates à être des termes s'effectue selon les systèmes, soit par des critères syntaxiques, soit par des critères essentiellement statistiques, soit par une approche mêlant ces deux aspects (cf. par exemple[Bou94,Dai94]).

Une seconde direction concerne l'analyse sémantique de ces séquences. L'objectif des travaux de ce domaine est fréquemment de trouver la relation prédicative qui lie les constituants des composés. La difficulté du problème abordé tient au fait qu'une part importante de l'information sémantique contenue dans les séquences composées est implicite, ce qui nécessite de rendre compte d'inférences complexes. Par exemple, un *interpréteur de commandes* sert à *interpréter* des commandes (*relation explicite*) alors qu'un *parc à munitions* sert à *entreposer* des munitions (*relation implicite*). Le caractère implicite est de plus source d'ambiguïtés : *milk disease* est une maladie *causée* par le lait alors que *plant disease* est une maladie *affectant* une plante. De très nombreux travaux ont été consacrés, tant en linguistique qu'en intelligence artificielle, à la question de la détermination automatique du sens des séquences complexes à partir de la représentation sémantique des éléments simples qui les composent. Dans le domaine du TALN, deux types de modèles s'opposent : ceux qui dépendent d'un domaine, et ceux qui se consacrent à l'interprétation hors domaine des composés. C'est dans ce dernier cadre que se situent nos travaux. Les systèmes hors domaine proposent diverses stratégies. Une d'entre elles consiste à fonder le calcul de la sémantique des séquences complexes sur des règles générales d'interprétation, qui associent des prédicats à certains noms simples (c'est-à-dire non déverbaux) et font jouer à l'autre constituant de la séquence un rôle dans la structure argumentale de ce prédicat. Cette approche, initialisée par Finin[Fin80], trouve des prolongements dans les travaux développés au sein de notre équipe, où un modèle général d'interprétation hors domaine des composés, basé sur les travaux de Lieber[Lie83] et Selkirk [Sel82] pour traiter les composés déverbaux, et sur le modèle du Lexique Génératif de Pustejovsky[Pus95] pour

[Bou94] D. Bourigault, *Acquisition de terminologie*, thèse de doctorat, EHESS, 1994.

[Dai94] B. Daille, *Approche mixte pour l'extraction de terminologie : statistiques lexicales et filtres linguistiques*, thèse de doctorat, Université Paris 7, 1994.

[Fin80] T. Finin, *The Semantic Interpretation of Compound Nominals*, thèse de doctorat, University of Illinois, 1980.

[Lie83] R. Lieber, « Argument Linking and Compounds in English », *Linguistic Inquiry 2*, 14, 1983, p. 251–285.

[Sel82] E. Selkirk, « The Syntax of Words », *MIT Press*, 1982.

[Pus95] J. Pustejovsky, *The Generative Lexicon*, Cambridge:MIT Press, 1995.

interpréter les séquences complexes à relation implicite, a été développé. Plus précisément, nous avons mis au point un système qui permet de déterminer automatiquement la relation qu'entretiennent les constituants d'une séquence binominale de la forme *Nom Nom* en anglais et *Nom à/de (déterminant) Nom* en français, en se basant uniquement sur la forme de la séquence et sur la sémantique des mots qui la composent. Pour les composés contenant un constituant déverbal (*truck-driver, séquençage de l'ADN*), notre calcul automatique se base sur la satisfaction de la structure argumentale du prédicat verbal sous-jacent. Les composés sans constituant déverbal sont traités en généralisant la notion d'attachement d'information prédicative aux noms simples, en faisant appel à une représentation lexicale élaborée intégrant des informations pragmatiques telle que la met en œuvre Pustejovsky dans le lexique génératif. Dans ce formalisme, la structure des *qualia* représente un mot en termes de rôles sémantiques – fonctionnel, agentif, constitutif, formel[2] – qui rendent explicites les différents éléments de sens nécessaires à sa définition, rôles qui, pour un nom, sont fréquemment tenus par des verbes.

Quelle que soit la méthode utilisée pour définir des mécanismes de calcul de la sémantique des séquences composées, l'interprétation passe par l'étude précise de la sémantique nominale. Les lexiques correspondants ne peuvent pas être construits manuellement pour chaque application et ces informations lexicales doivent donc être acquises automatiquement à partir de corpus de textes du domaine de l'application visée.

3.3.3 Acquisition automatique d'informations lexicales à partir de corpus

Le développement de travaux d'acquisition automatique d'informations lexicales à partir de corpus connaît un essor considérable depuis le début des années 90 [HNS97].

Outre les travaux en extraction de terminologie présentés plus haut, l'acquisition consiste principalement à rechercher par des techniques statistiques les informations sur les unités extraites. Celles-ci sont de deux types : syntagmatiques et paradigmatiques.

Les informations syntagmatiques concernent les capacités d'association d'un mot : étant donné un mot, on cherche à découvrir les mots qui apparaissent dans le même contexte. Les travaux de ce type s'intéressent par exemple à trouver la structure argumentale de prédicats, à repérer des verbes typiquement associés à des noms, etc. Les informations paradigmatiques concernent les similarités entre les mots : étant donné un mot, on cherche à découvrir les mots qui ont des comportement les plus proches, c'est-à-dire, en se basant sur les thèses de Harris[HGR+89], ceux qui génèrent les mêmes contextes. Les travaux de ce type cherchent par exemple à constituer automatiquement des classes sémantiques ou à découvrir des relations lexicales (synonymie, antonymie, etc.) entre des mots.

Cependant depuis quelques années, des méthodes d'apprentissage symbolique sur corpus sont également utilisées pour acquérir des informations lexicales sémantiques. Par exemple,

2. Le rôle fonctionnel indique la fonction typique de l'objet dénoté, l'agentif le mode de création, le constitutif ses éléments constitutifs et le formel sa catégorie sémantique.

[HNS97] B. Habert, A. Nazarenko, A. Salem, *Les linguistiques de corpus*, Armand Collin/Masson, Paris, 1997.

[HGR+89] Z. Harris, M. Gottfried, T. Ryckman, P. M. Jr, A. Daladier, T. Harris, S. Harris, «The Form of Information in Science, Analysis of Immunology Sublanguage», *Boston Studies in the Philosophy of Science 104*, 1989.

au sein de notre projet, nous utilisons la programmation logique inductive pour inférer des éléments du lexique génératif de Pustejovsky [Pus95].

4 Domaines d'applications

4.1 Panorama

Résumé : *Les principaux domaines d'application des travaux de recherche menés dans le projet sont la génomique, la supervision de réseaux de télécommunication et la recherche d'information. Plus récemment le «monitoring» de l'activité cardiaque ainsi que la surveillance dans le domaine de l'environnement : transfert de polluants tels que pesticides et nitrates, surveillance de l'évolution des parcelles agricoles. D'autres applications sont abordées dans des domaines connexes tels que l'étude des séquences de mots en reconnaissance de la parole.*

4.2 La génomique

Mots clés : automate, bio-informatique, analyse linguistique.

Résumé : *L'objectif est, à partir d'un ensemble d'observations de suites produites par un mécanisme mal connu (des macro-molécules biologiques), de mettre en évidence des sous-séquences ou des structures qui apportent des indices pour la compréhension de ce mécanisme.*

L'objectif est, à partir d'un ensemble d'observations de suites produites par un mécanisme mal connu (des macro-molécules biologiques), de mettre en évidence des sous-séquences ou des structures qui apportent des indices pour la compréhension de ce mécanisme.

Ceci peut s'effectuer par la recherche des sous-séquences «surprenantes». En effet, l'existence de «signaux» biologiques se repère dans une séquence génétique par des sous-séquences particulières anormalement répétées ou enchaînées de façon précise. Nous avons ainsi étudié le mécanisme d'initiation de la traduction chez *E. coli* et nous proposons de même d'étudier l'enchaînement de motifs particuliers tout au long du génome. Le but peut être également de modéliser un mécanisme particulier en établissant une correspondance entre séquence, structure et fonction. Ainsi, nous avons commencé à étudier le phénomène de régulation dans les gènes impliqués dans la lipogénèse sur les vertébrés, qui fait intervenir des motifs encore peu connus, de taille très réduite et donc difficiles à repérer individuellement mais dont la structure d'enchaînement est relativement précise (e.g. palindromes faiblement espacés). Bien que le domaine des séquences biologiques soit un domaine d'intérêt privilégié, la classe d'applications permet d'envisager des domaines très variés où les mêmes techniques sont utilisables. Nous avons ainsi un contrat FT R&D en cours sur l'inférence de la syntaxe en reconnaissance de la parole et d'autres projets de recherches possibles en collaboration avec des industriels (modélisation de la stratégie d'un apprenant dans la résolution d'un problème par étapes ou dans son parcours d'un logiciel d'enseignement, automate d'accès à un service à partir de séquences).

[Pus95] J. Pustejovsky, *The Generative Lexicon*, Cambridge:MIT Press, 1995.

Les difficultés peuvent provenir de la taille des séquences, de l'existence d'interactions à longue distance, et de la superposition de nombreuses contraintes indépendantes pour aboutir à la séquence observée. Comme dans tout domaine réel, il faut aussi résoudre des problèmes d'approximation ou de bruit sur les observations. La modélisation s'attache à décrire les séquences à un niveau lexical, syntaxique et éventuellement sémantique.

4.3 Surveillance de systèmes physiques

Mots clés : surveillance, diagnostic, reconnaissance et acquisition de scénarios, diagnostiqueur, réseaux de télécommunications, surveillance cardiaque, systèmes natures, parcelles agricoles.

Résumé :

L'objectif est de surveiller à partir d'un ensemble d'observations (souvent datées ou au moins ordonnées) le fonctionnement d'un système physique, de détecter un éventuel dysfonctionnement, de le caractériser en localisant le ou les composants défectueux et de préconiser l'action (ou la suite d'actions) qui semble la plus appropriée au maintien des fonctionnalités du système. Nous nous appuyons sur les méthodes utilisant des modèles de fonctionnement ou de dysfonctionnement des systèmes surveillés (approches de type model-based) tout en cherchant à construire des systèmes efficaces utilisables en temps réel.

L'objectif est de surveiller à partir d'un ensemble d'observations (souvent datées ou au moins ordonnées) le fonctionnement d'un système physique, de détecter un éventuel dysfonctionnement, de le caractériser en localisant le ou les composants défectueux et de préconiser l'action (ou la suite d'actions) qui semble la plus appropriée au maintien des fonctionnalités du système.

Les systèmes auxquels nous nous intéressons étant pour la plupart dynamiques (évolution dans le temps), les modélisations sur lesquelles nous nous focalisons permettent de tenir compte de la dimension temporelle : automates communicants temporels, graphes causaux temporels, logiques du changement et logiques de l'action. Nous appliquons nos méthodes à la surveillance de systèmes physiques aussi bien artefacts (tels que les réseaux de télécommunications) que naturels (tels que le système cardiaque ou les systèmes écologiques) :

- **Surveillance de réseaux de télécommunications.** Deux types d'approches sont expérimentées pour la surveillance de ces réseaux. La première approche est de type reconnaissance de scénarios et tire parti de l'efficacité de ce type de méthodes pour satisfaire aux contraintes temps réel. Un des points importants est l'acquisition automatique de ces scénarios afin en particulier de pouvoir prendre en compte l'évolution technologique rapide des systèmes considérés. Nous privilégions une approche de type apprentissage supervisé en nous appuyant sur les modèles décrivant leur fonctionnement. L'acquisition des scénarios se fait à partir des données résultant de la simulation de dysfonctionnements et fait appel à des techniques d'apprentissage de type PLI (programmation logique inductive et, plus particulièrement, PLI avec contraintes). Nous appliquons cette approche

à la surveillance de réseaux de télécommunications dans le cadre d'une collaboration avec FT R&D : projets GASPAR (contrat de type CTI) et MAGDA (contrat RNRT). Une autre approche consiste à *compiler* le modèle du système, représenté par un graphe causal temporel, en un ensemble de scénarios. Le modèle est utilisé de manière déductive et l'interaction entre pannes multiples est prise en compte. Cette approche est appliquée au diagnostic de pompes primaires dans le cadre d'un contrat avec EDF.

La seconde approche consiste à produire directement un automate diagnostiqueur à partir de l'automate modélisant le comportement du système. Les travaux actuels ont pour objectif la construction de diagnostiqueurs génériques (pour ne pas avoir à représenter l'ensemble des comportements instanciés de chacun des composants mais uniquement leurs classes de comportement), ainsi que de diagnostiqueurs décentralisés (afin de pouvoir répartir une partie du diagnostic au niveau des composants eux-mêmes en s'appuyant sur des diagnostiqueurs locaux). Cette approche est appliquée aux réseaux de télécommunications au sein du projet MAGDA (contrat RNRT).

– **Surveillance cardiaque.** La technique de reconnaissance de scénarios est utilisée pour la surveillance, à partir de leur électrocardiogramme, de patients souffrant de problèmes cardiaques. Les scénarios sont obtenus par apprentissage automatique (PLI) sur des données provenant de simulations et de signaux réels. Nous prévoyons d'étendre la méthode dans le cadre de la conception d'une prothèse cardiaque (pacemaker - défibrillateur) «intelligente» afin d'analyser plus finement les dysfonctionnements constatés et de produire une stimulation mieux située dans le cycle cardiaque.

– **Surveillance de systèmes naturels.** Une première application porte sur la surveillance de parcelles agricoles et s'appuie sur une suite d'images satellitales et aériennes. Après une étape de classification de ces images (classification des parcelles), les résultats sont améliorés en tirant parti de modèles de l'évolution de la couverture de ces zones agricoles. Ces modèles d'évolution sont décrits dans le formalisme des automates temporels et utilisent plus précisément le formalisme de Kronos [Yov97]. Les résultats obtenus montrent une amélioration notable dans la précision des identifications des parcelles traitées.

  Nous avons aussi abordé dans le cadre d'une collaboration avec l'INRA (Unité Sciences du Sol et Agronomie de Rennes-Quimper) deux études portant sur la modélisation du transfert du nitrate au niveau d'un bassin versant d'une part, et de pesticides au niveau d'une parcelle agricole d'autre part. Dans les deux cas, nous avons choisi de nous appuyer sur les modèles quantitatifs classiquement utilisés afin de construire des modèles qualitatifs, plus adaptés à une prise de décision. Deux prototypes ont été construits et sont en cours de validation.

[Yov97] S. Yovine, «Kronos: A verification tool for real-time systems», *International Journal of Software Tools for Technology Transfer 1*, 1997.

4.4 La recherche d'information et l'accès à des bases de documents ou de services

Mots clés : recherche d'information, sémantique lexicale.

Résumé : *La recherche d'information constitue le domaine global d'application de nos travaux. Nous avons intégré certains de nos résultats sur le calcul sémantique des composés à un serveur d'interrogation de services télématiques dans le cadre d'un contrat CTI avec France Télécom R&D. Deux voies ont été explorées : l'utilisation du contexte de la séquence binominale pour mettre en œuvre une désambiguïsation plus efficace des mots et le réordonnancement des réponses proposées en favorisant celles obtenues en suivant les liens de modification nominale.*

L'amélioration de la qualité de service d'une application passe par l'adaptation de l'interaction au comportement de l'utilisateur. Notre approche consiste à interpréter les actions de cet utilisateur de façon à suivre l'évolution de ses buts et ses intentions. Ces connaissances sont modélisées par une logique modale.

4.4.1 Recherche d'information

Nous explorons trois voies complémentaires pour améliorer les performances des systèmes : le développement d'un modèle d'interprétation hors domaine des séquences binominales, l'étude de la variation sémantique des termes, c'est-à-dire la reconnaissance de l'équivalence conceptuelle de deux structures différentes, et l'inférence de connaissances lexicales à partir de corpus pour obtenir des lexiques sémantiques nécessaires au fonctionnement du modèle d'interprétation.

Une première application concrète des méthodes développées a été faite dans le cadre d'un contrat CTI avec France Télécom R&D, dans laquelle nous avons intégré certains résultats sur le calcul sémantique des composés à un serveur d'interrogation de services télématiques. Compte tenu des contraintes du système, deux voies ont été explorées : l'utilisation du contexte de la séquence binominale pour mettre en œuvre une désambiguïsation plus efficace des mots et pour rechercher des liens de paraphrase sémantique entre la requête adressée au système et les services de la base indexée, et le réordonnancement des réponses proposées aux utilisateurs en favorisant celles obtenues en suivant les liens sémantiques de nature syntagmatique (liens de modification nominale) qui unissent les constituants des séquences complexes.

Toujours dans cette optique, nous travaillons actuellement, dans le cadre d'une Action de Recherche Partagée de l'AUF (Agence Universitaire de la Francophonie) en collaboration avec Pierrette Bouillon (ISSCO Genève), Laurence Jacqmin (Université libre de Bruxelles) et Cécile Fabre (ERSS Toulouse) à un projet dont l'objectif est de déterminer, grâce à l'utilisation de méthodes d'apprentissage sur des corpus étiquetés, les prédicats témoins des différentes facettes sémantiques des noms d'un domaine, telles qu'elles sont définies dans le lexique génératif de Pustejovsky, et de montrer l'apport de l'usage d'un tel lexique dans un système de recherche d'information.

Le travail que nous avons réalisé sur le calcul sémantique des séquences complexes peut trouver des applications dans des domaines autres que la recherche d'information, tels que la

structuration de données terminologiques, le résumé automatique de textes ou la traduction automatique. De même, les méthodes que nous avons développées pour acquérir des informations de sémantique lexicale sur corpus ont des retombées en extraction d'informations (par exemple, déterminer des zones de textes où des mots entretiennent un type de relation prédéfini) ou dans diverses applications nécessitant des connaissances sémantiques liées à un domaine.

4.4.2 Système coopératif d'accès à un ensemble de services

Il s'agit d'interpréter les actions d'un utilisateur d'un service automatisé de façon coopérative, en tenant compte de l'évolution des buts et des intentions de cet utilisateur. Ce travail trouve une application particulière, en coopération avec FT R&D, au sein d'un service d'interrogation orale d'un serveur d'informations. Les séquences à interpréter sont alors l'historique du dialogue. Une logique modale complexe, combinant divers systèmes modaux, dont certains très classiques comme KD45, est déjà utilisée comme langage de représentation des connaissances. L'historique du dialogue est ainsi traduit sous la forme d'une formule modale de plus en plus volumineuse, représentant l'état de croyance actuel du système, lequel conserve ainsi également la mémoire de ses croyances passées, à chaque stade du dialogue. Il convient de tenir compte d'erreurs toujours possibles, soit parce que la requête de l'utilisateur est effectivement erronée (*donnez moi le serveur de météo marine pour l'Orne*, par exemple), soit par la suite d'une erreur du système «en amont» (de reconnaissance vocale par exemple). Il faut aussi tenir compte de l'évolution possible de la requête de l'utilisateur, qui réagit en fonction des réponses que le système lui a déjà données.

5 Résultats nouveaux

5.1 Modélisation de systèmes (ou d'activités complexes) évoluant dans le temps en vue de leur surveillance en ligne

Mots clés : modélisation, supervision, diagnostic, acquisition de scénarios, apprentissage par PLI, décision en univers incertain.

Résumé : *Dans le cadre de l'aide à la surveillance de systèmes ou d'activités complexes, nous nous intéressons plus spécifiquement au cas de la surveillance par analyse de séquences d'alarmes reçues par l'opérateur. Nous cherchons non seulement à détecter les dysfonctionnements mais à les caractériser en localisant les composants responsables (diagnostic). Nous utilisons pour cela des modèles du système, en particulier des modèles de pannes, qui sont décrits dans le formalisme des automates communicants temporels pour les deux applications de surveillance des réseaux (télécommunications et distribution d'électricité) que nous traitons, ainsi que dans celui des graphes causaux temporels.*

Les activités du projet dans ce thème portent sur trois points : l'acquisition de scénarios à partir de modèles, la construction d'automates diagnostiqueurs et l'interaction diagnostic/décision dans un univers incertain. Ces travaux de recherche s'appuient principalement à la surveillance de réseaux de télécommunications dans

le cadre d'un contrat CTI avec FT R&D (en collaboration avec l'université Paris-Nord) et la participation au projet RNRT MAGDA (en collaboration avec les projets Sigma2 et Pampa, l'université Paris-Nord ainsi qu'Ilog et Alcatel). Une autre application est en cours avec l'ENSAR pour la surveillance de terrains agricoles à partir d'images satellitales.

5.1.1 Surveillance de parcelles agricoles

Participants :
Marie-Odile Cordier, Christine Largouët.

Dans le cadre d'une collaboration avec l'Ensar, nous avons abordé le problème de la surveillance de parcelles agricoles à partir d'une série d'images aériennes et satellitales avec pour objectif la maîtrise de la qualité de l'eau. Le site de l'étude est le bassin versant Chèze-Canut, d'une surface de 8000 hectares, qui alimente en eau la ville de Rennes. L'objectif du projet est de fournir, trois fois par an, une carte thématique qui résume les différentes occupations du sol (prairie, maïs, blé, etc.) des parcelles agricoles de cette région. La classification des images par des méthodes statistiques traditionnelles (maximum de vraisemblance, nuées dynamiques, analyse discriminante) donne des résultats globalement corrects mais comportant néanmoins des anomalies ou des incohérences apparaissant sur la carte thématique résultat. Les anomalies correspondent à la dispersion de pixels isolés d'une certaine culture dans une parcelle connue comme appartenant à une autre culture. Les incohérences, détectables, si l'on compare plusieurs cartes résultats à des dates différentes, ont pour origine l'ambiguïté possible entre deux ou plusieurs cultures ayant des signatures spectrales proches.

Partant de ce constat, notre objectif est de proposer une méthode d'interprétation d'un territoire agricole par classification «intelligente» sur une séquence d'image s. Nous orientons notre démarche selon deux axes : une préclassification sur la parcelle et non plus sur le pixel et la discrimination des occupations du sol à l'aide d'un modèle d'évolution de la parcelle. La préclassification a pour objectif de fournir les occupations du sol possibles pour chaque parcelle. Cette préclassification est ensuite précisée à l'aide des connaissances sur les cycles culturaux et de l'historique des observations. La préclassification est réalisée à l'aide du logiciel Arkémie (développé par la société Arkémie Toulouse) qui propose une méthode simple de classification par parcelle. La modélisation de l'évolution de la parcelle agricole est réalisée à l'aide du formalisme des automates temporisés. La démarche consiste à confronter une suite d'observations, issues des images, avec une suite d'états, proposés par la simulation du système dynamique, dans le but de restreindre le nombre d'états susceptibles de représenter l'occupation du sol. Les automates temporisés sont généralement employés pour la représentation des systèmes temps-réel mais s'adaptent bien dans ce cadre puisqu'ils permettent l'expression des contraintes temporelles et des cycles caractéristiques de l'évolution de la parcelle agricole. Les occupations du sol correspondent aux états de l'automate reliés par des transitions munies de contraintes temporelles exprimées à l'aide d'horloges.

La discrimination des occupations du sol consiste à comparer les observations, dérivées des images par la préclassification, à l'état attendu par la simulation de l'automate. Ce problème est

abordé comme un problème de vérification et résolu à l'aide de techniques de model-checking. Le principe de reconnaissance de l'occupation du sol sur une série d'images est spécifié en termes de propriétés d'atteignabilité. La mise en oeuvre de la méthode se fait dans un système NosyBe, faisant appel à l'outil de model-checking Kronos, développé à Verimag.

L'expérimentation, réalisée sur une séquence de cinq images du site de l'étude, a donné des résultats encourageants. Dans cette application le formalisme utilisé pour représenter l'incertitude des informations est celui des ensembles. Récemment, nous avons proposé une extension de la méthode aux probabilités afin de tenir compte, lorsqu'une ambiguïté subsiste sur la classe d'une parcelle, du poids de la simulation dans le choix de la classe finalement affectée. Les résultats obtenus présentent une classification de qualité légèrement supérieure par rapport à l'approche précédente.

5.1.2 Graphes causaux temporels

Participants :
Marie-Odile Cordier, Irène Grosclaude, René Quiniou.

Au cours de nos travaux sur l'utilisation déductive des graphes causaux temporels afin d'obtenir les scénarios des pannes nous avons montré que, même en supposant l'absence d'effets contraires ou additifs dans le graphe causal, des interactions sont possibles entre les effets de plusieurs pannes. Elles correspondent à des phénomènes de recouvrements temporels d'occurrences d'effets identiques. Ces recouvrements peuvent conduire à des observations anormales pendant une durée plus longue que celle correspondant à la superposition des durées provoquées isolément par chaque panne. Ces interactions ne représentent qu'une partie des interactions pouvant survenir dans la réalité. En effet, les interactions entre pannes peuvent être bien plus complexes ; en particulier, les effets de certaines pannes peuvent empêcher, accélérer ou retarder la survenue des effets d'autres pannes.

Nous proposons d'étendre le formalisme classique des graphes causaux afin de traiter les interactions non monotones en permettant l'expression d'effets négatifs. En cas d'effets opposés, nous utilisons certaines propriétés des causes (instantanées ou continues) et des effets (persistants ou non) pour déduire le résultat de l'interaction. Nous proposons un algorithme de recherche de panne prenant en compte ces interactions [25]. Notre méthode utilise un modèle intermédiaire, calculé automatiquement à partir du graphe causal dans le formalisme étendu, et contenant la représentation concise et explicite de tous les phénomènes, causaux ou interactifs. La prise en compte de l'interaction augmente la complexité du diagnostic. L'efficacité de l'algorithme de recherche de panne est donc un point crucial. Nous l'améliorons par l'utilisation du modèle intermédiaire pré-compilé, et par un traitement spécialisé des informations temporelles à l'aide d'un gestionnaire de contraintes temporelles. Notre méthode permet ainsi d'expliquer efficacement un ensemble d'observations anormales, dues à une ou plusieurs pannes, éventuellement intermittentes, indépendantes ou interagissant.

L'algorithme a été implémenté en Java et intégré au logiciel *CAÏD*, développé pour la compilation des scénarios de pannes à partir du graphe causal, et permettant une visualisation du graphe causal et du modèle intermédiaire.

5.1.3 Monitoring en cardiologie

Participants :

Marie-Odile Cordier, René Quiniou, Véronique Masson, Sophie Robin.

Nous étudions, en collaboration avec le LTSI (unité INSERM, université de Rennes 1), l'application en cardiologie de la surveillance par reconnaissance de chroniques. Il s'agit d'analyser le signal provenant des différentes voies d'un monitoring cardiaque afin d'y détecter et de caractériser les arythmies cardiaques d'un patient sous surveillance. La nature, les caractéristiques et la fréquence des arythmies détectées permettent ensuite de proposer une attitude thérapeutique adaptée, par exemple un traitement médicamenteux ou la pose d'un pacemaker.

Une arythmie cardiaque peut se caractériser sur l'électrocardiogramme (ECG) par la succession d'ondes P et QRS respectant un certain nombre de contraintes temporelles. Il est donc naturel d'associer une arythmie à une chronique. Un premier objectif consiste à définir un ensemble de chroniques discriminantes, efficaces et adaptées à un patient donné. Nous utilisons pour ce faire une méthode d'apprentissage automatique du type programmation logique inductive (PLI) qui produit, en particulier, des représentations du premier ordre, nécessaires pour prendre en compte les aspects temporels. Ces représentations de haut niveau sont, de plus, facilement interprétables par les spécialistes qui peuvent ainsi les valider directement. L'adaptation au patient est réalisée en utilisant une base d'apprentissage contenant des ECG enregistrés sur ce patient ou des exemples d'ECG représentatifs d'arythmies que ce patient est susceptible de développer. Cette année nous avons continué à étudier l'apprentissage à partir de la voie principale de l'ECG [42]. Nous étudions également l'apprentissage à partir de plusieurs voies afin d'améliorer la résistance au bruit des chroniques apprises.

Un deuxième objectif consiste à adapter les techniques de reconnaissance de chroniques afin qu'elles prennent en compte les aspects multivoies. Diverses méthodes de contrôle de la reconnaissance ont été envisagées : reconnaissance globale de tous les événements, reconnaissance hiérarchique privilégiant l'une des voies, reconnaissance sur une voie et confirmation sur les autres voies, etc. Ces méthodes sont en cours de mise en œuvre. Elles seront évaluées dans différents contextes : fortement bruité dans le cas de système d'enregistrement Holter ou peu bruité dans le cadre de signaux issus de prothèses multisites multifonctions.

Cette étude est développée dans le cadre de l'Action Concertée Incitative *Télémédecine et Technologies pour la Santé* du MENRT (cf. 6.2).

5.1.4 Extension de l'approche diagnostiqueur

Participants :

Marie-Odile Cordier, Yannick Pencolé, Sophie Robin, Laurence Rozé.

La méthode des automates diagnostiqueurs s'inspire des travaux de [SSL+95,SSL+94] et s'applique aux systèmes à événements discrets. Partant d'un modèle de fonctionnement d'un système décrit en terme d'automates, elle consiste à construire directement un automate particulier appelé diagnostiqueur. Les transitions de cet automate correspondent aux événements observables et ses états décrivent les pannes du système. Diagnostiquer le système consiste à parcourir le diagnostiqueur au fur et à mesure de l'arrivée d'événements observables. Nous avons étendu cette méthode pour l'appliquer à l'interprétation d'alarmes de réseaux de télécommunications. Le réseau est modélisé en utilisant le formalisme des automates communicants temporels. L'approche diagnostiqueur développée dans [SSL+95,SSL+94] a dû ainsi être adaptée et étendue à ce formalisme ainsi qu'aux exigences de l'application traitée.

Un prototype de ce diagnostiqueur, baptisé Dyp, a été réalisé. Il permet de construire un modèle global du système à partir d'une description en termes de composants élémentaires et d'interconnexions; construire l'automate diagnostiqueur à partir du modèle global du système; visionner les diagnostics réalisés au fur et à mesure de l'arrivée d'alarmes. Ce prototype est diffusé dans le cadre d'un appel à logiciels lancé par le réseau d'excellence européen MONET (Model-based systems and qualitative reasoning - `http://monet.aber.ac.uk`).

Les extensions de l'approche diagnostiqueur sont les suivantes :

- réalisation d'un prototype DypGen permettant d'effectuer des diagnostics de façon générique. Les deux idées clés de DypGen sont de ne modéliser que des parties génériques du réseau (par exemple une branche dans le cadre d'un réseau hiérarchique) et de ne pas traiter séparément chacune de ces parties mais de traiter des ensembles de parties ayant le même comportement.

- intégration des contraintes temporelles dans Dyp. Les spécifications et l'analyse ont été effectuées pour les modules de composition et de construction du diagnostiqueur.

- pré-étude de la réalisation de diagnostiqueurs symboliques. Les automates utilisés ci-dessus sont définis de façon explicite (états et transitions) et les algorithmes de construction sont énumératifs. Or il existe des techniques symboliques, couramment utilisées dans les approches de type "model checking", permettant d'utiliser des représentations implicites du modèle sous forme de relations. Notre idée, à long terme, serait d'utiliser ces techniques pour construire un diagnostiqueur symbolique.

- développement d'une approche décentralisée du diagnostiqueur (cf section suivante).

Parallèlement à ces extensions, un éditeur de modèles a été réalisé. En effet, Dyp et DypGen s'appuient tout deux sur une description textuelle des modules et automates décrivant la

[SSL+95] M. Sampath, R. Sengupta, S. Lafortune, K. Sinnamohideen, D. Teneketzis, « Diagnosability of discrete event systems », *in : Proceedings of the International Conference on Analysis and Optimization of Systems, 40*, p. 1555–1575, 1995.

[SSL+94] M. Sampath, R. Sengupta, S. Lafortune, K. Sinnamohideen, D. Teneketzis, « A discrete event systems approach to failure diagnosis», *in : Proceedings of the Fifth international workshop on principles of diagnosis (DX'94)*, p. 269–277, 1994.

topologie du réseau et le fonctionnement de ses composants. Un éditeur de modèles a donc été développé, permettant d'effectuer une saisie graphique des modèles utilisés.

Ces travaux s'effectuent dans le cadre de la poursuite du contrat CTI avec FT R&D (Projet Gaspar) ainsi que dans le cadre du contrat RNRT qui rassemble sous le nom de Magda des industriels et des équipes de recherche sur le thème de l'interprétation des alarmes dans les réseaux de télécommunications (cf. 6.3). Pour le contrat FT R&D, la réalisation de l'éditeur de modèles a permis la construction du diagnostiqueur d'une branche complète du réseau. Pour le contrat RNRT, la modélisation d'un anneau SDH se termine.

5.1.5 Approche décentralisée du diagnostic

Participants :
Marie-Odile Cordier, Laurence Rozé, Yannick Pencolé.

Le problème considéré est la supervision de systèmes tels que les réseaux de télécommunications. Étant donnée la taille d'un tel système, une approche diagnostiqueur du type centralisé n'est pas implantable car elle nécessite la mise en place d'un modèle global du système. Nous avons donc décrit un système de diagnostic non plus fondé sur un unique diagnostiqueur mais sur un ensemble de diagnostiqueurs. Contrairement à l'approche proposée par [DLT00] qui nécessite la construction du modèle global, l'idée est ici de construire un automate diagnostiqueur s'appuyant uniquement sur le modèle local d'un composant du réseau supervisé. Chaque diagnostiqueur est en mesure d'établir un diagnostic local au composant en fonction des alarmes de ce composant reçues par le superviseur. Une fois établi l'ensemble des diagnostics locaux, la seconde étape est la coordination des diagnostics locaux en vue de construire le diagnostic global du réseau supervisé. Cette coordination des diagnostics locaux est effectuée après la mise en place d'une stratégie de reconstruction fondée sur les interactions possibles entre les diagnostics locaux. Cette stratégie est une étape nécessaire afin d'optimiser le calcul de coordination [BLPZ99] et obtenir ainsi un diagnostic en un temps satisfaisant pour le superviseur. Les travaux sur l'approche diagnostiqueur décentralisé ont été décrits dans [37] et dans [36]. Du point de vue de la mise en œuvre du système, nous avons terminé la programmation des diagnostiqueurs locaux. Cette mise en œuvre s'appuie sur les bibliothèques du projet *Dyp*, ce qui assure la compatibilité avec les approches centralisée et génériques concernant la description du modèle en entrée du système (éditeur de modèle utilisable dans cette approche).Ces travaux s'effectuent également dans le cadre du contrat CTI avec FR&D (Projet Gaspar) ainsi que dans le cadre du contrat RNRT Magda (cf. 6.3). Dans le cadre du contrat FT R&D, nous avons utilisé les modèles développés afin de mettre en place l'approche décentralisée. Pour le contrat RNRT, la mise en place d'une démonstration de l'approche décentralisée est en cours. Cette démonstration est fondée sur le modèle SDH décrit avec l'éditeur de modèle (cf. section 5.1.4).

[DLT00] R. Debouk, S. Lafortune, D. Teneketzis, « Coordinated Decentralized Protocols for Failure Diagnosis of Discrete Event Systems», *Discrete Event Dynamic Systems 10*, 1-2, 2000, p. 33–86.

[BLPZ99] P. Baroni, G. Lamperti, P. Pogliano, M. Zanella, « Diagnosis of large active systems », *Artificial Intelligence 110*, 1999, p. 135–183.

5.2 Apprentissage automatique et structuration de données

Participants : Catherine Belleannée, François Coste, Daniel Fredouille, Israël-César Lerman, Yoann Mescam, Konan Lemée, Jacques Nicolas, Basavaneppa Tallur, Raoul Vorc'h.

Mots clés : inférence grammaticale, analyse de données, classification automatique.

Résumé : *L'automatisation de la construction de modèles de systèmes complexes est au cœur des motivations des recherches effectuées ici. Nous focalisons nos travaux pour le traitement de données qui se présentent sous forme de séquences discrètes finies. L'analyse de ces séquences passe généralement par deux étapes : une étape de prétraitement d'analyse à un niveau lexical et éventuellement syntaxique, où il faut regrouper les séquences ou sous-séquences similaires, et une étape d'inférence grammaticale qui conduit au modèle souhaité.*
Nous traitons également des problèmes importants associés au développement pratique de ces outils : la réduction de la complexité d'un système descriptif et la comparaison de modèles structurant un même ensemble d'objets.

Notons cette année une forte action de prospective en faveur d'une réorientation d'une partie du projet vers la bio-informatique. Ceci s'est traduit concrètement par la conception et la création d'un nouveau DEA «Informatique et Génomique» ainsi qu'un projet de génopole Ouest dans lequel l'équipe tient une place centrale. Ce projet doit être expertisé au début de l'année 2001 et bénéficie d'ores et déjà d'un soutien régional (cf. 7.1 Actions régionales).

5.2.1 Classification hiérarchique de gros ensembles libre ou sous contrainte de contiguïté

Participant : Israël-César Lerman.

Il est maintenant bien admis et depuis longtemps que la technique de recherche des plus proches voisins réciproques est cruciale pour la conception d'algorithmes de construction ascendante hiérarchique d'arbres de classification sur de «gros» ensembles. La situation spécifique considérée et qui se retrouve dans nombre d'applications est celle où, pour la formation des classes, une contrainte de contiguïté doit être respectée. On suppose de plus que le nombre d'objets contigus à un objet donné reste limité par une constante fixée à l'avance. C'est typiquement la situation pour la classification des pixels d'une image numérisée. K. Bachar [ESSCA, Angers] avait, notamment dans le cadre de sa thèse [université de Rennes 1, décembre 1994], élaboré et analysé sur les plans théorique et expérimental, un algorithme CAHCVR (Classification Ascendante Hiérarchique sous Contraintes utilisant les Voisins Réciproques). On démontre et on vérifie dans la pratique que la complexité moyenne en temps de calcul devient linéaire, au lieu de quadratique dans le cas général, en fonction du nombre d'objets, ce qui est optimal.

Dans ces conditions, il importe d'approfondir l'étude algorithmique par rapport aux critères de type inertiel déjà mis en oeuvre; mais aussi par rapport aux critères de dissimilarité "informationnelle" issus de la méthode AVL de la vraisemblance des liens. D'autre part, une telle

algorithmique doit permettre des agrégations multiples se produisant à un même niveau de l'arbre des classifications. Les expériences qui ont été menées ont pu montrer la quasi monotonie des critères issus de l'AVL par rapport à ceux inertiels où des inversions peuvent beaucoup plus facilement se présenter. Cette recherche devrait se poursuivre en relation avec K. Bachar et en relation avec la problématique fondamentale de la classification hiérarchique de "très gros ensembles" (sujet de DEA posé).

Ces ensembles se retrouvent communément dans le domaine de la Fouille des Données ("Data Mining"). Plus directement, ils peuvent être fournis par le projet CAPS pour l'analyse du comportement de l'exécution de très gros programmes comportant jusqu'à des centaines de milliards d'instructions (benchmarks SPEC95). Il s'agit de simuler un tel comportement à partir de l'exécution d'une faible part du programme formée de la réunion de tranches connexes d'instructions, toutes de même taille, mutuellement disjointes et en nombre limité. À partir d'une description, la classification automatique de l'ensemble de toutes les tranches et la détermination de tranches centrales et représentatives de classes "bien" déterminées, permet de répondre de façon efficace au problème posé. Telle est l'idée développée dans la thèse de Thierry Lafage (projet CAPS) qui a appliqué avec intérêt le programme CHAVL (Classification Hiérarchique par Analyse de la Vraisemblance du Lien). L'application de cet algorithme qui comprend bien le principe des voisins réciproques, s'est avérée performante compte tenu du fait que le nombre total des tranches était inférieur à 10.000. En multipliant par 10, 100 voire 1000 le nombre de tranches, on aboutit à un "très gros ensemble" pour lequel une algorithmique spécifique s'avère nécessaire.

5.2.2 Inférence grammaticale

Participants :
François Coste, Daniel Fredouille, Jacques Nicolas.

Notre activité en inférence grammaticale s'est poursuivie selon trois directions, l'inférence par fission, l'inférence d'automates non déterministes et l'inférence de grammaires de Lambek. La première étude s'inscrit dans le cadre d'une collaboration avec l'équipe Cordial de Lannion (L. Miclet) et fait suite au stage de DEA de D. Fredouille concernant l'inférence d'automates déterministes par fission. La fission semble une stratégie de recherche séduisante par rapport à la stratégie habituelle de fusion, car elle travaille sur des automates plus petits, potentiellement plus proches de la solution. Nous avons cependant mis en valeur les faiblesses inhérentes à l'approche par fission, notamment la difficulté d'obtenir des heuristiques permettant à la fois d'empêcher l'explosion combinatoire du nombre de fissions à tester à chaque étape de l'algorithme et, de guider efficacement la recherche [24].

L'inférence d'automates non déterministes est un second axe de recherche explorant cette fois le mode de représentation des langages réguliers recherchés. Considérer l'inférence d'automates non déterministes permet d'obtenir une représentation de certains langages réguliers exponentiellement plus compacte que celle qui serait obtenue par inférence d'automates déterministes. Dans cette optique, nous espérons pouvoir identifier ces langages grâce à un ensemble de données de départ de taille inférieure à celui necessité par l'inférence d'automates détermi-

nistes. Cette année a vu l'achèvement du travail de François Coste proposant et formalisant l'extension du cadre de l'inférence grammaticale régulière à l'inférence d'automates classifieurs [11, 23], automates permettant de classifier plusieurs langages simultanément. Ces résultats ont été étendus au thème de l'inférence d'automates classifieurs non déterministes, pour lequel nous avons étudié le problème de la détection efficace d'ambiguité [22].

Enfin, dans le cadre d'un stage de DEA (R. Bonato), nous avons également développé un travail sur le traitement de l'inférence grammaticale de grammaires de Lambek, grammaires particulièrement adaptées au traitement du langage naturel, en collaboration avec C. Rétoré de l'équipe Paragraphe. Il s'agit d'une extension des AB-grammaires que nous avions étudiées l'année dernière.

Nos perspectives de recherche concernent l'application de l'inférence grammaticale à la recherche de signatures dans les séquences génomiques. Ainsi, un travail est en cours afin d'appliquer ces algorithmes à la discrimination de sites promoteurs dans le génome d'*Escherichia coli*. Une autre application, la prédiction de structures secondaires de protéines, fait l'objet d'une thèse qui démarre en novembre, en collaboration avec l'université de Vérone (R. Bonato).

5.2.3 Recherche de variants génétiques discriminants dans l'homéostasie du fer

Participants :
Israël-César Lerman, Jacques Nicolas.

Cette action qui débute concerne le projet FER de l'UPR41 (Recombinaisons génétiques). Elle est menée en relation étroite avec Jean Mosser et Véronique Douabin (UPR41). Elle s'articule autour d'une thèse préparée par Véronique Douabin et co-dirigée par Jean Mosser et I.C. Lerman. Le problème général consiste à déterminer des profils génétiques responsables ou accompagnant l'hémochromatose (surcharge en fer). À cet égard, un premier ensemble d'apprentissage ou échantillon (au sens statistique du terme) est en cours de constitution. L'étape suivante consistera alors à déterminer une stratégie optimale de classification et d'analyse combinatoire des données pour découvrir des régressions intéressantes.

5.3 Acquisition d'informations lexicales sémantiques sur corpus et applications

Participants : Philippe Besnard, Vincent Claveau, Israël-César Lerman, Jacques Nicolas, Ronan Pichon, Pascale Sébillot.

Résumé : *Dans le cadre du développement de méthodes et outils linguistiques permettant d'augmenter les possibilités de détecter des concepts équivalents entre une requête et une base de documents indexés, nous nous intéressons à l'acquisition automatique de deux types de lexiques sémantiques à partir de corpus : acquisition d'éléments du lexique génératif de Pustejovsky par des méthodes de programmation logique inductive et acquisition de lexiques basés sur la sémantique différentielle*

de Rastier par des méthodes de classification. L'utilisation de ce dernier type de lexique permet de réaliser l'expansion des index représentant les textes par ajout de synomymes, de variantes morpho-syntaxiques, etc. Le modèle du lexique génératif est, quant à lui, particulièrement bien adapté à la génération de variantes en privilégiant plusieurs types de liens sémantiques (cf. [13]). Ainsi, si une requête contient la séquence jaugeur de carburant, le fait de disposer d'un lien entre le nom jaugeur et le verbe mesurer (fonction typiquement associée) permet, par exemple d'étendre la recherche aux séquences voisines mesure du carburant ou mesurer le carburant. Par ailleurs, les méthodes d'acquisition de lexiques sur corpus que nous développons sont la base d'un logiciel d'aide à l'argumentation dont nous avons élaboré les premières versions.

5.3.1 Acquisition automatique d'éléments du Lexique Génératif de Pustejovsky par programmation logique inductive

Participants :
Pascale Sébillot, Vincent Claveau, Jacques Nicolas.

Nous nous intéressons, dans le cadre d'une action de recherche partagée de l'AUF de 2 ans débutée en 1999, en collaboration avec Pierrette Bouillon (ISSCO Genève), Cécile Fabre (ERSS Toulouse) et Laurence Jacqmin (Université de Bruxelles), à l'acquisition automatique, à partir de corpus de textes, d'éléments du Lexique Génératif de Pustejovsky grâce à des techniques d'apprentissage symbolique. Notre objectif principal cette année a été de chercher à améliorer la qualité de la méthode de type programmation logique inductive (PLI) que nous avions développée au cours de l'année 99. Le but de cette méthode est de permettre, dans un corpus de textes techniques étiqueté catégoriellement, de distinguer automatiquement les couples nom-verbe (N-V par la suite) liés par une relation sémantique codée dans le lexique génératif des autres couples N-V. Pour ce faire, 4000 exemples positifs et 7000 exemples négatifs constitués à l'aide des contextes d'apparition de ces N et V dans les phrases (tels que la catégorie grammaticale du mot avant et après le nom) sont générés automatiquement et fournis en entrée de Progol, mise en œuvre de la PLI développée par Muggleton, qui produit alors des clauses par généralisation de certains exemples positifs. L'an dernier, les clauses générales obtenues couvraient 88% d'exemples positifs et 5% d'exemples négatifs (coefficient de Pearson de 0.84). La méthode d'apprentissage avait également été validée empiriquement en utilisant ces clauses générales pour étiqueter les couples N-V du corpus et en comparant la pertinence des décisions prises par rapport à un étiquetage manuel, les résultats obtenus étant largement meilleurs que ceux de tests de type Khi2 mais encore bruités.

Les améliorations apportées à cette méthode d'apprentissage (cf. [39, 40]) portent essentiellement sur deux points. Le premier concerne le développement d'un outil permettant de détecter automatiquement la forme des exemples positifs et négatifs à fournir au logiciel d'apprentissage pour obtenir les meilleurs résultats. Différents types d'éléments de contexte sont proposés à cet outil (étiquettes catégorielles, sémantiques (cf. ci-dessous), diverses distances, etc.) qui choisit les plus pertinents. Le second porte sur la mise au point d'un étiquetage

sémantique du corpus technique étudié (manuels de maintenance d'hélicoptères fournis par Matra-CCR) afin d'améliorer l'apprentissage. Ceci a tout d'abord nécessité le développement d'un lexique recensant, pour chaque mot, ses étiquettes sémantiques possibles. Pour classifier les noms, nous avons utilisé les classes les plus génériques de la base lexicale WordNet que nous avons éventuellement raffinées ; 33 classes de noms ont été recensées, 5 de verbes, 4 d'adjectifs et 11 de prépositions. Ces informations sémantiques sont ensuite projetées sur le corpus étiqueté catégoriellement. La désambiguïsation est effectuée à l'aide du logiciel Tatoo de type HMM de l'Issco. Seuls 1.18% d'erreur d'étiquetage subsistent à l'issue de cette étape, essentiellement dûs à des prépositions. La méthode d'apprentissage a alors été appliquée au corpus étiqueté sémantiquement. Les clauses générales obtenues permettent de couvrir 90% des exemples positifs et seulement 0.7% d'exemples négatifs (Pearson de 0.91). De même, le coefficient de Pearson lors de l'évaluation empirique croît de 10%. Parallèlement, nous avons également testé la portabilité de notre méthode d'apprentissage sur un second type de corpus, le corpus Euro, qui contient des textes décrivant la mise en place de l'Euro et ses conséquences. La validation théorique de cette méthode sur le corpus étiqueté catégoriellement présente un coefficient de Pearson de plus de 0.90. Enfin, nous avons mis au point une interface de visualisation des résultats, c'est-à-dire une interface qui, pour chaque lettre de l'alphabet, fournit la liste de noms étudiés et, pour chaque nom choisi, fournit la liste des verbes appris comme faisant partie de sa structure des qualia. La partie du corpus où apparaît une première occurrence de la paire N-V concernée s'affiche dans une fenêtre, le nom et le verbe en une couleur distincte du reste du texte ; il est alors possible de passer directement à l'occurrence suivante. Cet outil, outre son rôle de visualisation, permet également d'interroger des documentalistes sur la pertinence des verbes de la qualia d'un nom pour la recherche d'information. Un premier test de ce type a été réalisé avec une maquette de l'outil de visualisation au département documentation de la banque Fortisbank à Bruxelles.

5.3.2 Acquisition automatique de lexiques basés sur la sémantique différentielle de Rastier

Participants :
Pascale Sébillot, Israël-César Lerman, Ronan Pichon.

Nous nous intéressons, à l'aide de méthodes de classification, à l'acquisition automatique de lexiques sémantiques basés sur la sémantique différentielle (SC) de Rastier, théorie linguistique dans laquelle l'accent est mis sur les relations entre les significations des mots au sein d'un lexique, et dont une des thèses est que ces relations sont fortement dépendantes d'observations d'utilisation des mots en corpus.

Dans SC, la signification d'un mot est définie par les différences qu'elle entretient avec les autres significations présentes dans le lexique. Ces différences sont représentées par des sèmes (ou traits sémantiques). Au sein d'une même classe sémantique, correspondant à un groupe de mots partageant certains traits sémantiques et pouvant être échangés dans certains contextes, les éléments possèdent des sèmes génériques correspondant aux contextes dans lesquels ils peuvent effectivement être échangés, et des sèmes spécifiques, correspondant aux

autres contextes. Pour Rastier, le sens d'un mot est totalement déterminé par le co-texte qui l'entoure, et deux types de contextes sont fondamentaux pour caractériser les relations de signification lexicales : le thème de l'unité de texte dans laquelle est située l'occurrence étudiée et son voisinage. La présence d'un thème peut être caractérisée par la co-présence, dans une unité de texte, de quelques mots typiques de ce sujet.

Lors de l'année 99, nous avons, par l'étude de la distribution relative des noms dans les différents paragraphes d'un corpus (Le Monde Diplomatique, corpus global de 7.8 millions de mots, dont 1 million ont servi à cette expérience), mis en évidence à l'aide d'une méthode de classification hiérarchique (analyse de la vraisemblance du lien, AVL) les groupes de noms dénotant des thèmes principaux du corpus. La co-présence de certains de ces mots dans des paragraphes permet ensuite d'affecter chaque paragraphe du corpus global à un ou plusieurs thèmes et de découper ce corpus en sous-corpus thématiques. Enfin, à l'intérieur de chacun de ces thèmes, nous avons, pour chaque nom, construit son vecteur de voisinage formé des noms et adjectifs apparaissant dans une fenêtre de 5 mots avant et après chacune de ses occurrences. Ce vecteur permet de regrouper, par classification ascendante hiérarchique, les mots interchangeables dans les mêmes contextes au sein de classes sémantiquement homogènes. Nous avons étudié les similarités et dissimilarités de sens entre deux occurrences d'un même mot dans deux thèmes différents, ou entre deux mots dans un même thème, en détaillant les similarités et dissimilarités entre leurs vecteurs de voisinage. Cette étude se fait par calcul de l'intersection ou de la différence ensembliste entre ces vecteurs de voisinage, et nous avons interprété les ensembles de mots ainsi obtenus en y recherchant des séquences caractérisant une différence entre la signification de mots.

Notre contribution de cette année sur l'acquisition de ce type de lexique a été essentiellement d'ordre génie logiciel : nous avons choisi de mettre en place des outils paramétrisables permettant d'effectuer des tests multiples sur différents sous-corpus thématiques à un moindre coût. Par exemple, pour effectuer le regroupement des mots en classes sémantiques, différents filtres sont appliqués au sous-corpus, éliminant par leurs applications successives les informations contextuelles jugées non pertinentes et organisant celles restant dans un format interprétable par le programme calculant la similarité. Cette procédure doit être répétée à chaque modification des critères de similarité dans ce dernier programme, puisque les informations contextuelles en dépendent. Les filtres sont quant à eux éventuellement modifiés en fonction des informations recherchées. Nous avons donc choisi de créer une banque de données intermédiaire entre le corpus et le seul vecteur de contexte, le fichier de voisinage, contenant plus d'informations sur l'environnement de chaque nom et permettant de tester différents calculs de similarité et choix de contexte, et produit un programme Perl paramétré permettant de modifier ces divers critères. De même, jusqu'ici, les résultats de la classification des noms étaient extraits manuellement (décodage des classes et choix des classes produites pertinentes). Par conséquent, seules quelques classes sémantiques choisies manuellement avaient été traitées, c'est-à-dire avaient vu leurs caractéristiques en termes de sèmes génériques et spécifiques étudiées par intersection et différence ensembliste. Pour automatiser le traitement de l'ensemble des classes (choix des pertinentes par rapport à la méthode de classification, passage de la représentation polonaise préfixée avec des codes pour chaque nom à la production effective de classes de mots, accès aux informations contextuelles de ces mots, etc.) des logiciels, là aussi paramétrables, ont été créés. Ces outils seront la base de notre étude fine des différents sens, fondée sur une structuration

des éléments contextuels.

5.3.3 Aide à la production d'arguments

Participants :
Pascale Sébillot, Philippe Besnard.

Disposer de bases lexicales porteuses d'une information sémantique riche et liée au domaine d'étude ouvre des voies applicatives multiples. Nous avons utilisé nos travaux sur le sujet dans le cadre d'un travail développé avec Thomson-CSF portant sur la modélisation de la théorie de l'argumentation de Perelman et la production d'une maquette de logiciel qui, étant donné le début d'un argument, produit automatiquement sa fin. Par exemple, un argument de type réciprocité peut être utilisé dans une conversation pour convaincre son auditoire ; à la saisie de la première partie de l'argument, par exemple "Si les vendre n'est pas honteux pour vous", la fin doit s'afficher, à savoir ici "les acheter ne l'est pas non plus pour nous". Nous avons, avec notre partenaire industriel, choisi de représenter ces arguments sous forme de transformations de triplets dont les éléments doivent respecter certaines contraintes sémantiques. Par exemple, un argument de réciprocité peut se modéliser sous la forme $\langle$ X,Y,Z $\rangle \longrightarrow \langle$ U,V,W $\rangle$ pourvu que (X,U) et (Y,V) et (Z,W) comportent deux couples d'antonymes, l'autre étant un couple de synonymes. Nous avons ainsi formalisé un grand nombre des arguments proposés par Perelman et énoncé des contraintes sémantiques de types divers sur les éléments des triplets. Nous avons réalisé une maquette du logiciel en prenant pour base lexicale d'entrée WordNet. Nous avons ainsi pu montrer la faisabilité du produit tout en démontrant que l'utilisation d'une base généraliste ne pouvait résoudre certaines des contraintes sémantiques utiles ; il est nécessaire de construire des lexiques liés au domaine pour pouvoir disposer des relations sémantiques aussi précises et fines que celles requises par le logiciel.

5.4 EIAO (Assistants intelligents pour l'enseignement)

Participants : Jacques Nicolas, Dominique Py, Romuald Texier.

5.4.1 Individualisation des logiciels de formation

Participants :
Jacques Nicolas, Dominique Py, Romuald Texier.

Nous menons une collaboration avec la société IDP qui développe des dispositifs de formation professionnelle pour adultes. La thèse de Romuald Texier porte sur l'intégration d'assistants intelligents au sein des logiciels d'auto-formation. L'objectif est de proposer une approche générique apprentissage/coopération permettant la conception d'outils d'aide à l'apprenant, pour les logiciels de formation. Cette année, Romuald Texier a défini plus précisément la problématique de l'individualisation au sein des architectures logicielles destinées à la formation et recouvrant des contenus disciplinaires hétérogènes. Deux pistes de recherches sont favorisées :

d'une part, la modélisation de la motivation, facteur important de l'autodirection, dont la prise en compte dans les modèles de l'apprenant a peu été étudiée; d'autre part la conception de modèles génériques de connaissances intégrant les notions de motivation et de métacognition. La plateforme de formation à distance, en cours de développement à IDP, constitue un domaine d'application privilégié pour ce travail.

5.4.2 Interaction dans les EIAO de calcul formel

Participants :
Jacques Nicolas, Dominique Py, Romuald Texier.

L'arrivée d'outils de calcul formel fiables et performants dans l'enseignement des mathématiques pose la question de leur adéquation aux situations pédagogiques. Dans le cadre d'un projet piloté par l'INRP, nous nous intéressons à la conception d'environnements d'apprentissage basés sur des outils de calcul formel, et à la modélisation de l'interaction au sein de ces environnements. Le groupe de Rennes travaille plus précisément sur le domaine de l'étude des variations d'une fonction réelle. Dans un premier temps, nous avons analysé l'interaction d'élèves de première avec le logiciel Derive dans le cadre de tâches d'étude de fonctions, et nous avons montré comment l'élève navigue entre différents registres (graphique, numérique, symbolique, papier-crayon).

Cette année, notre groupe a poursuivi la réalisation d'un environnement logiciel dédié à l'étude des variations d'une fonction rationnelle définie et continue sur un intervalle fini. L'environnement est conçu comme une surcouche d'un logiciel de calcul formel, le noyau de Derive. Dans cette maquette, l'élève a pour tâche de remplir un tableau de variations interactif, en s'aidant d'outils d'exploration et de transformation, puis de justifier les valeurs entrées dans ce tableau. Nous avons particulièrement étudié la notion de preuve. En effet, les logiciels de calcul formel ne permettent pas à l'élève de construire une preuve. Notre hypothèse est que l'introduction d'une activité de preuve explicite dans l'environnement permet à l'élève d'élaborer des techniques de démonstration et favorise sa réflexion sur les résultats produits. Une preuve est considérée comme une enchaînement de preuves élémentaires, portant sur les propriétés des expressions manipulées. L'élève a la possibilité de conjecturer et d'opérer des justifications aussi bien à partir de conjectures que de propriétés établies. Le logiciel conserve la trace des déductions intermédiaires et valide automatiquement les résultats déduits de conjectures justifiées. La maquette, développée en Visual C++, est en cours d'expérimentation.

5.5 Raisonnements et logiques non classiques

Participants : Philippe Besnard, Yves Moinard, Dominique Py, Raymond Rolland.

Mots clés : logiques non classiques, logique modale, logique temporelle.

Glossaire :
circonscription logique de modèles minimaux particulière décrivant précisément l'ajout automatique d'axiomes formalisant la notion d'exception.

inférence préférentielle logique de modèles minimaux étendue où on s'autorise à considérer une relation non plus directement sur les modèles mais sur des états, ou copies de modèles, ou encore sur des ensembles de modèles.

5.5.1 Révision de connaissances pour le dialogue coopératif

Participants :
Yves moinard, Dominique Py, Philippe Besnard.

Il s'agit d'interpréter les requêtes d'un utilisateur dans un système de dialogue coopératif homme-machine. Des méthodes d'intelligence artificielle de *révision des connaissances* et de *raisonnement par défaut* ont donc été mises en oeuvre. La coopération a démarré en avril 1997 et a duré trois ans (cf. 6.8)

5.5.2 Inférence préférentielle et Circonscription

Participants :
Yves Moinard, Raymond Rolland.

L'étude de la notion la plus générale d'inférence préférentielle nous a permis de compléter nos précédants résultats. En effet, nous avons maintenant une caractérisation en termes de propriétés logiques de toutes les variantes de cette notion utilisées dans la littérature ([35], et, en combinaison avec nos précédants résultats [43]). Nous décrivons ici ce que signifie cette dernière phrase.

La variante la plus générale utilise une relation binaire sur des "états" qui sont des copies d'ensembles de modèles. Un ensemble de données est traduit par un ensemble de formules de logiques classiques, correspondant aux données certaines, et par une relation binaire entre états, qui permet d'obtenir des conclusions supplémentaires dites "par défaut" (qu'il est raisonnable de croire au vu de l'ensemble des données sures dont on dispose, mais qui pourront être rétractées au vu de données sûres supplémentaires). Précisément, on obtient les conclusions supplémentaires en ne conservant, parmi les états associés à l'ensemble de formules, que ceux qui sont minimaux pour la relation. En fait, la plupart des études portent sur les variantes les plus "simples" de l'inférence préférentielle, ne considérant que les singletons comme ensembles de modèles. Or, nous avons établi un double résultat qui peut faire douter de la pertinence de cette restriction dans tous les cas. Nous avons en effet montré que la variante la plus générale correspond exactement aux inférences qui satisfont une propriété naturelle, souvent considérée comme utile, appelée "cumulativité transitive": si on peut déduire A de B alors, on ne peut rien déduire de plus si on ajoute A à B. Une conséquence de cette caractérisation montre qu'admettre des copies d'ensembles de modèles n'ajoute rien à la notion plus simple d'inférence préférentielle dont les états sont des ensembles, sans nécessiter de copies.

L'étude plus particulière de la circonscription a porté sur deux aspects. D'une part, nous avons approfondi [17] nos résultats précédants sur l'expressivité de la circonscription et sur

les relations entre la circonscription classique, fondée sur l'inclusion ensembliste, et la circonscription par cardinalité. D'autre part, nous avons sensiblement amélioré nos résultats sur les ensembles de formules qui fournissent la même circonscription [32, 33] et en particulier sur les ensembles de formules les plus petits permettant de dérire une circonscription donnée [34]. Ces deux études ont un double but: en représentation des connaissances, il s'agit de savoir si la circonscription est appropriée, et si oui laquelle. En automatisation de la circonscription, il s'agit d'examiner si certains des ensembles équivalents peuvent simplifier les calculs. Rappelons enfin que nos résultats sur les inférences préférentielles les plus générales montrent que les plus fréquemment utilisées peuvent s'exprimer en termes de circonscription, ce qui augmente encore l'utilité de "circonscripteurs automatiques" efficaces.

6 Contrats industriels (nationaux, européens et internationaux)

6.1 Inférence grammaticale régulière pour l'apprentissage de la syntaxe en reconnaissance de la parole

Participants :
Jacques Nicolas, Laurent Miclet, François Coste.

Il s'agit d'une convention CTI avec FT R&D (CCTP LAA/TSS/RCP/860) de décembre 1997 à décembre 2000. En collaboration avec l'équipe Cordial de Lannion (Irisa-Enssat) et l'université de Saint-Étienne, le projet cherche à améliorer les techniques de reconnaissance de la parole continue actuellement développées à FT R&D. Il s'intéresse pour cela à un modèle de langages contraignant l'ensemble des phrases admissibles par la reconnaissance. L'objectif est de passer de la modélisation actuelle par trigramme à une modélisation par automates stochastiques réguliers construits de manière automatique par inférence grammaticale à partir d'exemples.

6.2 Conception et contrôle de stimulateurs-défibrillateurs cardiaques intégrés

Participants :
Marie-Odile Cordier, René Quiniou.

L'Action Concertée Incitative 8899 *Télémédecine et Technologies pour la Santé* du MENRT, d'une durée de 2 ans, réunit le département de Cardiologie du CHU de Rennes, Ela-Recherche, le LTSI de l'université de Rennes 1 et l'Irisa. Elle a pour objectif l'amélioration des prothèses cardiaques notamment en leur apportant des capacités multisites (contrôle à partir plusieurs sondes) et multifonctions (capacité à gérer des problèmes hémodynamiques et rythmiques). Le projet Aïda est chargé d'affiner la classification des arythmies en utilisant des nouveaux électrogrammes issus d'implantation multisites de sondes et la conception de nouveaux algorithmes de contrôle basés sur la technique de reconnaissance de scénarios.

6.3 Modélisation, diagnostic et supervision de réseaux de télécommunication

Participants :
Marie-Odile Cordier, Laurence Rozé, Emmanuel Mayer, Yannick Pencolé.

La convention CTI avec FT R&D concernant la surveillance de réseaux de télécommunications se poursuit en coopération avec le LIPN. La participation du projet Aïda se focalise sur deux points :

- L'acquisition automatique de scénarios de pannes discriminants. Nous avons choisi d'utiliser pour cela des techniques d'apprentissage automatique de type PLI. Le principe consiste à rechercher, pour chaque panne, un scénario discriminant qui accepte les séquences d'alarmes correspondant à la panne et rejette les séquences d'alarmes relatives aux autres pannes. Les séquences d'alarmes sont obtenues en simulant le modèle à base d'automates communicants décrivant le fonctionnement du réseau. Ce travail s'inscrit dans le cadre de la réalisation du projet GASPAR (module de discrimination).

- La construction d'automates diagnostiqueurs. Ce travail a pour objectif de construire, à partir du modèle de fonctionnement du système de gestion d'alarmes, un automate capable d'analyser les alarmes reçues par le superviseur et d'en inférer les pannes possibles. La principale difficulté est liée à la taille de cet automate et nous étudions la construction de diagnostiqueurs génériques, profitant de la structure hiérarchique du réseau, ainsi que de diagnostiqueurs décentralisés.

Ces travaux sont expérimentés sur le réseau de transmission de données Transpac ainsi que sur le réseau ATM par le LIPN. De plus, un projet a démarré dans le cadre des projets RNRT en collaboration avec Alcatel CIT, FT R&D et Ilog côté industriels, avec le LIPN/Université Paris-Nord côté universitaires et au sein de l'IRISA en collaboration entre les projets Pampa, Sigma2 et Aïda. Ce projet a pour nom MAGDA (Modélisation et Apprentissage pour une Gestion Distribuée des Alarmes) et a pour objectif l'étude d'une chaîne complète de supervision d'un réseau de télécommunication. Il s'agit de développer et d'expérimenter de nouvelles méthodes de gestion des alarmes et, plus précisément, de permettre une meilleure compréhension des défaillances ou des pannes, à l'aide d'outils d'acquisition d'expertise (modélisation, apprentissage), puis de reconnaître en ligne des situations à risques par des outils de corrélation d'alarmes et de diagnostic. Aïda est plus particulièrement concerné par le développement des outils de diagnostic. L'approche *diagnostiqueur*(voir 5.1.4) a été étudiée dans ce contexte et une définition de diagnostiqueurs décentralisés plus adaptés à cette application est en cours de développement.

6.4 Développement d'assistants intelligents au sein des logiciels de formation professionnelle

Participants :

Dominique Py, Romuald Texier, Jacques Nicolas.

Collaboration avec la société IDP (Ingéniérie et développement en pédagogie), à Rennes, pour le développement d'assistants intelligents au sein des logiciels de formation professionnelle. Cette coopération est matérialisée par une bourse Cifre (R. Texier).

6.5 L'interaction dans les EIAO intégrant des instruments de calcul formel

Participant :
Dominique Py.

Participation au contrat INRP "L'interaction dans les EIAO intégrant des instruments de calcul formel" dont l'objet est de concevoir des environnements d'apprentissage autour de logiciels de calcul formel.

6.6 Acquisition automatique d'éléments du Lexique Génératif pour améliorer les performances de systèmes de recherche d'information

Participant :
Pascale Sébillot.

Il s'agit d'un contrat de 2 ans obtenu en décembre 98 dans le cadre des Actions de Recherche Partagée de l'AUF, thème 1 : Ressources Linguistiques et évaluation/outils informatiques et formalismes linguistiques. En collaboration avec Pierrette Bouillon (ISSCO Genève), Laurence Jacqmin (Université Libre de Bruxelles) et Cécile Fabre (ERSS Toulouse), le projet a pour objectif de déterminer, grâce à l'utilisation de méthodes d'apprentissage sur des corpus étiquetés, les prédicats témoins des différentes facettes sémantiques des noms d'un domaine, telles qu'elles sont définies dans le Lexique Génératif de Pustejovsky, et de montrer l'apport de l'usage d'un tel lexique dans un système de recherche d'information.

6.7 Analyse linguistique pour la conception d'un logiciel d'aide à l'argumentation

Participants :
Pascale Sébillot, Philippe Besnard.

Il s'agit d'un contrat avec Thomson-CSF Communications Gennevilliers, de 5.5 mois obtenu en février 2000. Son objectif est de modéliser les arguments proposés par Perelman dans son Traité de l'argumentation et de produire une maquette de logiciel qui, étant donné le début d'un argument, produit automatiquement sa fin. Par exemple, un argument de type réciprocité peut être utilisé dans une conversation pour convaincre son auditoire ; à la saisie de la première partie de l'argument, par exemple "Si les vendre n'est pas honteux pour vous", la fin doit s'afficher, à savoir ici "les acheter ne l'est pas non plus pour nous". Une représentation schématique

des arguments sous forme de transformation de triplets avec contraintes sémantiques sur les éléments de ces triplets est utilisée.

6.8 Définition et mise en œuvre d'une théorie de la révision des croyances dans le contexte d'un dialogue coopératif

Participants :
Philippe Besnard, Yves Moinard.

Contrat FT R&D 97 1B 046 de mars 1997 à mars 2000 dont l'objectif est de définir une théorie de la révision des croyances pour un agent rationnel dialoguant avec un utilisateur des services audiotel. L'acquisition des modèles nécessaires passe par l'utilisation de techniques de révision des connaissances et de raisonnement par défaut. Il s'agit de décrire précisément le processus cognitif de changement d'état mental et d'identifier les opérations logiques mises en jeu, de manière à spécifier complètement le processus logique de reconstruction des croyances.

7 Actions régionales, nationales et internationales

7.1 Actions régionales

Nous avons développé cette année une forte action de prospective en faveur d'une réorientation d'une partie du projet vers la bio-informatique. Ceci s'est traduit concrètement par deux actions importantes :

- la création d'un nouveau DEA "Informatique et Génomique" dans lequel interviennent de façon significative les membres de l'équipe. J. Nicolas a fortement participé à sa conception et en est actuellement responsable adjoint (responsables C. Delamarche et D. Lavenier).

- un projet de génopole Ouest dans lequel l'équipe tient une place centrale. Le projet implique les régions Bretagne et Pays de Loire, et coordonne les actions des différents laboratoires impliqués en génomique, post-génomique et bioinformatique dans le Grand Ouest. Il est financé en grande partie dans le cadre des contrats de plan Etat-Région. Sont particulièrement actives les villes de Rennes, Nantes, Angers et Roscoff/Brest. J. Nicolas est responsable de la plate-forme de bio-informatique qui sera en charge de la gestion et du traitement des données de la génopole. Des premiers financements viennent d'être obtenus dans le cadre du BQR Université, d'un soutien du réseau national IMPG ainsi qu'un appel d'offre spécifique CNRS/INRA/INSERM/INRIA sur la bio-informatique. Ce projet doit être expertisé au début de l'année 2001 par un jury international.

Ces projets ont également été l'occasion d'établir de nombreuses coopérations avec les laboratoires de biologie rennais, principalement : Inserm (U435 et U522), Inra (Labo de Génétique Animale et labo SCRIBE) et CNRS (UMR6026, UPR41).

7.2 Actions nationales

- Participation au groupe IMALAIA du GdR Automatique et du GdR-PRC I3 et groupe de travail AFIA(M.-O. Cordier)

- Participation à l'action concertée Remag de recherche de motifs dans les séquences génétiques (J. Nicolas, F. Coste, D. Fredouille) : `http://www.loria.fr/projets/REMAG`.

- Participation au groupe IHMC du GdR-PRC I3 (D. Py)

- Participation au groupe de travail *A3CTE* : Application, Apprentissage, Acquisition de Connaissances à partir de Textes Électroniques du GdR-PRC I3 (P. Sébillot)

- Participation au groupe Colex (centre-ouest lexique) pour l'étude de la structuration d'un lexique pour l'anglais (P. Sébillot)

7.3 Réseaux et groupes de travail internationaux

- Participation au réseau d'excellence européen MONET (Model-Based and Qualitative Reasoning) (M.-O. Cordier). M.-O. Cordier est membre du «Industrial Liaison and Dissemination Committee» du réseau d'excellence européen MONET (Model-based and Qualitative Reasoning).

7.4 Relations bilatérales internationales

- Projet PROCOPE no 99027 «Fondations pour le traitement de contradictions dans les systèmes d'information intelligents» entre l'université de Potsdam et l'IRISA (Ph. Besnard, M.-O. Cordier)

7.5 Accueils de chercheurs étrangers

- Visite de Torsten Schaub (université de Potsdam) pendant une semaine en octobre 2000.

- Invitation de Edward Stabler (Los Angeles) pendant trois jours en mai 2000 (apprentissage de grammaires minimalistes).

- Accueil de Roberto Bonato (Università di Verona, échange Erasmus), mémoire de « Tesi di Laurea» sur l'apprentissage des grammaires catégorielles (octobre 1999 – février 2000).

8 Diffusion de résultats

8.1 Animation de la communauté scientifique

- M.-O. Cordier est co-responsable du groupe IMALAIA du GdR Automatique, du GdR-PRC I3 et groupe de travail AFIA.

- M.-O. Cordier est rédactrice en chef de RIA (*Revue d'intelligence artificielle*) et membre du comité de rédaction de AAI (*journal of Applied Artificial Intelligence*); membre du comité de programme de DX'99; membre du comité de programme de ECAI'2000; responsable de l'organisation des workshops de l'ECAI'2000.
- I.-C. Lerman est éditeur associé de la revue *RO-Operations Research*, membre des comités de rédaction des revues suivantes : *Mathématique, & sciences humaines* (édité par le centre d'Analyse et de Mathématiques Sociales) ; *La revue de modulad* (Editeur Inria).
- I.-C. Lerman est membre des sociétés organisatrices de l'IFCS 2000 (7th Conference of the International Federation of Classification Societies, 11-14 juillet 2000, Namur, Belgique.
- I.-C. Lerman est membre du comité scientifique du XII International Symposium on Applied Mathematical Methods to the Sciences, Costa Rica, 11-14 janvier 2000.
- L. Miclet et J. Nicolas ont été membres du comité de programme des conférences ICGI'2000 (Lisbonne, sept. 2000) et CAP2K (St-Etienne, juin 2000).
- R. Quiniou est trésorier-adjoint de l'AFIA et modérateur du «bulletin électronique de l'AFIA».
- R. Quiniou a été membre du comité de pilotage RFIA'2000.
- P. Sébillot a été membre du comité de programme de RFIA'2000.
- P. Sébillot est membre du comité de lecture de la revue In Cognito.
- P. Sébillot a été membre du comité de lecture du numéro spécial de la revue TAL "Traitement automatique des langues pour la recherche d'information".
- I.-C. Lerman est membre du comité de programme des journées EGC'2001, "Extraction et Gestion des Connaissances", 18-19 janvier 2001, Nantes.

8.2 Enseignement universitaire

- Option du DEA d'informatique, du DESS-ISA IFSIC et 5^e année Insa-Rennes : *module RATS : raisonnement temporel et spatial* (M.-O. Cordier, Y. Moinard, R. Quiniou).
- Option du DEA d'informatique IFSIC et 5^e année Insa-Rennes : *module CLAP : classification et apprentissage* (I.C. Lerman, L. Miclet).
- Cours en DIIC3 IFSIC : *images numériques : approche statistique de la reconnaissance des formes* (I.C. Lerman).
- Cours en $5^{ème}$ année d'informatique de l'Insa de Rennes : *traitement automatique des langues* (P. Sébillot).
- Cours en DEA "Informatique et génome", école doctorale Vie-Santé de l'université de Rennes 1 (J. Nicolas, F. Coste, I.C. Lerman, B. Tallur)
- Encadrement de projets de maîtrise IFSIC (J. Nicolas).

8.3 Participation à des colloques, séminaires, invitations

- Conférence invitée de J. Nicolas et Y. Mescam à la journée Science Art Multimédia : «Bioinformatique et génomique : l'informatique et la découverte des mécanismes du vivant au niveau moléculaire», Saint-Brieuc, 27 janvier 2000.

- Conférence invitée de J. Nicolas au 1er colloque sur la formation de technicien supérieur en Bio-informatique «Caractérisation et discrimination d'ensembles de séquences génomiques», Clermont-Ferrand, 16-17 mars 2000.

- Exposés de F. Coste, J. Nicolas, et P. Sébillot lors du workshop GRAL, Grammar and Logic: natural language analysis, generation and learning, ("Grammatical inference", "Learning categorial grammars" et "Acquiring Elements of Pustejovsky's Generative Lexicon using Inductive Logic Programming"), Rennes, 2-5 mai 2000.

- Conférence invitée de I.-C. Lerman au workshop "Dealing with structured data in machine learning and statistics" attached to ECML2000 (European Conference on Machine Learning), 30 mai 2000.

- Séminaire de P. Sébillot, au Loria de Nancy, invitée par l'équipe Langue et Dialogue, sur le thème "Apprentissage en corpus d'éléments d'un lexique génératif", juillet 2000.

- Exposé de J. Nicolas dans le cadre de la journée anniversaire CNRS/Inria à l'Irisa "Perspectives en bio-informatique : recherche de motifs dans les séquences biologiques", 13 octobre 2000.

- Conférences de J. Nicolas dans le cadre de la semaine "fête de la science" ("La bioinformatique : défis, difficultés et espoirs") 16-22 octobre 2000

- Participation de R. Quiniou au colloque Acm/Mda (Acquisition Conduite par le Modèle) en novembre 2000 à Grenoble (présentation d'un poster).

9 Bibliographie

Ouvrages et articles de référence de l'équipe

[1] P. Besnard, M.-O. Cordier, «Explanatory Diagnoses and their Characterization by Circumscription», *Annals of Mathematics and Artificial Intelligence 11*, 1994, p. 75–96.

[2] P. Boucher, P. Sébillot, « Interprétation et génération automatiques de noms composés anglais à l'aide de formes logiques», *Traitement Automatique des Langues 34*, 2, 1993, p. 89–104.

[3] P. Bouillon, C. Fabre, P. Sébillot, L. Jacqmin, «Apprentissage de ressources lexicales pour l'extension de requêtes», *TAL (traitement automatique des langues), numéro spécial traitement automatique des langues pour la recherche d'information 41*, 2, 2000.

[4] M.-O. Cordier, P. Siégel, «Prioritized transitions for Updates», *in : Proceedings of European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, C. Froidevaux, J. Kohlas (éditeurs), *LNAI 946*, Springer, p. 142–151, 1995.

[5] I. Lerman, « Construction d'un indice de similarité entre objets décrits par des variables d'un type quelconque. Application au problème de consensus en Classification», *Revue de Statistique Appliquée XXXV*, 2, 1987, p. 39–60.

[6] I. Lerman, « Conception et analyse d'une famille de coefficients statistiques d'association entre variables relationnelles I, II», *Revue Mathématiques, Informatique et Sciences Humaines 30*, 118 et 119, 1992, p. 35–52, 75–100.

[7] Y. Moinard, R. Rolland, « Around a Powerful Property of Circumscriptions», *in : Actes de JELIA'94, LNCS No 838*, Springer-Verlag, p. 34–49, 1994.

[8] Y. Moinard, R. Rolland, «Preferential entailments for circumscriptions», *in : KR'94*, J. Doyle, E. Sandewall, P. Torasso (éditeurs), Morgan Kaufmann, p. 461–472, Bonn, mai 1994.

[9] Y. Moinard, R. Rolland, « Propositional circumscriptions », *rapport de recherche*, INRIA Research Report RR-3538, également Publication Interne IRISA 1211, Rennes, France, octobre 1998, `http://www.irisa.fr/EXTERNE/bibli/pi/1211/1211.html`.

[10] S. Thiébaux, M.-O. Cordier, O. Jehl, J.-P. Krivine, « Supply Restoration in Power Distribution Systems — A Case Study in Integrating Model-Based Diagnosis and Repair Planning», *in : Actes de UAI-96*, p. 525–532, 1996.

Thèses et habilitations à diriger des recherches

[11] F. Coste, *Apprentissage d'automates classifieurs en inférence grammaticale*, thèse de doctorat, université de Rennes 1, jan 2000.

[12] C. Largouët, *Aide à l'interprétation d'une séquence d'images par la modélisation du système observé. Application à la reconnaissance de l'occupation du sol*, thèse de doctorat, Université de Rennes I, nov 2000.

Articles et chapitres de livre

[13] P. Bouillon, C. Fabre, P. Sébillot, L. Jacqmin, «Apprentissage de ressources lexicales pour l'extension de requêtes», *TAL (traitement automatique des langues), numéro spécial traitement automatique des langues pour la recherche d'information, à paraître 41*, 2, 2000.

[14] I.-C. Lerman, V. Rouat, *Data Analysis*, Springer, 2000, ch. New Results in Cutting Seriation for Approximate #SAT.

[15] I.-C. Lerman, F. Rouxel, « Comparing classification tree structures: A special case of comparing q-ary relations II», *RAIRO Operations Research 34*, 3, July/Sept 2000, p. 251–281.

[16] I.-C. Lerman, «Comparing taxonomic data», *Revue Mathématiques et Sciences Humaines*, 151, Nov 2000, p. 15 pages.

[17] Y. Moinard, « Note about cardinality-based circumscription», *Artificial Intelligence 119*, 1-2, May 2000, p. 259–273, `http://www.elsevier.nl:80/inca/publications/store/5/0/5/6/0/1/`.

Communications à des congrès, colloques, etc.

[18] P. Besnard, M.-O. Cordier, «Explications causales», *in : Proceedings of RFIA'2000*, p. 169–177, feb 2000.

[19] M.-O. Cordier, P. Dague, M. Dumas, F. Lévy, J. Montmain, M. Staroswiecki, L. Travé-Massuyès, « AI and Automatic Control Theory approaches of model-based diagnosis : links and underlying hypotheses », *in : Proceedings of Safeprocess'2000*, p. 274–279, june 2000.

[20] M.-O. Cordier, P. Dague, M. Dumas, F. Lévy, J. Montmain, M. Staroswiecki, L. Travé-Massuyès, « AI and Automatic Control Theory approaches of model-based diagnosis : links and underlying hypotheses», *in : Proceedings of the Eleventh International Workshop on Principles of diagnosis (DX'00)*, p. 33–40, june 2000.

[21] M.-O. Cordier, P. Dague, M. Dumas, F. Lévy, J. Montmain, M. Staroswiecki, L. Travé-Massuyès, «A comparative analysis of AI and control theory approaches to model-based diagnosis», *in : Proceedings of ECAI'2000*, p. 136–140, august 2000.

[22] F. Coste, D. Fredouille, «Efficient ambiguity detection in C-NFA, a step toward inference of non deterministic automata», *in : ICGI 2000, Grammatical inference: algorithms and applications*, A. L. Oliveira (éditeur), p. 25–38, Lisbonne, september 2000.

[23] F. Coste, «De l'inférence régulière à l'apprentissage d'automates classifieurspour la discrimination de séquences», *in : Conférence d'Apprentissage 2000 (CAp 2000)*, june 2000.

[24] D. Fredouille, « Expériences sur l'inférence de langage par spécialisation», *in : CAp 2000*, p. 117–130, juin 2000, `http://www.irisa.fr/aida/aida-new/dataFiles/dfredoui/cap2000.ps`.

[25] I. Grosclaude, R. Quiniou, «Dealing with interacting faults in temporal abductive diagnosis», *in : DX'2000: International workshop on model-based diagnosis*, june 2000, `http://www.irisa.fr/aida/aida-new/dataFiles/quiniou/dx2000.pdf`.

[26] I. Grosclaude, «Diagnostic abductif temporel de pannes interagissant», *in : RJCIA,00(5ème Rencontres nationales des Jeunes Chercheurs en Intelligence Artificielle)*, Lyon, France, 2000.

[27] C. Largouët, M.-O. Cordier, « Combining Observations and Expectations: Application to the Refinement of an Image Sequence», *in : Workshop UAI*, Standford, CA, USA, 30 Juin 2000, `http://www.irisa.fr/aida/aida-new/dataFiles/clargoue/uai.ps`.

[28] C. Largouët, M.-O. Cordier, «Improving the landcover Classification using Domain Knowledge», *in : XIX th Congress of the International Society for Photogrammetry and Remote Sensing (ISPRS)*, *4B*, 538-545, Amsterdam, 17-23 Juillet 2000, `http://www.irisa.fr/aida/aida-new/dataFiles/clargoue/isprs.ps`.

[29] C. Largouët, M.-O. Cordier, «Improving the Landcover Classification using Domain Knowledge », *in : BESAI'2000 (Workshop on Binding Environmental Sciences and Artificial Intelligence)*, p. (6–1)–(6–7), Berlin, Allemagne, Août 2000, `http://www.irisa.fr/aida/aida-new/dataFiles/clargoue/besai.ps`.

[30] C. Largouët, M.-O. Cordier, «Modélisation par automate temporisé pour aider à l'identification de l'occupation du sol», *in : RFIA'2000 : Reconnaissance des Formes et Intelligence Artificielle, II*, p. 285–294, Paris, France, 1-3 Février 2000, `http://www.irisa.fr/aida/aida-new/dataFiles/clargoue/rfia.ps`.

[31] C. Largouët, M.-O. Cordier, « Timed Automata Model to Improve the Classification of a Sequence of Images », *in : ECAI'2000 (European Conference on Artificial Intelligence)*, p. 156–160, Berlin, Allemagne, 20-25 Août 2000, `http://www.irisa.fr/aida/aida-new/dataFiles/clargoue/ecai.ps`.

[32] Y. Moinard, R. Rolland, « Ensembles de formules équivalents pour la circonscription », *in : RFIA'2000 : Reconnaissance des Formes et Intelligence Artificielle*, R. Deriche, M.-C. Rousset (éditeurs), p. II: 189–198, Paris, Février 2000, `http://www.irisa.fr/aida/aida-new/dataFiles/moinard/circetouRFIA.ps`.

[33] Y. Moinard, R. Rolland, « Equivalent sets of formulas for circumscriptions », *in : 14th European Conference on Artificial Intelligence*, W. Horn (éditeur), IOS Press, Amsterdam, p. 479–483, Berlin, August 2000, `http://www.irisa.fr/aida/aida-new/dataFiles/moinard/equivecpubel.ps`.

[34] Y. Moinard, R. Rolland, «Smallest Equivalent Sets for Finite Propositional Formula Circumscription », *in : First International Conference on Computational Logic*, J. Lloyd, al. (éditeurs), *LNAI*, 1861, Spinger-Verlag, p. 897–911, London, July 2000, `http://www.irisa.fr/aida/aida-new/dataFiles/moinard/equiv3CLsite.pdf`.

[35] Y. Moinard, «Characterizing general preferential entailments», *in : 14th European Conference on Artificial Intelligence*, W. Horn (éditeur), IOS Press, p. 474–478, Berlin, August 2000, `http://www.irisa.fr/aida/aida-new/dataFiles/moinard/cargpecpubel2.ps`.

[36] Y. Pencolé, « Approche diagnostiqueur décentralisé : application aux réseaux de télécommunication », *in : RJCIA'2000 (5èmes Rencontres nationales des Jeunes Chercheurs en Intelligence Artificielle)*, Lyon, France, september 2000, `http://www.irisa.fr/aida/aida-new/dataFiles/ypencole/RJCIA2000.pdf`.

[37] Y. Pencolé, «Decentralized diagnoser approach: application to telecommunication networks», *in : working notes of the 11th International Workshop on Principles of Diagnosis DX'00*, p. 185–192, june 2000, `http://www.irisa.fr/aida/aida-new/dataFiles/ypencole/DX00.pdf`.

[38] D. Py, « Exploration guidée dans un tuteur intelligent », *in : RFIA'2000 : Reconnaissance des Formes et Intelligence Artificielle*, Paris, feb 2000.

[39] P. Sébillot, P. Bouillon, V. Claveau, C. Fabre, L. Jacqmin, J. Nicolas, «Apprentissage en corpus de couples nom-verbe pour la construction d'un lexique génératif», *in : JADT 2000 (journées d'analyse de données textuelles)*, Lausanne, Suisse, mars 2000.

[40] P. Sebillot, P. Bouillon, C. Fabre, « Inductive Logic Programming for Corpus-Based Acquisition of Semantic Lexicons», *in : LLL-2000 (learning language in logic)*, Lisbonne, Portugal, septembre 2000.

[41] B. Tallur, « Une Stratégie de classification hiérarchique des éléments d'un tableau de contingence à trois dimensions », *in : Actes des XXXIIe Journées de Statistique*, Mai 2000, `http://www.irisa.fr/aida/aida-new/dataFiles/tallur/fes2000.ps`.

[42] F. Wang, G. Carrault, R. Quiniou, M.-O. Cordier, P. Mabo, « Fusion de méthodes d'apprentissage automatique et de traitement de signal pour la reconnaissance d'arythmies », *in : 10ème Forum des Jeunes Chercheurs du Génie Biologique et Médical*, juin 2000, `http://www.irisa.fr/aida/aida-new/dataFiles/quiniou/itbm.pdf`.

Rapports de recherche et publications internes

[43] Y. Moinard, R. Rolland, « Characterizations of preferential entailments », *rapport de recherche*, INRIA, Research Report RR-3928, IRISA, Publication Interne 1326, Rennes, France, April 2000, `http://www.irisa.fr/bibli/publi/pi/2000/1326/1326.html`.

Divers

[44] P. Besnard, P. Sébillot, « Analyse linguistique pour la conception d'un logiciel d'aide à l'argumentation », mars,avril,juin,juillet 2000, Rapport de la convention de recherche Thomson-CSF - Inria Rennes 1 00 C 0155 00 31327 01 2.

[45] P. Bouillon, C. Fabre, P. Sébillot, L. Jacqmin, « Acquisition automatique d'éléments du Lexique Génératif pour améliorer les performances de systèmes de recherche d'information », mars 2000, Rapport de l'action de recherche partagée AUF - Université de Rennes 1, Convention X/1.20.09.01.1/98.16.1 (réseau Francil).