INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team MOSTRARE

Modeling Tree Structures, Machine Learning, and Information Extraction

Futurs

THEME SYM

Activity Report

2004

Table of contents

1. Team

MOSTRARE *is a joint project with the* LIFL *(UMR 8022 of CNRS and University of Lille 1) and the* GRAPPA *Group (EA 3588 of the University of Lille 3).*

Head of project
Rémi Gilleron [professor, University of Lille 3]

Administrative assistant
Karine Lewandowski [shared with 2 other projects]

Staff member INRIA
Joachim Niehren [senior researcher (DR2), UR Futurs]

Staff member Lille 3 University
Aurélien Lemay [assistant professor]
Isabelle Tellier [assistant professor]
Marc Tommasi [assistant professor]
Fabien Torre [assistant professor]

Staff member Lille 1 University
Anne-Cécile Caron [assistant professor]
Yves Roos [assistant professor]
Jean-Marc Talbot [assistant professor]
Sophie Tison [professor]

Ph. D. student
Iovka Boneva [MESR fellowship, since October 2002]
Julien Carme [MESR fellowship, since October 2002]
Denis Debarbieux [MESR fellowship, since October 2002]
Patrick Marty [INRIA and Région Nord-Pas-de-Calais fellowship, since October 2003]
Florent Jousse [INRIA and Région Nord-Pas-de-Calais fellowship, since October 2004]
Laurent Planque [MESR fellowship, since October 2004]

2. Overall Objectives

MOSTRARE was successfully evaluated during 2003 so that it became an INRIA project in April 2004. MOSTRARE is bi-located, at the group STC at the LIFL of the University of Lille 1 (UMR 8022 of CNRS) and at the group GRAPPA of the University of Lille 3 (EA 3588).

The objective of MOSTRARE is to develop adaptive information extraction systems for semi-structured documents, that can fully exploit available tree structure. We approach this goal in two research lines:

Modeling Tree Structure for Information Extraction: define and investigate models of tree structures as needed by information extraction; develop corresponding algorithms and software components.

Machine Learning for Information Extraction: develop learning algorithms that induce models of tree structures and apply them to information extraction. Combine learning algorithms for tree and string models so that they apply to diverse data formats, and possibly to heterogeneous data.

3. Scientific Foundations

3.1. Modeling Tree Structure

Keywords: *XML*, *queries in trees*, *semi-structured documents*, *tree automata and logic*, *tree wrapper*.

The evolution of XML into the major document exchange format has reawaken a strong interest in modeling tree structures. The main objectives are to query for nodes in trees (XPath), to validity of documents, i.e., membership to a set of valid trees (DTD), and to transform trees into others (XSLT). Modelling approaches rely on tree automata, monadic second-order logics, modal logics, pattern languages, attribute grammars, or tree transducers.

The main goal of the project is to design adaptive information extraction systems that fully exploit the tree structures of XML or HTML documents. In our approach we want to combine novel models of tree structures and machine learning techniques. Appropriate languages for modeling node queries in tree satisfy a number of properties required for adaptive information extraction. Possible trade-offs between expressiveness, learnability, and efficiency are to be understood. We thus propose new query languages in trees and investigate them from these perspectives.

We consider tree automata for ranked and unranked trees, that may be ordered or unordered. We design efficient algorithms for querying semi-structured data by automata. We investigate logical query languages such as monadic Datalog, monadic second-order logic, and modal logical languages.

3.2. Learning Queries in Trees, Wrapper Induction

Keywords: *grammatical inference*, *semi-structured documents*, *statistical learning*, *wrapper induction*.

We suppose that a collection of tree-structured documents is given as input together with annotated positions of interest. The task is to learn a node queries in trees – tree wrapper – from this collection that can identify relevant positions in unseen documents. We search for new algorithms that learn models of languages of semi-structured data. We extend results on grammatical inference for regular (tree) languages to the case of unranked and/or unordered trees.

Because of the presence of noise in real datasets, we search for extensions from the string case to the tree case of statistical wrapper induction algorithms. Also, we study combination of wrapper induction algorithms because data often stem from heterogeneous sources.

4. Application Domains

Keywords: *Business Intelligence*, *Information Retrieval*, *Knowledge Management*, *Multimedia*, *Web Intelligence*.

The main objective is to develop wrappers for data intensive Web servers, CGI-based Web servers and Web services, that is sets of semi-structured documents whose structure is quite uniform. Wrappers are used in information mediator services, in information retrieval tools and in text mining tools. No specific application domain is targeted so far but wrappers are useful in Business Intelligence and Knowledge Management.

5. Software

5.1. Squirrel : Tree Automata for Query Induction

Keywords: *Mozilla*, *grammatical inference*, *query*, *wrapper induction*.

Participants: Julien Carme [correspondent], Aurélien Lemay, Joachim Niehren.

SQUIRREL is a tool for inducing monadic queries in HTML trees that are represented by tree automata. The inference algorithm of Squirrel is based on methods from grammatical inference. Wrapping amounts to answering queries in trees. Squirrel relies on a package for tree automata that we implemented in 2003.

Squirrel comes with a visual interactive interface for annotating documents, inferring and testing wrapper. We have implemented this interface as an extension for the Web navigator Mozilla Firefox. For annotation, the user clicks on elements of a Web page that he wants to extract. Squirrel infers a query that is compatible with the given examples and then indicates which elements the inferred wrapper would extract on the whole document. The user can then correct this extraction in an interactive process until the inferred wrapper performs correctly. The user can also test this query by applying it to unseen pages.

The early prototype of Squirrel seems highly promising; improvements of the algorithms, adequate heuristics, and extension to n-ary queries are under development.

5.2. CafeIn : Statistical Classification for Information Extraction

Keywords: *HTML documents*, *supervised classification*, *texts*, *wrapper induction*.

Participants: Patrick Marty [correspondent], Rémi Gilleron, Fabien Torre.

CafeIn generalized the Boosted Text Wrapper Induction prototype from December 2003 to a platform for statistical wrapper generation. It is parameterized by a document representation model and a supervised classification algorithm that can operate on that model. Currently CafeIn includes a number of feature-sets for textual and structured documents, that can be easily customized or extended. It is planned to integrate possibilities such as boosting and co-learning.

CafeIn is currently participating in the Pascal Challenge on machine learning for information extraction. This challenge will end on Friday 17th December 2004.

6. New Results

6.1. Modeling Tree Structures

6.1.1. Queries in Ordered Trees

Keywords: *monadic second-order logic*, *monadic Datalog*, *n-ary queries in unranked trees*, *tree automata*.

Participants: Julien Carme, Joachim Niehren [correspondent], Laurent Planque, Jean-Marc Talbot, Marc Tommasi, Sophie Tison, Alain Terlutte [collaborator].

Information extraction from semi-structured documents requires to find n-ary node queries in trees that define appropriate sets of n-tuples of nodes. Monadic queries for sets of nodes in trees recently have received considerable interest in the area of databases.

In [28], we propose a new representation formalisms for n-ary queries by tree automata that we prove to capture MSO. We then investigate n-ary queries by unambiguous tree automata which are relevant for query induction in multi-slot information extraction. We show that this representation formalism captures the class of n-ary queries that are finite unions of Cartesian closed queries, a property we have proved decidable. Future works include a study of learnability of n-ary queries.

In [17], we present *node selecting tree transducer* (NSTT) for representing monadic queries. NSTTs are the query formalism underlying our Squirrel tool. We have shown that deterministic NSTTs can be identified with monadic queries by unambiguous tree automata.

In [18], we propose *stepwise tree automata* for querying unranked trees equally to binary trees. Stepwise tree automata are traditional tree automata. Stepwise tree automata simplify the often-needed query transfer ranked to unranked trees. This comes with their algebra nature.

We have investigated related logics and constraints in ordered trees that are relevant to subtyping in programming languages [12][24] and for underspecified semantics of natural language [14].

6.1.2. Queries in Unordered Trees

Keywords: *modal logic*, *queries*, *semi-structured data*, *unordered unranked trees*.

Participants: Iovka Boneva, Jean-Marc Talbot [correspondent], Sophie Tison.

It is often useful to ignore the ordering of elements in semi-structured documents, so that these become unordered trees. The canonical logical language for querying unordered trees is monadic second-order logic (MSO).

In [15], we investigate an alternative modal logical querying formalism, the spatial logic TQL proposed by Cardelli and Ghelli in the context of the ambient calculus. We exhaustively study combined and data complexity of the model checking problem of TQL and its fragments, and equally investigate satisfiability. We characterize fragments of TQL that have the same expressive power as MSO.

In [27][26], we link TQL to languages with counting constraints such as Presburger monadic second-order logic (PMSO) introduced by Muscholl, Seidl and Schwentick or counting monadic second-order logic (CMSO) introduced by Courcelle in a uniform tree automata framework (see also [11]).

6.1.3. Queries in Graphs

Keywords: *path constraints*, *rewriting*, *semi-structured documents*.

Participants: Anne-Cécile Caron, Denis Debarbieux, Yves Roos, Sophie Tison [correspondent], Yves André [collaborator].

Semi-structured documents with hyper-links or other references are best modeled by rooted edge-labeled digraphs. We study inclusion constraints for such graphs, that were introduced by Abiteboul and Vianu (1997) in the context of query optimization. A inclusion constraint $p \preceq q$ with regular path expressions p and q means that the set of nodes reachable by path in p is included in the set of nodes reachable by path in q.

In [13], we give a PSPACE decision algorithm for the implication problem of a constraint $p \preceq q$ by a set of constraints $p_i \preceq u_i$, where p, q, the p_i's are regular path expressions, and u_i's are non empty words. We thereby improve a previous EXPSPACE algorithms of Abiteboul and Vianu (1997), and on a EXPTIME algorithm by de Rijke et. al(2003).

In [20], we study the existence of exact models for a given set of path constraints: an exact model of a set of path constraints C satisfies the inclusion constraint $p \preceq q$ if and only if the inclusion constraint is entailed by C. We propose a decidable characterization of sets C of path inclusions $p \preceq u$ where p is a regular set of paths and u is a singleton path, which have a finite exact model. We present an effective way of computing such a model when it exists. (see also [19]).

6.2. Learning Queries in Trees, Wrapper Induction

6.2.1. Tree Automata for Query Induction

Keywords: *grammatical inference*, *monadic queries*, *ordered trees*, *tree automata*, *wrapper induction*.

Participants: Aurélien Lemay [correspondent], Julien Carme, Rémi Gilleron, Joachim Niehren, Marc Tommasi, Alain Terlutte [collaborator].

Programming node queries in `HTML` trees manually is a difficult, tedious, and time consuming task. Even with visual wrapper induction tools such as Lixto, it still requires expertise in tree logics or other query language and on the many details of `HTML`. We hope to simplify the wrapper generation task by query induction from annotated examples. In contrast to previous query induction approaches, we consider learning queries in trees rather than strings. Queries in trees are represented by tree automata.

In [17] we have proposed *node selecting tree transducer* (NSTT) for modelling monadic queries in ordered trees – as already mentioned above. Node selecting transducer are particular tree automata that operate on trees with Boolean annotations. We have shown how to infer NSTTs from completely annotated examples; these are trees where all nodes are marked Booleans, stating whether the node is selected or not. Our learning algorithm adapts an induction algorithm for tree automata from Oncina and Garcia [32].

The Squirrel prototype has been developed on those ideas. End users annotates nodes in HTML documents through a standard Mozilla compatible Web-browser. Squirrel induces and tests NSTTs from these examples. The prototype already allows to deal with partially annotated examples.

6.2.2. Statistical Classification for Query Induction

Keywords: *attribute-value representation*, *semi-structured data*, *supervised classification*, *textual data*, *wrapper induction*.

Participants: Patrick Marty, Rémi Gilleron [correspondent], Marc Tommasi, Fabien Torre.

When considering heterogeneous HTML and XML documents as inputs of information extraction tasks – that may be either scattered into pieces or hidden in large parts of purely textual data – different tree wrappers need to be combined with textual wrappers. For classification tasks, some machine learning algorithms realize such combinations. We have started to examine how such methods for information extraction. In a first step, we reformulate information extraction as classification tasks. Second we want to apply known techniques and combine them with structural wrapper induction.

The CafeIn framework ([31],[23]) improve previous approaches based on supervised classification for information extraction ([30], [29]). CafeIn is a single-slot wrapper induction framework for text data or semi-structured data. It combines an adaptive attribute-value representation of documents and a supervised classification algorithm. For text data, the classification task consists in deciding whether a tuple of two positions in a text are respectively the beginning and the end of data to extract. The document representation use only textual features. For semi-structured data, like `XML` or `HTML` trees, data to extract are contained in leaves. Thus the classification task consists in deciding whether a leaf is to be extracted or not. In CafeIn, the attribute-value representation of the data is adaptive because it is based on different feature-sets, and it allows the integration of domain knowledge easily. Any supervised classification algorithm can be used in the CafeIn framework.

As mentioned above, we have developed a CafeIn prototype and examined its performance with different learning algorithms for classification (C4.5 [33], GloBoost [25]) and with different models of data representation. CafeIn is competitive with the others wrapper induction systems based on specialized learning algorithms. It will be continued with the combination of textual and structural informations, and with multi-slots wrappers induction.

6.2.3. Answer Extraction

Keywords: *Information Extraction*, *Machine Learning*, *Question Answering*.

Participants: Florent Jousse, Isabelle Tellier, Marc Tommasi [correspondent].

Question Answering (QA) systems are complex programs able to answer a question in natural language. Their source of information is a given corpus or the Web. To achieve their goal, these systems perform various subtasks among which the last one, called answer extraction, is very similar to an Information Extraction task. In most QA systems, the extraction rules used during answer extraction are hand-written patterns or rules. Writing these rules is a long fastidious task and usually results in very language-specific and domain-specific rules. Our objective it to adapt machine learning techniques to produce such rules, especially those defined for the Information Extraction task (see work of Patrick Marty). The specificities of QA systems need to be identified and exploited in this adaptation. This theme constitutes the framework of the PhD thesis of F. Jousse (starting October 04, see his master thesis on the subject).

Our original hypothesis is that extraction rules sometimes need to take into account the syntactic structure of chunks of natural language texts. Here we hope to benefit from our experience in natural language processing. Our recent contributions to this domain consist in using lexical semantic information in syntax learning. In [10], we have proved the learnability of subclasses of categorial grammars from typed examples (i.e. sentences enriched with lexical semantic information) and proposed a learning algorithm adapted to this kind of data. The strategy has been empirically evaluated on a real corpus [22]. We have also proved the learnability of subclasses of pregroup grammars, a new powerful syntax formalism [16].

7. Contracts and Grants with Industry

In 2004, we have intensified cooperations with the Lixto information extraction company in Vienna, a spin-off of G. Gottlob's database and artificial intelligence group at the technical university of Vienna. We have proposed to pursue this cooperation in form of an associated research team. Exchanges of post-doctoral students are envisaged.

We have continued our regular exchanges with B. Chidlovskii form the Xerox Research Center Europe XRCE in Grenoble. We will propose two master project – on PDF to XML conversion and on statistical tree automata – with the goal to intensify the cooperation.

8. Other Grants and Activities

8.1. French Actions

8.1.1. ACI Masse de Données ACIMDD

Participants: Julien Carme, Rémi Gilleron, Aurélien Lemay, Patrick Marty, Joachim Niehren, Alain Terlutte, Isabelle Tellier, Marc Tommasi [correspondent].

We participate in the French cooperation project "ACI masse de données – ACI-MDD – Accès au Contenu Informationnel pour les Masses de Données et Documents" (2003–2006). The aim of the project is the design of algorithmic tools for Information Retrieval, Information Extraction and Text Classification for semi-structured documents.

In particular, the ACI tries to deal with the problem of heterogeneity of data that can arise in large XML or HTML corpus, often due to the fact that data can come from several sources. To handle this, it is usually useful to be able to cluster data into homogeneous subsets. One objective of the ACI-MDD is therefore to propose clustering techniques adapted to semi-structured documents.

In that field, we proposed a new method, related to the emerging problem of subspace clustering, that aims at identifying clusters that may exist in different subspaces of the original space. And we now have started to explore the task of clustering for semi-structured data.

Also, as many unsupervised learning tasks, the evaluation of clustering tools is problematic as it is difficult to get freely available corpuses. As part of a joint work with other members of the ACI, we have started to produce XML datasets with various specificities accompanied by clustering tasks. Although, a comparison of the various approaches studied in this ACI on the common corpus is under preparation.

Our partners are: Patrick GALLINARI (Coordinator - LIP6) and Marie-Christine ROUSSET (LRI and GEMO INRIA project). More information about the project can be found on http://www.grappa.univ-lille3.fr/Acimdd

8.1.2. ACI TraLaLA: Transformation Languages, Logic and Application

Participants: Iovka Boneva, Anne-Cécile Caron [correspondent], Denis Debarbieux, Joachim Niehren, Yves Roos, Jean-Marc Talbot, Sophie Tison.

We are involved in the ACI "TraLaLA", (XML Transformation Languages, Logic and Application). This ACI is motivated by the increasing number of applications that produce, consume or handle large sets of data, or "datamasses". In many cases, these are either raw data or a collection of data from various sources, both of which lack uniform descriptive criteria. Such cases require more flexibility than the classical relational model can provide, and have given rise to the so-called semi-structured data model, of which XML is one of the most prominent examples. Our project intends to study the processing, querying and handling of large datamasses whenever data is available in XML format. We pay particular attention to the programming languages and query languages problems. We aim to cover in a uniform way a wide spectrum of different areas, namely: programming languages (expressiveness, typing, new programming primitives, query underlying logics, logical optimization), data access (streamed data, compression, access to secondary memory storages, persistency engines), implementation (pattern matching compiling, physical optimization, subtyping verification, execution models for streamed data).

Ours parters are: Giuseppe CASTAGNA (coordinator - LIENS) Luc SÉGOUFIN (GEMO INRIA project), Silvano DAL ZILIO (LIF) and Véronique BENZAKEN (LRI). More information about the project can be found on http://www.cduce.org/tralala.html.

8.1.3. Action RIP-WEB

Participants: Rémi Gilleron, Patrick Marty, Isabelle Tellier, Marc Tommasi.

We are member of a French research group on Question Answering “RIP-WEB: Recherche d’Information Précise sur le WEB: http://www.limsi.fr/Individu/monceaux/RIP-Web/rip-web.html” whose leader is Brigitte GRAU, in LIMSI, and whose purposes include to evaluate what machine learning techniques can bring to Question Answering systems. A workshop of TALN’04 has been organised as part of this action.

9. Dissemination

9.1. Scientific Animation

- **Program Committees:**
  S. TISON was member of the "SPECIF Best thesis Award" jury and of the editorial board of RAIRO - Theoretical Informatics and Applications.
  R. GILLERON was PC member of CAP’2004 (French conference on machine learning).
  M. TOMMASI was PC member of CAP’2004.
  J. NIEHREN was PC member of DEPENDENCY’2004, MOZ’2004, and ROMAND’2004.
  J. M. TALBOT was PC member of CSL’2004 (Conference of the European Association for Computer Science Logic).
- **French Scientific Responsibilities**
  S. TISON is, vice-director of the LIFL (computer science department in Lille), head of the research group STC of the LIFL, and director of the doctoral school SPI of the university Lille 1.
  R. GILLERON is head of the research group GRAPPA of the university of Lille 3, member of the scientific committee of Lille 3, and member of the scientific committee of the RTP STIC CNRS “découvrir et résumer” (French national action of the CNRS on machine learning and data mining).

9.2. Teaching and Scientific Diffusion

- master thesis lectures:
  - J. NIEHREN, J. M. TALBOT and S. TISON on Logic and Modelisation;
  - I. TELLIER and M. TOMMASI on Machine Learning for Information Extraction.
- master projects:
  - L. PLANQUE on n-ary Queries in Trees : Representations and Algorithms.
  - F. JOUSSE on Learning answer extraction pattern for Question Answering Systems.
- development: C. DUPRET and A. SAMSENESENA on the implementation of parts of Squirrel (CaML).
- student internships: E. FILIOT (magister Lyon) on libraries for CafeIn (CaML).
- direction of PhD thesis submitted in 2004:
  - D. Dudau on learning categorial grammars to simulate the acquisition of natural language with the help of semantic information (University of Lille I). Directed by R.Gilleron, I. Tellier, and Marc Tommasi.
  - Tim Priesnitz Satisfiability and Entailment of Subtype constraints. University of Saarbruecken. Directed by Joachim Niehren.

- PhD committees:
  R. GILLERON belonged to the committee of A. HABRARD (St. Etienne), L. DENOYER (Paris VI), D. DUDAU (Lille), and Y. ESPOSITO (Marseille). J. NIEHREN belonged to the committee of M. VILLARET (Barcelona/Girona). S. Tison was member of the committees of D. DUDAU (Lille), B. WEINBERG, J. LENOIR, and A. ALJER
- Habilitation committees: S. Tison was member of the habilitation committee of H. TOUZET (Lille).
- Evaluation committees : R. GILLERON is member of the scientific committee for the evaluation of LRI (computer science department of Orsay, Paris 11).

10. Bibliography

Major publications by the team in recent years

[1] S. ABITEBOUL, P. BUNEMAN, D. SUCIU. *Data on the Web*, Morgan Kaufmann Publishers, 2000.

[2] R. BAUMGARTNER, S. FLESCA, G. GOTTLOB. *Visual Web Information Extraction with Lixto*, in "The VLDB Journal", 2001, p. 119-128.

[3] L. CARDELLI, G. GHELLI. *A query language based on the ambient logic*, in "Proceedings of the 9th European Symposium on Programming ESOP'01", Lecture Notes in Computer Science, vol. 2028, 2001, p. 1–22.

[4] H. COMON, M. DAUCHET, R. GILLERON, F. JACQUEMARD, D. LUGIEZ, S. TISON, M. TOMMASI. *Tree Automata Techniques and Applications*, 1997, http://www.grappa.univ-lille3.fr/tata.

[5] G. GOTTLOB, C. KOCH. *Monadic Queries over Tree-Structured Data*, in "Proceedings of the 17th IEEE Symposium on Logic in Computer Science (LICS 2002), Copenhagen", Lecture Notes in Computer Science, 2002, p. 189–202.

[6] R. KOSALA, M. BRUYNOOGHE, J. V. DEN BUSSCHE, H. BLOCKEEL. *Information Extraction from web documents based on local unranked tree automaton inference*, in "Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-2003)", 2003, p. 403–408.

[7] N. KUSHMERICK. *Finite-state approaches to Web information extraction*, in "Proc. 3rd Summer Convention on Information Extraction", 2002.

[8] M. MÜLLER, J. NIEHREN, R. TREINEN. *The First-Order Theory of Ordering Constraints over Feature Trees*, in "Discrete Mathematics and Theoretical Computer Science", vol. 4, n^{o} 2, 2001, p. 193-234.

[9] F. NEVEN, T. SCHWENTICK. *Query automata over finite trees*, in "Theoretical Computer Science", vol. 275, n^{o} 1-2, 2002, p. 633–674.

Doctoral dissertations and Habilitation theses

[10] D. D. SOFRONIE. *Apprentissage de grammaires catégorielles pour simuler l'acquisition du langage naturel à l'aide d'informations sémantiques*, Ph. D. Thesis, Université Lille 1, avril 2004.

Articles in referred journals and book chapters

[11] I. BONEVA, J.-M. TALBOT. *When Ambients Cannot be Opened*, in "Theoretical Computer Science", 2004, http://www.grappa.univ-lille3.fr/twiki/pub/Private/IovkaBoneva/BonevaTalbot-WhenAmbientsCannotBeOpened.pdf.

[12] Z. SU, A. AIKEN, J. NIEHREN, T. PRIESNITZ, R. TREINEN. *First-Order Theory of Subtyping Constraints*, in "ACM Transactions on Programming Languages and Systems", to appear, 2005, http://www.ps.uni-sb.de/Papers/abstracts/sub-journal.html.

Publications in Conferences and Workshops

[13] Y. ANDRÉ, A.-C. CARON, D. DEBARBIEUX, Y. ROOS, S. TISON. *Extraction and Implication of Path Constraints*, in "Proceedings of the 29th Symposium on Mathematical Foundations of Computer Science, Prague (Czech Republic)", Lecture Notes in Computer Science, vol. 3153, Springer Verlag, august 2004, p. 863-875, http://www.lifl.fr/~debarbie/DOC/ExtractionAndImplicationOfPathConstraints.pdf.

[14] M. BODIRSKY, D. DUCHIER, S. MIELE, J. NIEHREN. *A New Algorithm for Normal Dominance Constraints*, in "ACM-SIAM Symposium on Discrete Algorithms", January 2004, p. 54-78, http://www.ps.uni-sb.de/Papers/abstracts/wndc.pdf.

[15] I. BONEVA, J.-M. TALBOT. *On Complexity of Model-Checking for the TQL Logic*, in "3rd IFIP International Conference on Theoretical Computer Science", Kluwer, 2004, http://www.grappa.univ-lille3.fr/twiki/pub/Private/IovkaBoneva/BonevaTalbot-ModelChecking.pdf.

[16] D. BÉCHET, A. FORET, I. TELLIER. *Learnability of Pregroup Grammars*, in "7th International Colloquium on Grammatical Inference", Lecture Notes in Artificial Intelligence, Springer Verlag, 2004, p. 65–76.

[17] J. CARME, A. LEMAY, J. NIEHREN. *Learning Node Selecting Tree Transducer from Completely Annotated Examples*, in "7th International Colloquium on Grammatical Inference", Lecture Notes in Artificial Intelligence, vol. 3264, Springer Verlag, 2004, p. 91–102, http://www.grappa.univ-lille3.fr/~carme/publi/nst.pdf.

[18] J. CARME, J. NIEHREN, M. TOMMASI. *Querying Unranked Trees with Stepwise Tree Automata*, in "International Conference on Rewriting Techniques and Applications", Lecture Notes in Computer Science, vol. 3091, Springer Verlag, 2004, p. 105 – 118, http://www.ps.uni-sb.de/Papers/abstracts/stepwise.html.

[19] D. DEBARBIEUX. *Données semi-structurées et contraintes de chemin*, in "MAJECSTIC'04", 2004.

[20] D. DEBARBIEUX, Y. ROOS, S. TISON. *Models of Path Constraints*, in "10ièmes Journées Montoises d'Informatique Théorique, Liege - Belgique", September 2004, http://www.lifl.fr/~debarbie/DOC/Mons04.pdf.

[21] D. DUDAU-SOFRONIE, I. TELLIER. *A Study of Learnability of Lambek Grammars from Typed Examples*, in "Proceedings of Categorial Grammars 04, Montpellier, France", juin 2004, p. 133-147, http://www.grappa.univ-lille3.fr/~tellier/CG2004.pdf.

[22] D. DUDAU-SOFRONIE, I. TELLIER. *Un modèle d'acquisition de la syntaxe à l'aide d'informations sémantiques*, in "actes de la 11ème Conférence TALN, Traitement Automatique du Langage Naturel, Fès, Maroc",

avril 2004, p. 137–146, http://www.grappa.univ-lille3.fr/~tellier/TALN2004.pdf.

[23] P. Marty, F. Torre. *Codages et connaissances en extraction d'information*, in "Actes de la Sixième Conférence Apprentissage CAp'2004", M. L. et Marc Sebban (editor)., Presses Universitaires de Grenoble, 2004, p. 207–222, http://www.grappa.univ-lille3.fr/~marty/Recherche/Publications/2004/EI-CAp2004.pdf.

[24] J. Niehren, T. Priesnitz, Z. Su. *Complexity of Subtype Satisfiability over Posets*, in "European Symposium on Programming", to appear, April 2005, http://www.ps.uni-sb.de/Papers/paper_info.php?label=pdl05.

[25] F. Torre. *GloBoost : Boosting de moindres généralisés*, in "Actes de la Sixième Conférence Apprentissage CAp'2004", M. L. et Marc Sebban (editor)., Presses Universitaires de Grenoble, 2004, p. 49–64, http://www.grappa.univ-lille3.fr/~torre/Recherche/Articles/2004/GloBoost-CAp2004.pdf.

Miscellaneous

[26] I. Boneva, J.-M. Talbot. *Automata and Logics for Unranked and Unordered Trees*, submitted, 2005.

[27] I. Boneva, J.-M. Talbot, S. Tison. *Expressiveness of Spatial Logic for Trees*, submitted, 2005.

[28] J. Niehren, L. Planque, J.-M. Talbot, S. Tison. *N-ary Queries by Tree Automata*, submitted, 2005, http://www.lifl.fr/~planque/n-ary-queries.pdf.

Bibliography in notes

[29] A. Finn, N. Kushmerick. *Multi-level boundary classification for information extraction*, in "In Proceedings of the European Conference on Machine Learning, Pisa, 2004.", 2004, http://citeseer.ist.psu.edu/634972.html.

[30] D. Freitag, N. Kushmerick. *Boosted Wrapper Induction*, in "Proceedings of the 17th National Conference on Artificial Intelligence AAAI-2000", 2000, p. 577-583.

[31] P. Marty, F. Torre. *Classer pour extraire : représentations et méthodes*, Technical report, n° Grappa report 0103, GRAPPA, december 2003, http://www.grappa.univ-lille3.fr/~torre/Recherche/Articles/2003/marty2003.pdf.

[32] J. Oncina, P. Garcia. *Inferring regular languages in polynomial update time*, in "Pattern Recognition and Image Analysis", 1992, p. 49–61.

[33] J. R. Quinlan. *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA, 1993.