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

François Coste [CR Inria]
Yves Moinard [CR Inria]
René Quiniou [CR Inria]
Rumen Andonov [Professeur, Université de Valenciennes, en détachement Inria depuis oct. 2001]
Michel Le Borgne [Maître de conférences, université de Rennes 1, en détachement Inria depuis oct. 2001]

Personnel Université de Rennes 1 et autres établissements d'enseignement

Catherine Belleannée [maître de conférences]
Laure Berti [maître de conférences]
Marie-Odile Cordier [professeur, délégation INRIA jusqu'en sept. 2001]
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 jusqu'à août 2001]
Basavaneppa Tallur [maître de conférences]
Raoul Vorc'h [maître de conférences]

Chercheurs doctorants

Vincent Claveau [bourse MENRT]
Daniel Fredouille [bourse MENRT]
Irène Grosclaude [bourse MENRT (jusqu'au 31 août 2001)]
Christine Largouët [AERC Ensar jusqu'au 31 août 2001)]
Yannick Pencolé [bourse MENRT, Ater depuis septembre 2001]
Mathias Rossignol [bourse MENRT (depuis octobre 2001)]
Stéphane Guyetant [bourse BDI CNRS/Région, depuis octobre 2001]
Aurélien Leroux [bourse INRIA/Région, depuis octobre 2001]
Ingrid Jacqmin [bourse MENRT, depuis octobre 2001]
Andre Floeter [cotutelle université de Potsdam, depuis juillet 2001]
Yoann Mescam [bourse INRIA cofinancée SIB Genève, à partir de décembre 2001]

Collaborateur extérieur

Philippe Besnard [DR CNRS, IRIT, Toulouse]

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 chroniques, graphe causal temporel, acquisition de chroniques.

Glossaire :

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

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

reconnaissance de chronique système permettant, à partir d'un ensemble de chroniques décrivant des situations (la base de chroniques), 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.
- 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).

[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.

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 une chronique (ou 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 chroniques. Son principe consiste en un suivi, en fonction des messages reçus, d'un ensemble de chroniques potentiels jusqu'à une reconnaissance complète d'un ou plusieurs d'entre eux. L'apport d'une base de chroniques est, dans ce cas, nécessaire au bon fonctionnement de la supervision. Cette base doit contenir l'ensemble des chroniques 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 chroniques. Les séquences étiquetées sont ensuite généralisées afin d'obtenir un ensemble de chroniques discriminants. Un système de reconnaissance de chroniques 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, deux idées sont étudiées : la généricité et la décentralisation. 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. Une autre possibilité explorée est celle de "diagnostiqueur distribué" afin de pouvoir travailler directement à partir des modèles des composants.

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 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).

[SR96] M. Sebag, C. Rouveirol, « Induction de clauses contraintes », *in : Reconnaissance des formes et intelligence artificielle (RFIA'96)*, p. 706–716, Rennes, 1996.

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

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

.

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* 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

[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.

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 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

[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.

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 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,

[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.

[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.

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

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, 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

[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.

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).

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 chroniques, diagnostiqueur, réseaux de télécommunications, surveillance cardiaque, systèmes naturels, 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 chroniques 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 chroniques 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 chroniques 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, projet MAGDA (contrat RNRT). Une autre approche consiste à *compiler* le modèle du système, représenté par un graphe causal temporel, en un ensemble de chroniques. 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 chroniques est utilisée pour la surveillance, à partir de leur électrocardiogramme, de patients souffrant de problèmes cardiaques. Les chroniques 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

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

sont en cours de validation.

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. Par ailleurs, nos travaux sur l'acquisition d'éléments lexicaux sur corpus nous permettent d'accéder à des informations sémantiques utilisables pour l'extension et la désambiguïsation des index représentant les contenus des documents et des requêtes.*

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 avons travaillé, dans le cadre d'une Action de Recherche Partagée de l'AUF (Agence Universitaire de la Francophonie) terminée fin mai 2001 en collaboration avec Pierrette Bouillon (ISSCO Genève), Laurence Jacqmin (Université libre de Bruxelles) et Cécile Fabre (ERSS Toulouse) à un projet dont l'objectif était 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 Logiciels

5.1 Dyp : logiciel de démonstration d'une approche centralisée du diagnostic de système à événements discrets

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

Mots clés : diagnostic à base de modèle, systèmes à événements discrets, diagnostiqueur.

Résumé : *Dans le cadre du réseau d'excellence européen MONET (Model-based systems and qualitative reasoning - `http://monet.aber.ac.uk`) et de son appel à logiciel, nous avons proposé un logiciel montrant le principe du diagnostic de systèmes à événements discrets à l'aide d'une extension de l'approche diagnostiqueur classique :* Dyp. *A partir d'un modèle de comportement en cas de pannes d'un système, le logiciel présente la transformation de ce modèle en une structure diagnostiqueur adaptée pour l'analyse en continu d'un flot d'événements observés.* Dyp *présente les trois étapes d'une telle approche :*

- *la construction du modèle du système par la composition de modèles élémentaires ;*
- *la construction du diagnostiqueur centralisé basé sur ce modèle ;*
- *l'analyse d'un flot d'observations et l'élaboration d'un diagnostic de pannes sur le système expliquant les observations.*

Ce logiciel a été distribué sur CD-ROM aux partenaires du réseau MONET avec l'accord de l'INRIA et de l'université de Rennes 1.

5.2 DDyp : plateforme de diagnostic décentralisé pour la supervision de réseaux de télécommunications

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

Mots clés : diagnostic à base de modèle, systèmes à événements discrets, algorithmes distribués, supervision de systèmes complexes.

Résumé : *L'approche diagnostic décentralisé (voir section 6.1.2 a été mise en œuvre dans le cadre du projet MAGDA (cf section 7.2) et a donné lieu à un logiciel :* DDyp. *Cette plate-forme fait partie de la chaine de diagnostic nécessaire à la supervision de réseau de télécommunications [33]. Elle a la charge de récupérer les alarmes reçues par un logiciel de gestion de réseau et d'établir un diagnostic du réseau. Ce diagnostic est ensuite envoyé sur un module de visualisation qui a la charge de présenter le résultat de façon ergonomique à l'opérateur de supervision. Cette plate-forme implante tous les aspects de l'approche décentralisée. A partir d'un modèle du réseau à base d'automates communicants, elle construit un ensemble de diagnostiqueurs locaux (phase hors-ligne). Puis, étant en-ligne, elle récupère les alarmes, produit des diagnostics locaux à l'aide des diagnostiqueurs puis fusionne les diagnostics avec une stratégie adéquate pour obtenir efficacement le diagnostic final. La plate-forme a été développée en C++/Java et de manière distribuée par l'utilisation de la technologie CORBA. Chaque tâche de diagnostic (acquisition d'alarmes, calcul d'un diagnostic local, coordination, fusions des diagnostics, interface graphique) est un processus séparé, ce qui permet de déployer la plate-forme en fonction des ressources informatiques disponibles et de la nature du réseau à superviser. Ce logiciel est en cours de finalisation. Une première démonstration a été effectuée lors de la revue finale du projet RNRT-MAGDA où a été présentée la chaîne complète pour la supervision et le diagnostic d'un réseau.*

5.3 caïd : diagnostic temporel à partir d'un modèle causal

Participant : Irène Grosclaude.

Mots clés : diagnostic à base de modèle, réseaux causaux temporels, pannes interactives, reconnaissance de chronique.

Résumé : *La représentation des connaissances de pannes d'un système dynamique est naturelle dans un réseau causal où les nœuds représentent des observations ou des pannes et les arcs une relation de causalité éventuellement affectée de contraintes temporelles entre cause et effet(s). Le logiciel* CAÏD *met en œuvre une procédure d'inférence abductive et déductive dans de tels réseaux pour le diagnostic de pannes. Pour ce faire, il s'appuie, en particulier, sur une propagation efficace des contraintes temporelles. Il permet également de visualiser et de manipuler des réseaux causaux, par exemple, pour la "compilation" des effets observables d'une ou plusieurs pannes (éventuellement interactives) en vue de la production de chroniques. Les chroniques ainsi produites sont ensuite utilisées en ligne par un reconnaisseur de chroniques pour le diagnostic ou le pronostic à base de modèles.*

6 Résultats nouveaux

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

Participants : Marie-Odile Cordier, Irène Grosclaude, Christine Largouët, Véronique Masson, Yannick Pencolé, René Quiniou, Sophie Robin, Laurence Rozé.

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 automatique de chroniques, la construction de diagnostiqueurs en particulier décentralisés et l'interaction diagnostic/décision dans un univers incertain. Les trois thèmes d'application privilégiés sont les suivants :

- *La surveillance de réseaux en particulier de réseaux de télécommunications dans le cadre du projet RNRT MAGDA (en collaboration avec les projets Sigma2 et Pampa, l'université Paris-Nord ainsi qu'Ilog et Alcatel).*
- *Le monitoring médical et en particulier la détection de troubles d'arrythmie cardiaque par l'analyse d'électrocardiogrammes.*
- *La surveillance de l'environnement et en particulier celle de parcelles agricoles à partir d'images satellitales, dans le cadre du programme Bretagne Eau Pure.*

6.1.1 Extension de l'approche diagnostiqueur

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

La méthode des automates diagnostiqueurs s'inspire des travaux de [SSL+94,SSL+95] 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+94,SSL+95] a dû ainsi être adaptée et étendue à ce formalisme ainsi qu'aux exigences de l'application traitée.

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.
- développement d'une approche décentralisée du diagnostiqueur (cf section 6.1.2).

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 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 du projet RNRT MAGDA qui rassemble des industriels et des équipes de recherche sur le thème de l'interprétation des alarmes dans les réseaux de télécommunications (voir section 7.2). La modélisation d'un anneau SDH a ainsi été réalisée et a permis de tester l'approche diagnostiqueur décentralisé décrite dans la section suivante.

Le travail actuel porte sur l'étude et 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 est d'utiliser ces techniques pour construire un diagnostiqueur symbolique et profiter ainsi de l'efficacité associée.

[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.

[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.

6.1.2 Approche décentralisée du diagnostic

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

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 [Pen00b,Pen00a] et dans [29]. Du point de vue de la mise en œuvre du système, un premier prototype a été réalisé et expérimenté sur l'application faisant l'objet du projet MAGDA (projet RNRT)(voir section 7.2). Il s'agit d'un anneau SDH qui a été modélisé et sur lequel un certain nombre de scénarios de pannes ont été testés. Cette mise en œuvre s'appuie sur les bibliothèques du projet *Dyp*, ce qui assure la compatibilité avec les approches centralisées et génériques concernant la description du modèle en entrée du système (éditeur de modèle utilisable dans cette approche).

6.1.3 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

[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.

[Pen00b] 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`.

[Pen00a] 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`.

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.

Les connaissances diagnostiques sont traditionnellement représentées par des graphes causaux liant les pannes à leurs effets. Afin de traiter les interactions non monotones, nous avons étendu le formalisme classique des graphes causaux en permettant l'expression d'effets négatifs. La représentation des règles causales est enrichie par des propriétés précisant si la ou les causes sont instantanées ou continues et si les effets sont persistants ou non. Dans bon nombre de cas, ces propriétés permettent de déduire le résultat de l'interaction de plusieurs pannes ayant des effets opposés.

L'algorithme de diagnostic de panne repose sur un raisonnement abductif qui prend en compte ces interactions [11]. Le graphe causal dans le formalisme étendu, contenant la représentation concise et explicite de tous les phénomènes, causaux ou interactifs, est traduit dans un formalisme intermédiaire basé sur le calcul des événements [24]. Cette représentation procure une sémantique claire au formalisme des graphes causaux étendus. La prise en compte de l'interaction augmente la complexité du diagnostic du fait de la combinatoire introduite. L'efficacité de l'algorithme de recherche de panne est donc un point crucial. Nous l'améliorons par une pré-compilation du modèle intermédiaire 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. Cet outil est utilisé dans le cadre d'une collaboration EDF sur le thème du pronostic pour la maintenance conditionnelle (cf. 7.1).

6.1.4 Pronostic pour la maintenance conditionnelle

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

Le pronostic est une notion essentiellement utilisée en médecine où il concerne la prédiction du cours et des effets d'une maladie sur l'état de santé d'un patient et l'estimation de sa durée de vie. Le pronostic joue un rôle important pour des tâches comme la planification des traitements administrés aux patients, aussi bien chirurgicaux que médicamenteux. La rationalisation de la prise en charge du patient nécessite de considérer la pertinence des interventions et leur coût. Pour des raisons d'économie, les soins administrés à un patient ne reposent plus seulement sur le diagnostic et le traitement de la maladie : les bénéfices attendus pour le patient et pour la communauté (qui finance) interviennent de plus en plus. Concrètement, la qualité de vie, les risques d'une intervention chirugicale, les effets indésirables des médicaments, l'espérance de vie, la récurrence de la maladie, les coûts financiers ou les limitations budgétaires entrent en ligne de compte.

Cette notion de pronostic peut être utilisée de manière quasiment identique dans le cadre de

la maintenance conditionnelle des systèmes industriels. L'objectif de la maintenance conditionnelle est de déterminer les interventions les plus rationnelles prenant en compte les éléments contextuels correspondant à ceux cités plus haut : risques et coût d'une intervention, durée de vie des composants mis en jeu, bénéfices espérés de la réparation, etc.

En milieu industriel le pronostic consiste donc à prédire quel pourra être l'état d'une installation dans le futur au vu de son état actuel et passé. Nous proposons d'utiliser les graphes causaux temporels pour accomplir cette tâche. A partir d'un ensemble d'observations sur le système supervisé il est possible d'émettre des hypothèses sur son état actuel (en état de fonctionnement ou sujet à une panne). Les hypothèses de pannes permettent d'envisager des effets futurs et d'estimer leur date. L'évaluation des effets (importance, dangérosité, ...) et des réparations (coût, difficulté de mise en œuvre) associées à une panne est utilisée ensuite pour déterminer la meilleure action de maintenance conditionnelle.

Cette recherche a pour cadre une collaboration avec EDF (cf. 7.1).

6.1.5 Monitoring en cardiologie

Participants : Marie-Odile Cordier, René Quiniou, 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 [30]. 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 [32].

Dans le cadre du DEA de Céline Fildier, nous avons commencé à étudier l'apprentissage à partir de plusieurs voies (deux voies ECG et une voie hémodynamique) afin d'améliorer la résistance au bruit des chroniques apprises. Nous examinons, de plus, l'adaptation des techniques de reconnaissance de chroniques afin qu'elles prennent en compte les aspects multivoies. Diverses méthodes de contrôle de la reconnaissance sont 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 constituent une première approche d'une collaboration étroite entre traitement

de signal et traitement de haut niveau représenté par la reconnaissance de chroniques.

Cette étude est développée dans le cadre de l'Action Concertée Incitative *Télémédecine et Technologies pour la Santé* du MENRT (cf. 7.3).

6.1.6 Surveillance de parcelles agricoles

Participants : Marie-Odile Cordier, Christine Largouët, Véronique Masson.

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'images [16]. 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 par celle. 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 [22]. 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. Nous avons proposé une extension de la méthode aux probabilités [26] 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. De plus, les données météorologiques ont été intégrées afin de préciser les contraintes temporelles et tenir compte de l'influence climatique.

Profitant de l'expérience acquise lors de ces travaux, nous avons d'autre part proposé l'utilisation des techniques de model-checking pour le diagnostic ([LC00]).

6.2 Apprentissage automatique et structuration de données

Participants : Catherine Belleannée, Laure Berti-Equille, Roberto Bonato, François Coste, Daniel Fredouille, Israël-César Lerman, Aurélien Leroux, Ingrid Jacquemin, Andre Floeter, Yoann Mescam, Jacques Nicolas, Basavaneppa Tallur, Raoul Vorc'h.

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

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.

Nous avons cette année finalisé et concrétisé notre action en faveur d'une réorientation d'une partie du projet vers la bio-informatique. Deux actions importantes doivent être mentionnées :

- *la création d'une huitième génopole, la génopole OUEST, acceptée pour 2 ans. 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 et Roscoff/Brest. J. Nicolas est responsable du comité bio-informatique qui sera en charge de la coordination de la gestion et du*

[LC00] 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`.

traitement des données de la génopole, ainsi que des recherches effectuées en bioinformatique. L'équipe doit bénéficier début 2002 du recrutement de deux ingénieurs experts, utilisés à 50% au service de la génopole et à 50% au service du projet, afin d'assurer sa contribution à la génopole.

- *une candidature déposée pour un projet INRIA de bioinformatique, Symbiose. Les* thèmes scientifiques *sur lesquels se focalise le projet découlent de notre choix de modéliser des systèmes biologiques complexes, en se plaçant dans un cadre linguistique et logique.*

  De façon plus précise, le projet s'articule en trois grandes directions :

  - **Analyse linguistique de séquences**

    Le premier apport des travaux en bioinformatique concerne la recherche des structures spatiales ou logiques pertinentes (i.e. fonctionnelles) dans les macro-molécules, qu'il s'agisse de modéliser des mécanismes biologiques généraux (transcription, épissage, frameshift...) ou des structures spatiales spécifiques (structures secondaires ou tertiaires de familles de protéines). Nous abordons ces problèmes dans le cadre de la théorie des langages, en nous posant des questions à la fois théoriques (comparaison de deux mots, classe utile de langages, classe apprenable) et pratiques (comment construire des analyseurs efficaces, comment inférer des langages à partir d'échantillons de phrases ?). Globalement, notre spécificité est le traitement de données de manière combinatoire, c'est-à-dire en s'appuyant sur le dénombrement de structures similaires pour rassembler ou caractériser plutôt que sur l'estimation ou l'adaptation de paramètres dans des modèles fixés. Les champs disciplinaires abordés sont les grammaires logiques, l'apprentissage automatique et l'analyse de données.

  - **Analyse et identification de systèmes dynamiques**

    Le point précédent s'intéresse à la caractérisation pour chaque organisme de leurs gènes pris individuellement. Le but final reste néanmoins d'intégrer toutes ces données en un modèle de fonctionnement global, explicitant et permettant de simuler les interactions majeures entre gènes dans des environnements donnés. Cette démarche est indispensable à la pleine compréhension de pathologies ou des mécanismes de régulation impliqués dans des chemins métaboliques donnés. Du fait de la difficulté d'obtenir des données expérimentalesprécises et de la complexité des mécanismes étudiés, on ne peut espérer développer actuellement dans ce domaine que des modélisations de nature qualitative. Nous visons l'aide à l'identification de tels modèles. Notre approche consiste à utiliser une modélisation logique, à la valider au moyen de techniques issues de la vérification de circuits et plus généralement de l'automatique, puis à le raffiner par apprentissage automatique. Les champs disciplinaires abordés sont l'analyse de données, la logique, l'apprentissage automatique et l'automatique.

  - **Parallélisme**

    Les traitements combinatoires que nous venons de décrire sur les données génomiques demandent une puissance de calcul d'autant plus grande que

celles-ci sont produites à un rythme soutenu (doublement de la taille des banques publiques tous les ans), aboutissant à des banques de données de plusieurs millions d'objets. L'accès rapide à des sélections complexes de ces données devient alors un enjeu scientifique stratégique. L'objectif principal de cet axe de recherche est de paralléliser ces traitements pour en accélérer fortement l'exécution. La mise en oeuvre vise plusieurs catégories de machines parallèles : les super calculateurs, les grilles de calcul et les machines spécialisées. Les champs disciplinaires abordés sont le parallélisme, le grid computing et l'architecture des machines.

6.2.1 Inférence grammaticale

Participants : Roberto Bonato, François Coste, Daniel Fredouille, Jacques Nicolas.

Notre activité en inférence grammaticale s'est poursuivie selon trois directions : l'inférence de grammaires minimalistes, l'inférence d'automates non déterministes et l'inférence de transducteurs.

La première étude s'inscrit dans le cadre d'une collaboration avec avec C. Rétoré de l'équipe Paragraphe et fait suite au stage de DEA de Roberto Bonato concernant l'inférence grammaticale de grammaires de Lambek. Nous avons proposé cette année [19] l'extension de l'algorithme développé lors du stage de DEA à la classe des Grammaires Minimalistes (rigide). Ces grammaires sont une formalisation mathematique du "Minimalist Program" de Chomsky par Edward Stabler (UCLA University), qui paraît prometteuse par rapport à la possibilité d'integrer dans le processus d'apprentissage des données de nature sémantique.

La seconde étude concerne l'inférence de grammaires régulières. Dans ce cadre, l'inférence d'automates non déterministes (AFNs) est un axe de recherche explorant le mode de représentation des langages réguliers. Considérer l'inférence d'AFNs permet d'obtenir une représentation de certains langages réguliers exponentiellement plus compacte que celle qui serait obtenue par inférence d'automates déterministes (AFDs). Nous avons étendu le travail [CF00] sur l'inférence d'automates classifieurs non déterministes et non ambigus en classification (dans lesquels un mot ne peut avoir qu'une classe) à l'inférence d'automates non ambigus en analyse syntaxique, ou ANAs (dans lesquels un mot ne peut être accepté que par un seul chemin dans l'automate). Les ANAs semblent en effet un bon compromis entre les AFDs, de taille potentiellement très importante, et les AFNs, de petite taille mais peu facilement manipulables. L'espace de recherche pour l'inférence d'ANAs a été caractérisé [23] et l'extension au cas non déterministe de la plate-forme d'apprentissage d'automates existante est en cours. La généricité et l'efficacité (pouvoir traiter des données réelles) visées par la plate-forme nous ont conduits à concevoir une bibliothèque générique de représentation et de manipulation des machines à états finis, implémentée sous la forme d'une extension à la librairie standard de C++. Cette bibliothèque permet une plus grande abstraction des machines à états finis et nous permet d'envisager notamment une extension plus facile des travaux effectués à l'apprentissage de transducteurs,

[CF00] 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.

sujet du paragraphe suivant.

Enfin dans le cadre de la thèse de R. Bonato nous étudions l'application des algorithmes d'inférence de transducteurs au problème de la prédiction de structures secondaires d'une protéine à partir de sa séquence d'acides aminés. Les transducteurs, en tant qu'automates à états finis formalisant un processus de traduction d'un language à un autre, semblent être un outil formel prometteur par rapport à d'autres approches heuristiques actuellement employées. Le travail effectué est axé sur deux points. En premier lieu, nous recherchons les sous classes de transducteurs rationnels suffisantes pour la modélisation des structures auquelles nous nous intéressons. En second lieu, nous étudions les aspects algorithmiques de l'inférence en cherchant à étendre les résultats connus en inférence d'automates finis. Le problème principal réside dans la gestion de transitions sur les mots au lieu de lettres qui entraîne de difficiles problèmes de segmentation.

Nos perspectives de recherche concernent l'application de l'inférence grammaticale aux problèmes de traitement de séquences génomiques. Nous allons ainsi développer la recherche de signatures dans ces séquences par inférence d'automates non déterministes. Concernant l'inférence de transducteurs, notre objectif est d'obtenir des résultats comparables aux algorithmes adaptés au problème de la prédiction de structures secondaires de protéines.

6.2.2 Analyse de la méthode AVL

Participants : Israël-César Lerman.

La méthode AVL (Analyse de la Vraisemblance des Liens) est peut être davantage connue pour la classification de l'ensemble V des variables descriptives que pour celle d'un ensemble O d'objets ou C de catégories décrits au moyen de V. Cependant cette méthode permet avec la même rigueur conceptuelle d'élaborer une classification ascendante hiérarchique sur O (respectivement sur C) et de fournir des coefficients d'"explication" compte tenu de l'organisation de V.

Nous avons repris avec la collaboration de Ph. Peter (Ecole Polytechnique de Nantes) l'analyse conceptuelle et expérimentale de la construction d'un indice de similarité probabiliste entre objets décrits par des variables de types quelconques. Ceci, afin de la comparer le plus finement possible, avec une approche due à W.D.Goodall qui conduit également à un indice probabiliste.

Notre méthode s'avère essentiellement distincte : plus souple, très générale et surtout, tenant étroitement compte de la sémantique des variables, notamment dans le cas qualitatif (conférence invitée à Porto et Publication Interne en cours de préparation).

6.2.3 Critères linéaires de validation d'une classification

Participants : Israël-César Lerman.

Nous avions établi un critère de validation d'une classification, bien fondé sur les plans formel et statistique. Ce critère confronte une partition donnée de l'ensemble E décrit avec une information de nature ordinale ou numérique quant aux ressemblances entre éléments de l'ensemble E. L'expression de ce critère est essentiellement quadratique par rapport à la taille

de E. Ce qui peut limiter son usage dans le cas de "très grosses données" tel qu'il s'en présente dans la Fouille de Données (Data Mining).

Suite à l'extension de l'algorithme des K-MEANS dans le cas de larges ensembles de données où les variables sont numériques ou qualitatives nominales [Hua98], nous avons bâti une expression linéarisée de notre critère qui s'applique dans le cas d'un ensemble d'objets décrits par des attributs numériques, booléens ou qualitatifs nominaux. Les résultats expérimentaux ont dès lors montré le grand intérêt de ce critère. Cette recherche est menée en collaboration avec J.P. Costa (Université de Porto).

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

Participants : 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 avait, notamment dans le cadre de sa thèse (Université de Rennes, 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. Une telle implémentation vient d'être réalisée dans le contexte du critère de l'inertie expliquée. Elle permet -en poussant l'expérimentation- de se rendre compte du gain de performance et des conditions les plus favorables à ce gain. Il s'agit également d'implémenter une telle algorithmque par rapport à la famille unidimensionnellement paramétrée, des critères de l'AVL. Il s'est avéré expérimentalement que pour de tels critères, l'arbre de classification sous contrainte de contiguïté ne présentait pour ainsi dire, pas d'inversions ; alors que des inversions pouvaient "facilement" se produire dans le cas du critère de l'inertie expliquée. À cet égard un résultat théorique vient d'être établi. Cette recherche est menée et doit se poursuivre en collaboration avec K. Bachar (sus cité), par rapport au problème fondamental de la classification hiérarchique de "très gros ensembles".

[Hua98] Z. Huang, « Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values », *Data Mining and Knowledge Discovery, Kluwer 2*, 1998.

6.2.5 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. Il sera formé de 1000 sujets bretons normaux. À cet échantillon sera confronté le moment venu, un ensemble de même taille formé de personnes carencées. Sur chaque individu de l'ensemble d'apprentissage des paramètres quantitatifs tels que le fer sérique, le coefficient de saturation de la transferrine et la ferritine, sont mesurés. D'autre part, on détermine l'haplotype de chacun de ces individus relativement à 15 gènes et à 15 sites polymorphiques par gène, chacun des sites pouvant être occupé par l'un d'entre deux nucléotides. L'étape actuelle est une étape de constitution des données. Cette constitution suppose actuellement une instrumentation et une manipulation longues et délicates. L'acquisition d'un robot par l'UPR41 devrait permettre d'accélérer la collecte des données. L'étape suivante consistera alors à déterminer une stratégie optimale de classification et d'analyse combinatoire des données pour découvrir des régréssions intéressantes.

6.2.6 Classification prédictive de protéines MIP

Participants : Basavanneppa Tallur.

La méthodologie AVL pour la classification hiérarchique est basée sur une notion de similarité probabilistique ou la "vraisemblance du lien" qui peut s'appliquer à des données de types divers en respectant leur représentation mathématique. Les séquences biologiques sont des données de type très particulier et leur classification nécessite la construction d'un indice de similarité approprié. Ce problème a été abordé et les expériences concluantes ont été menées sur différentes familles de séquences. Cet indice de similarité entre séquences dépend de la matrice de similarité entre acides aminés dont il existe plusieurs versions connues et utilisées par les biologistes et les résultats de la classification dépendent de la matrice choisie. Les biologistes sont actuellement intéressés par une vieille famille de protéines, MIP (Major Intrinsic Protein), qui ont typiquement l'une des deux fonctions reconnues : (1) les aquaporines (AQP) qui laissent passer l'eau et (2) les facilitateurs de glycérole (GlpF). Nous avons proposé une méthode de sélection de la meilleure matrice de similarité entre acides aminés pour classifier les séquences MIP et montré expérimentalement son efficacité pour la prédiction de la fonction des protéines de cette famille. L'utilisation des arbres de décision semble intéressante pour la prédiction de fonction. Nous avons mené une expérience avec la méthode de type CART dont les résultats sont prometteurs. Ce travail a donné lieu à une communication aux 8èmes Rencontres de la Société Francophone de Classification [31].

6.2.7 Intégration et nettoyage de données biologiques

Participants : Laure Berti-Equille.

Les banques de données biologiques proposent en masse des données hétérogènes et distribuées souvent redondantes, imprécises et contradictoires, sans traçabilité ni complément d'information quant à leur qualité (absence de méta-données et d'estimations statistiques décrivant la production, la gestion ou l'utilisation des données stockées).

Dans ce contexte, le sujet de recherche consiste à définir des méta-données adaptées, à spécifier des techniques de calculs et d'extraction pour faciliter l'intégration des données biologiques et en évaluer la qualité.

Dans une démarche d'analyse et de conception d'une base de données spécialisée collectant de façon intensive (par mises à jour quotidienne) des données en masse issues de différentes sources, il s'agit en particulier de proposer :

- une modélisation des méta-données pour les données biologiques,
- une technique permettant d'évaluer et de contrôler la qualité d'une banque de données biologiques,
- une technique permettant le nettoyage et l'intégration des données biologiques,
- le développement d'une plate-forme expérimentale pour le contrôle de la qualité des données biologiques.

Inspiré des travaux sur la qualité de service, l'objectif de cette action est ici de formaliser et de développer des solutions opérationnelles permettant de contrôler dynamiquement la qualité des données et d'orienter l'acquisition et l'intégration des données biologiques [18, 25]. La satisfaction d'un certain nombre de contraintes sur la qualité d'une source de données doit permettre le choix des données à accéder (et à rapatrier) ainsi que le choix de la stratégie la plus adaptée pour la conciliation des conflits de données avant l'intégration.

6.2.8 Qualité des données

Participants : Laure Berti-Equille.

Nous nous intéressons à l'évaluation et au contrôle de la qualité d'une base de données en proposant des techniques de génération automatique de méta-données décrivant la qualité des données stockées (détection des erreurs, incohérences par des contraintes statistiques).

Les activités liées à la veille technologique sont traditionnellement centrées sur la notion de validation de l'information par expertise. Jusqu'à présent aucun système d'information n'assure ni n'assiste l'analyse critique et qualitative de l'information qu'il stocke. Plus généralement, la plupart des systèmes d'information actuels stockent des données (1) dont la source est généralement unique, non connue ou non identifiée/authentifiée et (2) dont la qualité est inégale et/ou ignorée.

En collaboration avec la cellule de veille du Centre Technique des Systèmes Navals (DGA/ECN/CTSN), l'objectif de notre travail est de développer un environnement permettant la gestion des sources textuelles (en majorité des documents techniques), la gestion des données extraites de leur contenu (par structuration sélective et extraction des informations) et l'exploitation de méta-données de qualité par des mécanismes de recommandation adaptative [17].

6.3 Acquisition d'informations lexicales sémantiques sur corpus et applications

Participants : Philippe Besnard, Vincent Claveau, Israël-César Lerman, Jacques Nicolas, Mathias Rossignol, 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. [3]). 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. Ces méthodes d'acquisition de lexiques sur corpus que nous développons sont également la base d'un travail sur l'extraction d'information (interactions géniques) dans des textes de bioinformatique.*

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

Participants : Vincent Claveau, Jacques Nicolas, Pascale Sébillot.

Nous nous intéressons, dans le cadre d'une action de recherche partagée de l'AUF de 2 ans terminée en mai 2001, 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 et la portabilité de la méthode de type programmation logique inductive (PLI) que nous avons développée au cours des 2 années précédentes. Le but de cette méthode est de permettre, dans un corpus de textes techniques étiqueté catégoriellement et sémantiquement (corpus de manuels de maintenance d'hélicoptères fournis par Matra-ccr), 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 oeuvre de la PLI développée par Muggleton, qui produit alors des clauses par généralisation de certains exemples positifs. À l'aide du seul étiquetage catégoriel, 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 (Pearson de 0.52 contre 0.12). La prise en compte en 2000 de l'étiquetage sémantique des mots apparaissant dans le contexte du N et du V nous avait conduits à des résultats d'apprentissage théorique de 0.91 (90% des exemples positifs couverts et seulement 0.7% d'exemples négatifs) et empiriques de 0.58.

Au cours de l'année 2001, nos travaux sur cette méthode d'apprentissage ont porté sur 2 points :

1. Accroissement de l'efficacité de la méthode afin de traiter de manière performante toute l'information contextuelle (catégorielle et sémantique) disponible et éviter des difficultés d'ordre combinatoire. La prise en compte de l'étiquetage sémantique du corpus Matra-ccr en 2000 avait porté sur le seul voisinage des N et des V dans les phrases, et l'apprentissage par Progol fournissait des clauses généralisées concernant un nombre fixe d'éléments de contexte. Nous avons d'une part montré l'intérêt linguistique des clauses apprises [20] et, d'autre part, cherché à exploiter pleinement l'étiquetage sémantique en traitant également les étiquettes du N et du V. Ce volume d'information conséquent à gérer (par exemple 33 étiquettes sémantiques différentes pour les N) et la volonté d'un format plus libre des clauses généralisées nous ont conduits à abandonner Progol au profit d'Aleph, autre algorithme de PLI développé par Muggleton, présentant en particulier l'avantage d'une plus grande paramétrisation. Pour exploiter de manière efficace l'intégralité de l'information contextuelle catégorielle et sémantique, nous avons défini un opérateur de raffinement, basé sur les principes de la theta-subsomption sous identité objet, permettant un parcours plus efficace du treillis de recherche d'hypothèses, et qui est lié à la hiérarchie des étiquettes sémantiques. Cet opérateur nous permet d'obtenir avec Aleph des temps d'exécution similaires à ceux de la seule exploitation des informations catégorielles avec Progol. Enfin, nous avons étudié le regroupement des clauses généralisées obtenues à l'aide de treillis de Galois.

2. Test et évaluation de la portabilité de la méthode d'apprentissage. À l'aide de notre nouvelle méthode d'apprentissage, nous avons tenté d'évaluer le rapport qualité de l'apprentissage réalisé - coût effectif de mise-en-oeuvre et donc de portage d'un corpus à un autre [13]. Le coût de l'étiquetage sémantique manuel des N étant élevé (réalisation manuelle de la classification même si l'étiquetage est lui automatisable), nous avons comparé, sur le corpus Matra-ccr, les résultats d'apprentissage théorique et empirique en prenant en compte d'une part toutes les étiquettes catégorielles et sémantiques - coefficients de Pearson respectifs de 0.78 et 0.61 - et, d'autre part, la même information mais en se limitant à l'étiquetage catégoriel pour les N - 0.89 et 0.64. Ce dernier résultat nous conforte quant à la portabilité de notre méthode. Nous avons également réeffectué l'intégralité de nos étiquetages et expériences d'apprentissage (étiquetage sémantique réalisé par E. Galy et L. Delort de l'université de Toulouse le Mirail) sur un autre corpus moins technique, le corpus Euro qui contient des textes de la Fortisbank de Bruxelles décrivant la mise en place de l'Euro et ses conséquences pour des particuliers. Les résultats d'apprentissage

obtenus sont tout à fait comparables aux précédents, l'apprentissage permettant cependant de mettre là encore au jour des clauses descriptives du concept de paires N-V qualia liées à chaque corpus.

Pour terminer, signalons que, dans le cadre du stage de DES information et documentation de l'université libre de Bruxelles de L. Vandenbroucke, nous avons réalisé une enquête auprès de documentalistes de la Fortisbank pour évaluer la pertinence des couples N-V qualia obtenu par notre système pour leurs requêtes habituelles. Les résultats globalement positifs nous ont donné des pistes pour leur exploitation effective future au sein d'un système de recherche d'information.

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

Participants : Israël-César Lerman, Mathias Rossignol, Pascale Sébillot.

Nous nous intéressons, à l'aide de méthodes de classification, à l'acquisition automatique de lexiques sémantiques basés sur la sémantique différentielle (SD) 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 SD, 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.

Au cours des trois dernières années, nous avons mis en place une méthodologie d'acquisition de lexiques basés sur la SD. 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), nous avons 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 a essentiellement porté sur la consolidation de la première étape de l'acquisition de ces lexiques, à savoir la détection automatique des listes de mots dénotant de la présence d'un thème dans un corpus, mots dont la coprésence permet le découpage du corpus initial en sous-corpus thématiquement homogènes. Nous avons plus particulièrement travaillé à l'amélioration de l'adéquation de la méthode de classification des mots utilisée (AVL) pour pouvoir déterminer, avec le moindre recours possible à un expert, les classes effectivement porteuses d'un thème. Une algorithmique originale de recherche d'une classification valide et signifiante a été mise au point dans le cas d'un gros tableau de contingence particulièrement creux. L'objectif de la classification étant de regrouper les noms dont les distributions au sein des paragraphes sont les plus similaires, le travail, sur notre corpus test du Monde diplomatique, consiste en la classification de l'ensemble des lignes représentant les distibutions de noms à travers un ensemble de paragraphes. La taille du tableau de contingence étudié ici est de 383x8000, et près de 98% des cases sont vides. De plus, les cases non vides sont très faiblement chargées, de l'ordre de quelques unités. La méthode AVL à travers le programme Chavl a été largement mise à contribution, d'abord comme outil de densification du tableau de contingence par regroupement des colonnes représentant l'ensemble des paragraphes. Elle a été ensuite utilisée pour produire un arbre de classification sur l'ensemble des noms. Le résultat obtenu a défini un prétraitement significatif. C'est l'ossature de cet arbre qui a fourni l'argument de la nouvelle algorithmique de remise en cause partielle des associations, en tenant compte d'indices spécifiques de discrimination de classes de paragraphes par une classe de noms. Ces résultats ont été acceptés pour publication lors de la conférence JADT2002.

6.3.3 Caderige : Catégorisation automatique de documents pour l'extraction de réseaux d'interactions géniques

Participants : Michel Le Borgne, Jacques Nicolas, Pascale Sébillot.

Caderige est une action inter-EPST Bio-informatique CNRS, INSERM, INRA, INRIA, Ministère de la Recherche qui regroupe, outre ceux de l'équipe Aïda, des membres des laboratoires Leibniz de l'Imag, du LIPN, du LRI, et de deux laboratoires Inra : MIG et Inra-Ensar. Son objectif global est d'extraire automatiquement à partir de textes de bioinformatique provenant de bases telles Medline les zones de textes décrivant des interactions entre gènes et de modéliser l'interaction décrite. La modélisation nécessitant une analyse fine et coûteuse des phrases, elle ne doit être effectuée que sur des zones de textes susceptibles de contenir effectivement une interaction.

Cette année, nous avons débuté un travail d'intégration de nos résultats en PLI et classification à la détermination de ces zones de textes pertinentes. Nous avons en particulier amélioré l'étiquetage catégoriel et nettoyé le corpus de 2209 résumés extraits de Medline qui sert de base

aux traitements, et avons inséré des étiquettes sémantiques permettant de repérer les noms de gènes simples ou complexes. Nous avons également préparé des exemples positifs et négatifs qui seront fournis à Aleph pour tenter d'apprendre ce qui distingue les phrase contenant des interactions des autres. Par ailleurs, nous avons appliqué notre méthode de détection des thèmes sur ces différents paragraphes afin de déterminer la liste des mots porteurs du thème "interaction", mais ces premiers résultats ont encore besoin d'être affinés.

6.4 EIAO (Assistants intelligents pour l'enseignement)

Participants : Jacques Nicolas, Dominique Py.

6.4.1 Interaction dans les EIAO de calcul formel

Participants : Jacques Nicolas, Dominique Py.

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 sommes intéressés 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. Ce projet, débuté en 1997, s'est achevé en 2001.

Partant des limitations des systèmes de calcul formel (SCF) existants, nous explorons une autre voie qui consiste à réaliser des environnements d'apprentissage en mathématiques intégrant le calcul formel. Il s'agit de spécifier des fonctionnalités autour d'un noyau de calcul formel, afin de développer un environnement dans lequel l'élève puisse organiser en techniques les gestes disponibles et élaborer une réflexion sur ces techniques. Pour cela, nous mettons à la disposition de l'élève des outils voisins de ceux qu'il peut rencontrer dans les SCF existants, ainsi que des moyens de preuve et de guidage.

Deux environnements d'apprentissage ont été implémentés suivant ces principes. L'un concerne l'étude des variations et l'autre les limites de fonctions numériques. Ils ont donné lieu à des expérimentations qui ont permis de vérifier leur utilisabilité et l'intérêt du travail mathématique réalisé par les élèves. Il s'avère que ces environnements conservent certaines caractéristiques des SCF (immédiateté des gestes, nombreux observables, possibilité d'exploration au hasard), mais imposent une organisation de l'étude plus rigoureuse que dans la pratique habituelle.

Si des différences sont apparues en cours de réalisation, dues à la nature de chaque tâche et aux spécificités de l'analyse a priori, il semble néanmoins possible de dégager de cette expérience une méthodologie de conception d'environnements d'apprentissage, généralisable à d'autres domaines des mathématiques utilisant le calcul formel. En particulier, l'importance accordée à l'analyse de la tâche de l'élève et au repérage de gestes à partir d'observations en papier/crayon et avec le calcul formel apparaît déterminante.

6.5 Raisonnements et logiques non classiques

Participants : Philippe Besnard, Yves Moinard.

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.

6.5.1 Inférence préférentielle, circonscription et langages de description d'actions

Participants : Philippe Besnard, Yves Moinard.

Nous avons poursuivi l'étude de la notion générale d'inférence préférentielle dans deux directions, une seule ayant abouti pour l'instant à des résultats concrets. Ayant déjà établi que la version la plus générale, souvent considérée comme trop complexe, est en fait loin d'être si complexe qu'il n'y paraît, nous avons entrepris de réduire son calcul au calcul de circonscriptions. Nous avions précédemment étendu des travaux récents dans ce domaine, qui montraient comment une extension du vocabulaire permettait de traduire bien plus de variétés d'inférence préférentielle qu'il n'avait été affirmé dans la littérature.

Afin de nous faire comprendre, nous nous permettons de rappeler ici les principales définitions. L'inférence préférentielle consiste à se donner une représentation des données, lesquelles sont des formules logiques classiques. La représentation associe un ensemble d'"états" à chaque ensemble de données. Ces "états" peuvent être les modèles en terme de sémantique classique, ou des copies de modèles, ou des ensembles de modèles, voire même des copies d'ensembles de modèles (nous avons déjà montré que cette dernière généralisation n'apportait rien en fait). La relation binaire décrit quels états doivent être préférés : on ne garde que les états associés aux données du domaine qui sont minimaux pour la relation. L'inférence préférentielle sur laquelle on connaît le plus de résultats, et surtout celle que l'on sait le mieux calculer, la circonscription, est un cas très particulier du type le plus simple (où état = modèle classique). Il était considéré comme impossible de traduire une classe significative de l'inférence où les états sont des "copies de modèles" (ce qui apporte donc plus de possibilité pour la relation binaire) en termes de circonscriptions. Il a été démontré récemment qu'il est facile de lever cette impossibilité en étendant le vocabulaire. Nous avions étendu ce résultat en réduisant la taille du vocabulaire étendu, et aussi en démontrant que la classe d'inférences préférentielles considérée décrivait exactement ce que la méthode de traduction peut faire. Ce résultat dérivait facilement de nos précédentes études sur le comportement en termes de règles de raisonnement des différentes inférence préférentielles. Cela démontrait en particulier l'impossibilité de traduire une classe significative (non dégénérée en la classe inférieure) des inférences préférentielles les plus générales par ce type de méthode. Or, ce type de méthode peut être très intéressant puisque la circonscription est, dans ce domaine, ce que l'on sait le mieux calculer.

Nous sommes parvenus [28] à contourner cette impossibilité en modifiant le vocabulaire. Il ne s'agit donc plus d'une extension, même si le vocabulaire final est toujours plus grand que celui de départ. Le vocabulaire final est très grand, et la méthode de traduction est un peu plus compliquée. Toutefois, nous avons aussi montré qu'on peut en fait se contenter de considérer une sous-famille très simple de formules logiques dans ce grand vocabulaire, famille comportant exactement autant de formules que le vocabulaire de départ. Cela semble donc rendre la méthode réaliste, même s'il reste à étudier sa faisabilité effective, et en particulier sa complexité. Nous nous sommes attachés en tout cas à simplifier la traduction, qui est donnée sous une forme assez simple et très générale, pouvant permettre d'envisager d'autres applications. Et nous avons aussi utilisé nos résultats précédents sur les ensembles minimalement équivalents de formules à circonscrire, afin d'obtenir au final une circonscription qui soit le plus facile possible à calculer.

La seconde direction consiste à utiliser nos connaissances actuelles sur les règles de raisonnement des diverses inférences préférentielles, et en particulier de la circonscription, afin d'utiliser au mieux ces inférences dans un langage de description des actions comme le calcul situationnel ou le calcul des événements. Il s'agit d'un retour aux sources, puisque la circonscription a été introduite dès le départ afin de faciliter la traduction de "règles générales" comme l'"inertie" (au sens où, sauf indication directe ou déductible à partir de données, une propriété ne se modifie pas) dans des langages de ce type. Le problème s'est avéré plus complexe qu'envisagé au départ (fin des années 70). Il existe maintenant des propositions d'utilisation effective de la circonscription dans ces langages, et aussi des méthodes de calcul effectives de la circonscription ou de mécanismes proches. La description exacte des règles de raisonnements de ces formalismes doit permettre de mieux cerner la pertinence des propositions actuelles, et éventuellement de les améliorer.

7 Contrats industriels (nationaux, européens et internationaux)

7.1 Pronostic pour la maintenance conditionnelle

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

Il s'agit du contrat EDF 101C0002 d'une durée de 6 mois entre le service recherche d'EDF Chatou et le projet Aïda. Il a pour objectif l'étude de l'utilisation des techniques de diagnostic temporel pour le pronostic en vue d'améliorer la qualité et le coût des opérations de maintenance de centrale nucléaire. Nous avons proposé l'utilisation de chroniques temporelles pour la détection précoce de défaillance. Cette première étude s'est achevée en constatant la nécessité de compléter les connaissances en adjoignant des modèles de dégradation de composants ou d'évoluation de pannes au modèle causal de pannes actuellement disponible.

7.2 Modélisation, diagnostic et supervision de réseaux de télécommunication

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

La convention CTI avec le CNET concernant la surveillance de réseaux de télécommunications s'est terminée en 2000. Le projet RNRT, démarré en 1999, se poursuit. Il s'effectue en collaboration avec Alcatel CIT, le Cnet et Ilog côté industriels, et côté universitaires avec le LIPN/Université Paris-Nord et au sein de l'IRISA avec 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 identification des défaillances ou des pannes. Après modélisation du réseau, il s'agit de reconnaître en ligne des situations à risques en utilisant 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 6.1.1) a été étudiée dans ce contexte et une approche de type diagnostiqueurs décentralisés a été développé pour traiter cette application(voir section 6.1.2 et [29]).

7.3 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 (prolongée jusqu'en mai 2002), 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. Nous avons particulièrement travaillé cette année sur l'apprentissage et la reconnaissance d'électrocardiogrammes multivoies.

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

Participants : 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.

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

Participants : Vincent Claveau, Pascale Sébillot.

Il s'agit d'un contrat de 2 ans terminé fin mai 2001 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 avait 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.

7.6 Caderige : catégorisation automatique de documents pour l'extraction de réseaux d'interactions géniques

Participants : Michel Le Borgne, Jacques Nicolas, Pascale Sébillot.

Il s'agit d'un pré-contrat de 1 an obtenu en 2000, qui a été renouvelé pour 2 ans en octobre 2001. Cette action inter-EPST Bio-informatique CNRS, INSERM, INRA, INRIA, Ministère de la Recherche regroupe, outre ceux de l'équipe Aïda, des membres des laboratoires Leibniz de l'Imag, du LIPN, du LRI, et de deux laboratoires Inra : MIG et Inra-Ensar. Son objectif est de filtrer, dans des bases textuelles de bioinformatique telles que MedLine, les textes parlant spécifiquement d'interactions géniques, et d'extraire de ces textes des réseaux de telles interactions. La participation de notre équipe concerne d'une part la détection des zones de textes susceptibles de contenir une interaction et, d'autre part, la modélisation de l'interaction repérée.

7.7 Analyse des caractéristiques d'un auditoire en vue de la conception d'un logiciel d'argumentation

Participants : Philippe Besnard, Pascale Sébillot.

Il s'agit d'un contrat avec Thomson-CSF Communications Gennevilliers de 9 mois obtenu fin 2001. Un énoncé qui établit une conclusion sur la base de diverses hypothèses est un argument quand le destinataire du propos tend à être favorablement disposé vis-à-vis des hypothèses. Cet aspect de l'argumentation est qualifié d'accord avec l'auditoire. L'objectif de ce contrat est d'analyser le processus d'accord avec l'auditoire selon quatre points : l'interaction entre les prémisses de l'argumentation et la génération d'arguments, les relations entre les prémisses et les caractéristiques de l'auditoire, l'attribution et la validation d'une caractéristique pour un auditoire donné, et l'analyse des auditoires hétérogènes.

8 Actions régionales, nationales et internationales

8.1 Actions régionales

Nous avons cette année finalisé et concrétisé notre action en faveur d'une réorientation d'une partie du projet vers la bio-informatique. À cette occasion nous avons établi 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).

8.2 Actions nationales

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

8.3 Actions financées par la Commission Européenne

En novembre 2001, débute un contrat européen StressGenes (No Q5RS-2001-02211, Qality of Life and management of living Resources Area 5.1.2) consacré à l'étude de gènes impliqués dans la résistance au stress chez les poissons. Les partenaires incluent l'Inra (Scribe) à Rennes, porteur du projet, et les universités d'Aberdeen, de Galway, de Liverpool et d'Uppsala. Les études concerneront un organisme modèle pour la pisciculture, la truite arc-en-ciel. L'approche poursuivie est celle de la génomique fonctionnelle, via la fabrication de micro-arrays dédiés. Dans une première étape, il s'agira d'extraire les gènes candidats depuis les bases de données publiques. Nous comptons exploiter les résultats du projet Caderige d'analyse des résumés scientifiques pour cela. Vient ensuite l'activité de conception et de développement du système d'information qui sera créé pour gérer les données du projet. L'objectif est de se rapprocher d'un système de type LIMS (Laboratory Information Management System), en s'appuyant sur les outils et normes développés dans la communauté internationale. La deuxième phase du projet concernera l'exploitation des données d'expression produites à partir des micro-arrays. Il s'agit alors pour nous d'effectuer un transfert des techniques de classification et de découverte de motifs étudiées dans l'équipe au cas des données d'expression du génome de la truite.

8.4 Réseaux et groupes de travail internationaux

- Participation au réseau d'excellence européen MONET (Model-Based and Qualitative Reasoning). La suite de ce réseau MONET2 est en cours de création (M.-O. Cordier),
- Participation au groupe de travail européen BRIDGE dont l'objectif est d'étudier les liens entre les approches automatique et intelligence artificielle au problème du diagnostic.

Organisation d'une journée de travail dans le cadre du workshop international DX'01 (M.-O. Cordier).

8.5 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)

8.6 Accueils de chercheurs étrangers

- Visite de Torsten Schaub et André Floeter (Université de Potsdam) pendant une semaine.
- Visite de Robin Gras et David Hernandez (Université de Genève) pendant une semaine.

9 Diffusion de résultats

9.1 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 membre du comité de rédaction de AAI (*journal of Applied Artificial Intelligence*) et du journal européen AICOMS ; membre du comité de programme de DX'O1 ainsi que de celui de RFIA'02, ECAI'02 et KR'02.
- J. Nicolas a été membre du comité de programme de la conférence JOBIM'2001 (Toulouse, mai 2001).
- F. Coste a été membre du comité de programme de la conférence CAP2001 (Grenoble, mai 2001).
- 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 a été membre du comité de programme des journées EGC'2001, "Extraction et Gestion des Connaissances", 18-19 janvier 2001, Nantes.
- I.-C. Lerman est membre du comité de programme des 8-èmes Rencontres de la Société Francophone de Classification, SFC01, Guadeloupe, 17-21 décembre 2001
- I.-C. Lerman est membre du comité de programme des Journées Nationales EGC 2002, Extraction et Gestion de Connaissances, Montpellier, 21-23 janvier 2002
- P. Sébillot est membre du comité de lecture de la revue In Cognito.
- P. Sébillot a été membre du comité de lecture des numéros spéciaux de la revue TAL "Lexiques sémantiques dans les applications du TAL" et "Traitement automatique des langues et linguistique de corpus".
- P. Sébillot a été membre du comité de programme de GL'2001.
- P. Sébillot a été membre du comité de programme de LACL'2001.

9.2 Enseignement universitaire

- Module du tronc commun du DEA d'informatique : *module RATS : raisonnement temporel et spatial* (M.-O. Cordier, Y. Moinard, R. Quiniou).
- Option du DEA d'informatique : *module DIAG : diagnostic* (M.-O. Cordier, Y. Pencolé, S. Robin).
- Option du DEA d'informatique : *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 DIIC3 IFSIC : *Introduction à l'intelligence artificielle* (M.O. Cordier).
- Cours en DESS Méthodes Informatiques et Technologies de l'Information et de la Communication (Mitic) IFSIC : *indexation multimédia* (P. Sébillot)
- DESS Méthodes Informatiques et Technologie de l'Information et de la Communication (Mitic) : *méthodes de Data Mining ou fouille des données* (B. Tallur)
- DESS Mathématiques Appliquées : *Cours sur l'analyse des données par la classification* (B. Tallur).
- 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)
- Cours sur *Les arbres de décision par la méthode CART*, Journées Pédagogiques de Vannes, 13-15 Juin 2001 (B. Tallur)
- Cours *Introduction à la Classification Hiérarchique Non-Supervisée*, Journées Pédagogiques de Vannes, 13-15 Juin 2001 (I.C. Lerman).
- Option Bioinformatique en 5ème année INSA (L. Berti, J. Nicolas)

9.3 Participation à des colloques, séminaires, invitations

- Exposé de R. Bonato "Introduction to Categorial Grammars Learning". Procope Workshop "Constraints, Automata, and Applications to Natural Language", Université de Lille 3, France, 1-2 décembre 2001.
- Conférence invitée d'I.-C. Lerman et Philippe Peter aux Journées Portugaises de Classification "Indice Probabiliste de la Vraisemblance du Lien entre Objets Quelconques : Analyse Comparative entre deux approches et proposition du logiciel SIMOB", Porto, 8-10 Février 2001.
- I.-C. Lerman et K. Bachar "Agrégations Multiples et Contraintes de Contiguïté dans la Classification Ascendante Hiérarchique utilisant les Voisins Réciproques et le Critère de la Vraisemblance des Liens", 8-èmes Rencontres de la SFC, Guadeloupe, 17-21 décembre 2001, à pararaître dans les actes.
- Exposé de R. Quiniou "Learning temporal rules from the ECG" au workshop "Machine Learning with Spatial and Temporal Data" dans le cadre d'ICML-2001, Williamstown (USA), 29 juin 2001.
- Exposé de M.O. Cordier dans le cadre du workshop "Spatio-Temporal Reasoning and Geographic Information Systems", ECSQARU'01, Toulouse,septembre 2001.
- Séminaire de P. Sébillot, au Loria de Nancy, lors des journées Inria sur le Web sémantique

sur le thème "Acquérir automatiquement les éléments de sémantique lexicale pertinents pour la reformulation d'index", janvier 2001
- Séminaire de V. Claveau sur le thème "Indexation de documents textuels" lors de la journée "Indexation multimédia" du CNRT-Irisatech à l'Irisa, mai 2001
- Participation de P. Sébillot à l'atelier E-connaissance lors des journées du Grand Ouest organisées à La Rochelle par l'association Ouest-Atlantique, mai 2001
- Séminaire de V. Claveau à la journée "Ontologies et Web sémantique" du groupe de travail "Outils et Méthodes pour la Mémoire d'Entreprise" des clubs CRIN de l'association ECRINS organisée à Sophia-Antipolis, novembre 2001
- Séminaire de P. Sébillot, à Rocquencourt, lors de la journée Inria-Industrie sur le document, décembre 2001
- Exposé invité de M.O. Cordier aus journées scientifiques de l'ONERA sur "Les multiples utilisations des techniques de satisfaction et d'optimisation", novembre 2001.
- Exposé invité de J. Nicolas au séminaire Langage du Génome du LRI Orsay sr le thème 'Inférence grammaticale pour la découverte de signatures de familles de séquences.', 3 mai 2001.
- Séminaire de J. Nicolas au Max Planck Institüt de Golm, le 6 décembre 2001 "Grammatical inference and pattern discovery in biological sequences".
- Invitation de J. Nicolas 2 semaines à l'IEHS (Prof. M. Gromov) à Gif sur Yvette.
- Invitation de J. Nicolas 1 semaine au Max Planck Institüt de Golm et à l'université de Potsdam (Prof. T. Schaub, banlieue de Berlin).
- Exposé invité de Jacques Nicolas à l'Académie des technologies à Paris sur le thème 'Analyse de protéines membranaires : 1er bilan d'une expérience multidisciplinaire Biologie-Informatique', le 10 juillet 2001.
- Exposé de B. Tallur aux Journées Statistiques de l'INRIA Rennes IRISA sur "Classification et prédiction fonctionnelle des protéines à partir des séquences", Novembre 2001.

10 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 », *TAL (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, p. 367–393.

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

[8] 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`.

[9] 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/`.

[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] I. Grosclaude, *Diagnostic abductif temporel : scénarios de pannes, modèles causaux et traitement de l'interaction*, thèse de doctorat, université de Rennes 1, jun 2001.

[12] D. Py, *Environnements Interactifs d'Apprentissage et démonstration en géométrie*, thèse de doctorat, Habilitation à Diriger des Recherches, Université de Rennes 1, jul 2001.

Articles et chapitres de livre

[13] V. Claveau, P. Sébillot, P. Bouillon, C. Fabre, « Acquérir des éléments du lexique génératif : quels résultats et à quels coûts ? », *TAL (traitement automatique des langues), numéro spécial Lexiques sémantiques dans les applications du TAL, à paraître 42*, 3, 2001.

[14] B. Daille, C. Fabre, P. Sébillot, « Applications of Computational Morphology », *in : Many Morphologies, to appear*, Cascadilla Press, 2001.

[15] A. Elamrani, L. Marie, A. Aïnouche, J. Nicolas, I. Couée, « Genome wide distribution and potential regulatory functions of AtATE, a novel miniature inverted-repeat transposable element that is present in the promoter region of one of the Arginine Decarboxylase genes in Arabidopsis thaliana », *Soumis à Molecular Genetics and Genomics*, 2001.

[16] C. Largouët, M.-O. Cordier, « Improving the Landcover Classification using Domain Knowledge », *AI Communication special issue on Environmental Sciences and Artificial Intelligence 14*, 1, 2001, p. 35–43.

Communications à des congrès, colloques, etc.

[17] L. Berti-Equille, D. Graveleau, « Documents, données et méta-données : une approche mixte pour un système de veille », *in : Colloque Veille Stratégique, Scientifique et Technologique (VSST'01)*, *I*, p. 115–126, Barcelone, octobre 2001.

[18] L. Berti-Equille, « Integration of Biological Data and Quality-Driven Source Negotiation », *in : 20th. Intl. Conference on Conceptual Modeling (ER2001)*, *LNCS*, *2224*, p. 256–269, Yokohama, Japan, november 2001.

[19] R. Bonato, C. Retoré, « Learning Rigid Lambek Grammars and Minimalist Grammars from Structured Sentences », *in : Third Learning Language in Logic Workshop (LLL2001)*, L. Popelínský, M. Nepil (éditeurs), sept 2001.

[20] P. Bouillon, V. Claveau, C. Fabre, P. Sébillot, « Using Part-of-Speech and Semantic Tagging for the Corpus-Based Learning of Qualia Structure Elements », *in : First international workshop on Generative Approaches to the Lexicon (GL'2001)*, P. Bouillon, K. Kanzaki (éditeurs), apr 2001, `http://www.irisa.fr/aida/aida-new/dataFiles/vclaveau/GL2001.ps`.

[21] G. Carrault, F. Wang, R. Quiniou, M.-O. Cordier, « Apprentissage de séquences structurées : exemple en ECG », *in : 18e colloque GRETSI'01 sur le traitement du signal et l'image*, Toulouse, France, septembre 2001, `http://www.cta-congres.com/gretsi/INTRO.html`.

[22] M.-O. Cordier, C. Largouët, « Using model-checking techniques for diagnosing discrete-event systems », *in : Proceedings of the Twelve International Workshop on Principles of diagnosis (DX'01)*, p. 39–46, mar 2001, `http://www.irisa.fr/aida/aida-new/dataFiles/clargoue/dx_final.ps`.

[23] F. Coste, D. Fredouille, « Inférence d'AFNs : restriction de l'espace de recherche aux automates non ambigus », *in : Conférence d'apprentissage, CAp 2001*, Presses Universitaires de Grenoble, p. 75–84, juin 2001.

[24] I. Grosclaude, M.-O. Cordier, R. Quiniou, « Causal interaction : from a high-level representation to an operational event based representation », *in : IJCAI'01*, aug 2001, `http://www.irisa.fr/aida/aida-new/dataFiles/quiniou/ijcai01.pdf`.

[25] E. Guérin, F. Moussouni, L. Berti-Equille, « Intégration des données sur le transcriptome », *in : Actes de la Journée du GDR-PRC I3*, p. 219–228, Lyon, Décembre 2001.

[26] C. Largouët, M.-O. Cordier, « Adding Probabilities to timed automata to improve landcover classification », *in : ECSQARU-2001 Workshop*, IRIT - UMR 5505 CNRS - INP - UPS, 2001, `http://www.irisa.fr/aida/aida-new/dataFiles/clargoue/escaru.ps`.

[27] D. Lenne, J. Gélis, J. Lagrange, D. Py, « Modélisation de l'interaction dans les EIAO utilisant le calcul formel », *in : Actes des Journées EIAO*, *8*, 1-2, Hermès, 2001.

[28] Y. Moinard, « General preferential entailments as circumscriptions », *in : Symbolic and Quantitative Approaches to Reasoning and Uncertainty (ECSQARU 2001)*, S. Benferhat, P. Besnard (éditeurs), *Lecture Notes on Artificial Intelligence*, *2143*, Springer-Verlag, LNAI, p. 532–543, September 2001.

[29] Y. Pencolé, M.-O. Cordier, L. Rozé, « Incremental decentralized diagnosis approach for the supervision of a telecommunication network », *in : working notes of the 12th International Workshop on Principles of Diagnosis DX'01*, mar 2001, `http://www.irisa.fr/aida/aida-new/dataFiles/ypencole/DX01.pdf`.

[30] R. Quiniou, M.-O. Cordier, G. Carrault, F. Wang, « Application of ILP to cardiac arrhythmia characterization for chronicle recognition », *in : ILP'2001*, C. Rouveirol, M. Sebag (éditeurs), *LNCS*, Springer-Verlag, sept. 2001, `http://www.irisa.fr/aida/aida-new/dataFiles/quiniou/ilp01.pdf`.

[31] B. Tallur, « Méthodes d'apprentissage supervisé et non spervisé pour la prédiction des fonctions à partir des séquences protéiques », *in : SFC'01 (9èmes Rencontres de la Société Francophone de Classification)*, December 2001, `http://www.irisa.fr/aida/aida-new/dataFiles/tallur/sfc2001.ps`.

[32] F. Wang, R. Quiniou, G. Carrault, M.-O. Cordier, « Learning structural knowledge from the ECG », *in : ISMDA-2001*, *LNCS*, *1933*, Springer Verlag, oct. 2001, `http://www.irisa.fr/aida/aida-new/dataFiles/quiniou/ismda01.pdf`.

[33] Y.Pencolé, M-O.Cordier, L. Rozé, « A decentralized model-based diagnostic tool for complex systems », *in : thirteen IEEE international conference on tools with artificial intelligence*, IEEE computer society, p. 95–102, Novembre 2001.

Divers

[34] M.-O. Cordier, I. Grosclaude, R. Quiniou, S. Robin, « Étude du pronostic pour la maintenance conditionnelle », Rapport de fin de contrat EDF 101C0002, Décembre 2001.