Extraction d'entités dans des collections évolutives

Thierry Despeyroux Eduardo Fraschini
Anne-Marie Vercoustre

Inria, Rocquencourt

EGC 2007, Namur

INRIA

Au menu...

Le rapport d'activité de l'Inria

- Annexe technique
- 168 rapports d'équipes en 2005
- 4611 pages en 2005
- en anglais
- annuel

INRIA

Le rapport d'activité de l'Inria

- Annexe technique
- 168 rapports d'équipes en 2005
- 4611 pages en 2005
- en anglais
- annuel

INRIA

Le rapport d'activité de l'Inria

- Annexe technique
- 168 rapports d'équipes en 2005
- 4611 pages en 2005
- en anglais
- annuel

INRIA

Le rapport d'activité de l'Inria

- Annexe technique
- 168 rapports d'équipes en 2005
- 4611 pages en 2005
- en anglais
- annuel

INRIA

Le rapport d'activité de l'Inria

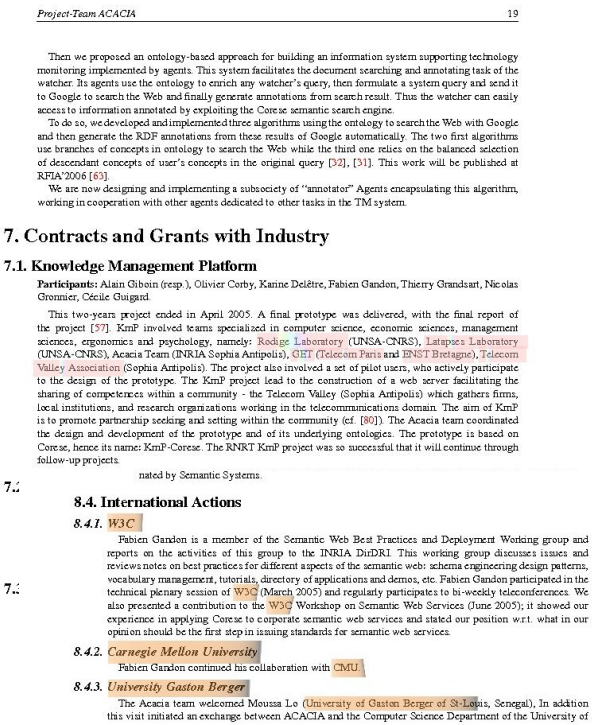
- Annexe technique
- 168 rapports d'équipes en 2005
- 4611 pages en 2005
- en anglais
- annuel

INRIA

Question

- Quels sont les partenaires de l'Inria ?

Réponse... dans le RA



Project-Team ACACIA 19

Then we proposed an ontology-based approach for building an information system supporting technology monitoring implemented by agents. This system facilitates the document searching and annotating task of the watcher. Its agents use the ontology to enrich any watcher's query, then formulate a system query and send it to Google to search the Web and finally generate annotations from search result. Thus the watcher can easily access to information annotated by exploiting the Corese semantic search engine.

To do so, we developed and implemented three algorithms using the ontology to search the Web with Google and then generate the RDF annotations from these results of Google automatically. The two first algorithms use branches of concepts in ontology to search the Web while the third one relies on the balanced selection of descendant concepts of user's concepts in the original query [32], [31]. This work will be published at RFIA'2006 [63].

We are now designing and implementing a subsociety of "annotator" Agents encapsulating this algorithm, working in cooperation with other agents dedicated to other tasks in the TM system.

7. Contracts and Grants with Industry

7.1. Knowledge Management Platform

Participants: Alain Giboin (resp.), Olivier Corby, Karine Delêtre, Fabien Gandon, Thierry Grandsart, Nicolas Gronnier, Cécile Guigard.

This two-years project ended in April 2005. A final prototype was delivered, with the final report of the project [57]. KmP involved teams specialized in computer science, economic sciences, management sciences, ergonomics and psychology, namely: Rodige Laboratory (UNSA-CNRS), Latapses Laboratory (UNSA-CNRS), Acacia Team (INRIA Sophia Antipolis), GET (Telecom Paris and ENST Bretagne), Telecom Valley Association (Sophia Antipolis). The project also involved a set of pilot users, who actively participate to the design of the prototype. The KmP project lead to the construction of a web server facilitating the sharing of competences within a community - the Telecom Valley (Sophia Antipolis) which gathers firms, local institutions, and research organizations working in the telecommunications domain. The aim of KmP is to promote partnership seeking and setting within the community (cf. [80]). The Acacia team coordinated the design and development of the prototype and of its underlying ontologies. The prototype is based on Corese, hence its name: KmP-Corese. The RNRT KmP project was so successful that it will continue through follow-up projects.

nated by Semantic Systems.

8.4. International Actions

8.4.1. W3C

Fabien Gandon is a member of the Semantic Web Best Practices and Deployment Working group and reports on the activities of this group to the INRIA DirDRI. This working group discusses issues and reviews notes on best practices for different aspects of the semantic web: schema engineering design patterns, vocabulary management, tutorials, directory of applications and demos, etc. Fabien Gandon participated in the technical plenary session of W3C (March 2005) and regularly participates to bi-weekly teleconferences. We also presented a contribution to the W3C Workshop on Semantic Web Services (June 2005); it showed our experience in applying Corese to corporate semantic web services and stated our position w.r.t. what in our opinion should be the first step in issuing standards for semantic web services.

8.4.2. Carnegie Mellon University

Fabien Gandon continued his collaboration with CMU.

8.4.3. University Gaston Berger

The Acacia team welcomed Moussa Lo (University of Gaston Berger of St-Louis, Senegal). In addition this visit initiated an exchange between ACACIA and the Computer Science Department of the University of

- Les parties “contrats” et “international” sont les plus pertinentes

INRIA

Réponse... dans le RA, oui mais...

16 *Activity Report INRIA 2005*

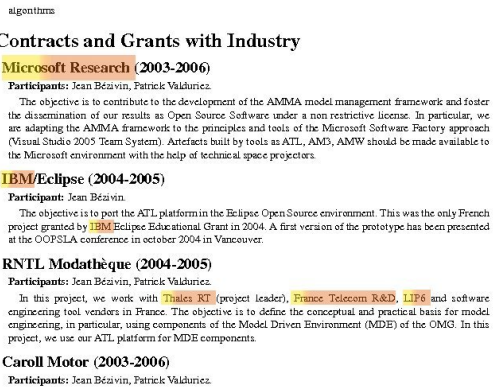
6. Contracts and Grants with Industry

6.1. Industrial Contracts

The Algorithms Project and Waterloo Maple Inc. (WMI) have developed a collaboration based on reciprocal interests. It is obviously interesting for the company to integrate functionalities at the forefront of the current research in computer algebra. Reciprocally, this integration makes our programs and our research visible to a very wide audience.

Numerous exchanges have thus taken place between the project and the company over the years. After more than 3 years within the project, J. Carette has been for several years Product Development Director at WMI, before going back to the academic world. Similarly, E. Murray, who worked for two years in the project developing the combstruct package is now working at WMI.

Thanks to all this activity, the company WMI considers INRIA as a special partner and grants it a free license for all of its research units. Moreover, a cooperation agreement has been signed between WMI and ALGO in 2001. In particular, one of the objectives is to replace all the routines dealing with asymptotic and series



algorithms

7. Contracts and Grants with Industry

7.1. Microsoft Research (2003-2006)

Participants: Jean Bézivin, Patrick Valduriez.

The objective is to contribute to the development of the AMMA model management framework and foster the dissemination of our results as Open Source Software under a non restrictive license. In particular, we are adapting the AMMA framework to the principles and tools of the Microsoft Software Factory approach (Visual Studio 2005 Team System). Artefacts built by tools as ATL, AM3, AMW should be made available to the Microsoft environment with the help of technical space projectors.

7.2. IBM/Eclipse (2004-2005)

Participant: Jean Bézivin.

The objective is to port the ATL platform in the Eclipse Open Source environment. This was the only French project granted by IBM Eclipse Educational Grant in 2004. A first version of the prototype has been presented at the OOPSLA conference in october 2004 in Vancouver.

7.3. RNTL Modathèque (2004-2005)

Participants: Jean Bézivin, Patrick Valduriez.

In this project, we work with Thales_RT (project leader), France_Telecom_R&D, LIP6 and software engineering tool vendors in France. The objective is to define the conceptual and practical basis for model engineering, in particular, using components of the Model Driven Environment (MDE) of the OMG. In this project, we use our ATL platform for MDE components.

7.4. Caroll Motor (2003-2006)

Participants: Jean Bézivin, Patrick Valduriez.

- Style très hétérogène
- Souvent peu rédigé
- Sigles plus ou moins développés
- Noms de réseaux, de labos, de localisation...

INRIA

Réponse... dans le RA, oui mais...

16 *Activity Report INRIA 2005*

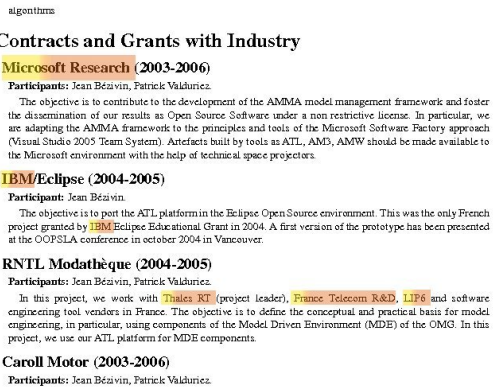
6. Contracts and Grants with Industry

6.1. Industrial Contracts

The Algorithms Project and Waterloo Maple Inc. (WMI) have developed a collaboration based on reciprocal interests. It is obviously interesting for the company to integrate functionalities at the forefront of the current research in computer algebra. Reciprocally, this integration makes our programs and our research visible to a very wide audience.

Numerous exchanges have thus taken place between the project and the company over the years. After more than 3 years within the project, J. Carette has been for several years Product Development Director at WMI, before going back to the academic world. Similarly, E. Murray, who worked for two years in the project developing the combstruct package is now working at WMI.

Thanks to all this activity, the company WMI considers INRIA as a special partner and grants it a free license for all of its research units. Moreover, a cooperation agreement has been signed between WMI and ALGO in 2001. In particular, one of the objectives is to replace all the routines dealing with asymptotic and series



algorithms

7. Contracts and Grants with Industry

7.1. Microsoft Research (2003-2006)

Participants: Jean Bézivin, Patrick Valduriez.

The objective is to contribute to the development of the AMMA model management framework and foster the dissemination of our results as Open Source Software under a non restrictive license. In particular, we are adapting the AMMA framework to the principles and tools of the Microsoft Software Factory approach (Visual Studio 2005 Team System). Artefacts built by tools as ATL, AM3, AMW should be made available to the Microsoft environment with the help of technical space projectors.

7.2. IBM/Eclipse (2004-2005)

Participant: Jean Bézivin.

The objective is to port the ATL platform in the Eclipse Open Source environment. This was the only French project granted by IBM Eclipse Educational Grant in 2004. A first version of the prototype has been presented at the OOPSLA conference in october 2004 in Vancouver.

7.3. RNTL Modathèque (2004-2005)

Participants: Jean Bézivin, Patrick Valduriez.

In this project, we work with Thales_RT (project leader), France_Telecom_R&D, LIP6 and software engineering tool vendors in France. The objective is to define the conceptual and practical basis for model engineering, in particular, using components of the Model Driven Environment (MDE) of the OMG. In this project, we use our ATL platform for MDE components.

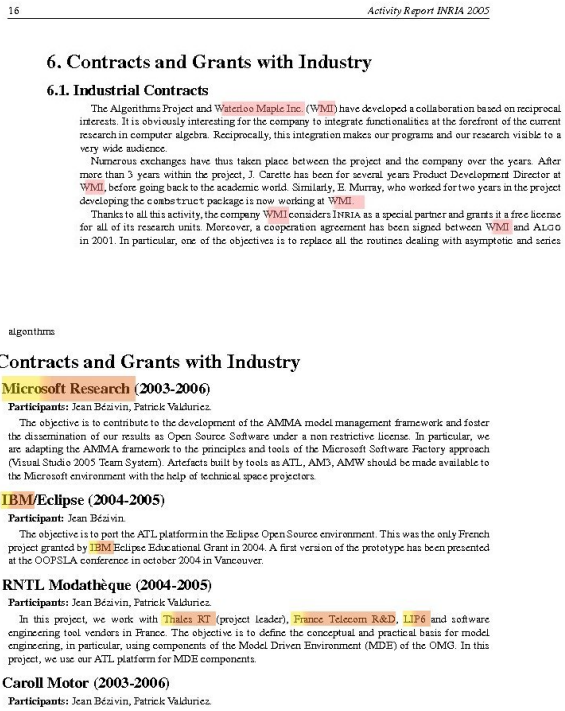
7.4. Caroll Motor (2003-2006)

Participants: Jean Bézivin, Patrick Valduriez.

- Style très hétérogène
- Souvent peu rédigé
- Sigles plus ou moins développés
- Noms de réseaux, de labos, de localisation...

INRIA

Réponse... dans le RA, oui mais...



16 *Activity Report INRIA 2005*

6. Contracts and Grants with Industry

6.1. Industrial Contracts

The Algorithms Project and Waterloo Maple Inc. (WMI) have developed a collaboration based on reciprocal interests. It is obviously interesting for the company to integrate functionalities at the forefront of the current research in computer algebra. Reciprocally, this integration makes our programs and our research visible to a very wide audience.

Numerous exchanges have thus taken place between the project and the company over the years. After more than 3 years within the project, J. Carette has been for several years Product Development Director at WMI, before going back to the academic world. Similarly, E. Murray, who worked for two years in the project developing the combstruct package is now working at WMI.

Thanks to all this activity, the company WMI considers INRIA as a special partner and grants it a free license for all of its research units. Moreover, a cooperation agreement has been signed between WMI and ALGO in 2001. In particular, one of the objectives is to replace all the routines dealing with asymptotic and series

algorithms

7. Contracts and Grants with Industry

7.1. Microsoft Research (2003-2006)

Participants: Jean Bézivin, Patrick Valduriez.

The objective is to contribute to the development of the AMMA model management framework and foster the dissemination of our results as Open Source Software under a non restrictive license. In particular, we are adapting the AMMA framework to the principles and tools of the Microsoft Software Factory approach (Visual Studio 2005 Team System). Artefacts built by tools as ATL, AM3, AMW should be made available to the Microsoft environment with the help of technical space projectors.

7.2. IBM/Eclipse (2004-2005)

Participant: Jean Bézivin.

The objective is to port the ATL platform in the Eclipse Open Source environment. This was the only French project granted by IBM Eclipse Educational Grant in 2004. A first version of the prototype has been presented at the OOPSLA conference in october 2004 in Vancouver.

7.3. RNTL Modathèque (2004-2005)

Participants: Jean Bézivin, Patrick Valduriez.

In this project, we work with Thales_RT (project leader), France_Telecom_R&D, LIP6 and software engineering tool vendors in France. The objective is to define the conceptual and practical basis for model engineering, in particular, using components of the Model Driven Environment (MDE) of the OMG. In this project, we use our ATL platform for MDE components.

7.4. Caroll Motor (2003-2006)

Participants: Jean Bézivin, Patrick Valduriez.

- Style très hétérogène
- Souvent peu rédigé
- Sigles plus ou moins développés
- Noms de réseaux, de labos, de localisation...

INRIA

Réponse... dans le RA, oui mais...

16 *Activity Report INRIA 2005*

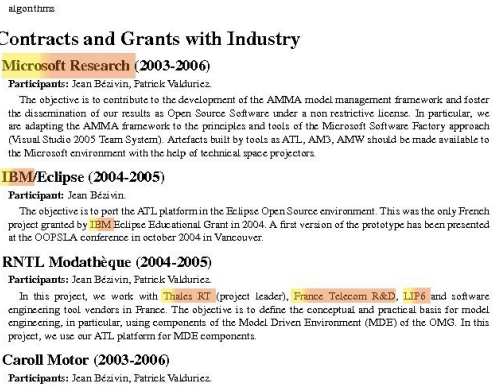
6. Contracts and Grants with Industry

6.1. Industrial Contracts

The Algorithms Project and Waterloo Maple Inc. (WMI) have developed a collaboration based on reciprocal interests. It is obviously interesting for the company to integrate functionalities at the forefront of the current research in computer algebra. Reciprocally, this integration makes our programs and our research visible to a very wide audience.

Numerous exchanges have thus taken place between the project and the company over the years. After more than 3 years within the project, J. Carette has been for several years Product Development Director at WMI, before going back to the academic world. Similarly, E. Murray, who worked for two years in the project developing the combstruct package is now working at WMI.

Thanks to all this activity, the company WMI considers INRIA as a special partner and grants it a free license for all of its research units. Moreover, a cooperation agreement has been signed between WMI and ALGO in 2001. In particular, one of the objectives is to replace all the routines dealing with asymptotic and series



algorithms

7. Contracts and Grants with Industry

7.1. Microsoft Research (2003-2006)

Participants: Jean Bézivin, Patrick Valduriez.

The objective is to contribute to the development of the AMMA model management framework and foster the dissemination of our results as Open Source Software under a non restrictive license. In particular, we are adapting the AMMA framework to the principles and tools of the Microsoft Software Factory approach (Visual Studio 2005 Team System). Artefacts built by tools as ATL, AM3, AMW should be made available to the Microsoft environment with the help of technical space projectors.

7.2. IBM/Eclipse (2004-2005)

Participant: Jean Bézivin.

The objective is to port the ATL platform in the Eclipse Open Source environment. This was the only French project granted by IBM Eclipse Educational Grant in 2004. A first version of the prototype has been presented at the OOPSLA conference in october 2004 in Vancouver.

7.3. RNTL Modathèque (2004-2005)

Participants: Jean Bézivin, Patrick Valduriez.

In this project, we work with Thales_RT (project leader), France_Telecom_R&D, LIP6 and software engineering tool vendors in France. The objective is to define the conceptual and practical basis for model engineering, in particular, using components of the Model Driven Environment (MDE) of the OMG. In this project, we use our ATL platform for MDE components.

7.4. Caroll Motor (2003-2006)

Participants: Jean Bézivin, Patrick Valduriez.

- Style très hétérogène
- Souvent peu rédigé
- Sigles plus ou moins développés
- Noms de réseaux, de labos, de localisation...

INRIA

Extraction d'entités nommées

- Entités nommées : noms de personnes, d'organisations, de lieux, dates, valeurs monétaires...
- Extraction : tester l'existence ou trouver toutes les occurences

Extraction d'entités nommées

- Entités nommées : noms de personnes, d'organisations, de lieux, dates, valeurs monétaires...
- Extraction : tester l'existence ou trouver toutes les occurences

Existant

- Outils commerciaux : Rosette Entity Extractor, Inxight SmartDiscovery, Convera-RetrievalWare, Xerox...
- Large communauté, ex : univ. de Sheffield (ANNIE, GATE)
- Utilisent généralement des ressources linguistiques importantes (dictionnaires), et de grandes collections pour faire l'apprentissage (méthodes statistiques)
- Nécessitent une adaptation manuelle pour un corpus particulier (par exemple écriture de règles)
- Xeros : 250 règles manuelles pour extraire des entités biologiques

Existant

- Outils commerciaux : Rosette Entity Extractor, Inxight SmartDiscovery, Convera-RetrievalWare, Xerox...
- Large communauté, ex : univ. de Sheffield (ANNIE, GATE)
- Utilisent généralement des ressources linguistiques importantes (dictionnaires), et de grandes collections pour faire l'apprentissage (méthodes statistiques)
- Nécessitent une adaptation manuelle pour un corpus particulier (par exemple écriture de règles)
- Xeros : 250 règles manuelles pour extraire des entités biologiques

Existant

- Outils commerciaux : Rosette Entity Extractor, Inxight SmartDiscovery, Convera-RetrievalWare, Xerox...
- Large communauté, ex : univ. de Sheffield (ANNIE, GATE)
- Utilisent généralement des ressources linguistiques importantes (dictionnaires), et de grandes collections pour faire l'apprentissage (méthodes statistiques)
- Nécessitent une adaptation manuelle pour un corpus particulier (par exemple écriture de règles)
- Xeros : 250 règles manuelles pour extraire des entités biologiques

Existant

- Outils commerciaux : Rosette Entity Extractor, Inxight SmartDiscovery, Convera-RetrievalWare, Xerox...
- Large communauté, ex : univ. de Sheffield (ANNIE, GATE)
- Utilisent généralement des ressources linguistiques importantes (dictionnaires), et de grandes collections pour faire l'apprentissage (méthodes statistiques)
- Nécessitent une adaptation manuelle pour un corpus particulier (par exemple écriture de règles)
- Xeros : 250 règles manuelles pour extraire des entités biologiques

Existant

- Outils commerciaux : Rosette Entity Extractor, Inxight SmartDiscovery, Convera-RetrievalWare, Xerox...
- Large communauté, ex : univ. de Sheffield (ANNIE, GATE)
- Utilisent généralement des ressources linguistiques importantes (dictionnaires), et de grandes collections pour faire l'apprentissage (méthodes statistiques)
- Nécessitent une adaptation manuelle pour un corpus particulier (par exemple écriture de règles)
- Xeros : 250 règles manuelles pour extraire des entités biologiques

Idées de base

- Utiliser une méthode similaire aux méthodes de wrapping
- On remplace les balises HTML par les syntagmes du langage (nom, verbe, etc.)
- Apprentissage à partir de peu d'exemples
- Ne pas utiliser notre connaissance de la langue
- Réinjecter les résultats d'une année l'année suivante

Idées de base

- Utiliser une méthode similaire aux méthodes de wrapping
- On remplace les balises HTML par les syntagmes du langage (nom, verbe, etc.)
- Apprentissage à partir de peu d'exemples
- Ne pas utiliser notre connaissance de la langue
- Réinjecter les résultats d'une année l'année suivante

Idées de base

- Utiliser une méthode similaire aux méthodes de wrapping
- On remplace les balises HTML par les syntagmes du langage (nom, verbe, etc.)
- Apprentissage à partir de peu d'exemples
- Ne pas utiliser notre connaissance de la langue
- Réinjecter les résultats d'une année l'année suivante

Idées de base

- Utiliser une méthode similaire aux méthodes de wrapping
- On remplace les balises HTML par les syntagmes du langage (nom, verbe, etc.)
- Apprentissage à partir de peu d'exemples
- Ne pas utiliser notre connaissance de la langue
- Réinjecter les résultats d'une année l'année suivante

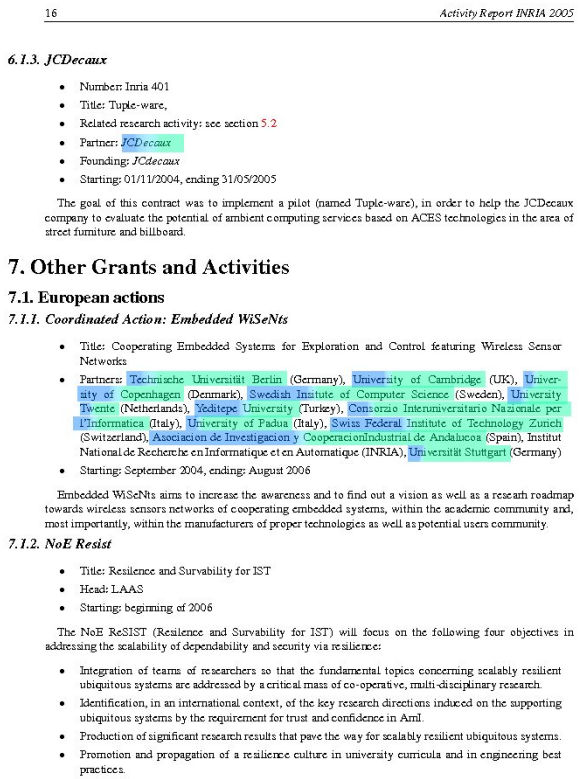
Idées de base

- Utiliser une méthode similaire aux méthodes de wrapping
- On remplace les balises HTML par les syntagmes du langage (nom, verbe, etc.)
- Apprentissage à partir de peu d'exemples
- Ne pas utiliser notre connaissance de la langue
- Réinjecter les résultats d'une année l'année suivante

Schémas de phrases

- ...by Texas Instruments because...
- IN NNP NNPS IN
- <in>by</in>
  <org><nnp>Texas</nnp>
  <nnps>Instruments</nnps></org>
  <in>because</in>
- IN ~ NNPS* % ~ IN

Schémas de listes



16 *Activity Report INRIA 2005*

6.1.3. JCDecaux

- Number: Inria 401
- Title: Tuple-ware,
- Related research activity: see section 5.2
- Partner: *JCDecaux*
- Founding: *JCdecaux*
- Starting: 01/11/2004, ending 31/05/2005

The goal of this contract was to implement a pilot (named Tuple-ware), in order to help the JCDecaux company to evaluate the potential of ambient computing services based on ACES technologies in the area of street furniture and billboard.

7. Other Grants and Activities

7.1. European actions

7.1.1. Coordinated Action: Embedded WiSeNts

- Title: Cooperating Embedded Systems for Exploration and Control featuring Wireless Sensor Networks
- Partners: Technische Universität Berlin (Germany), University of Cambridge (UK), University of Copenhagen (Denmark), Swedish Insitute of Computer Science (Sweden), University Twente (Netherlands), Yeditepe University (Turkey), Consorzio Interuniversitario Nazionale per l'Informatica (Italy), University of Padua (Italy), Swiss Federal Institute of Technology Zurich (Switzerland), Asociacion de Investigacion y CooperacionIndustrial de Andalucea (Spain), Institut National de Recherche en Informatique et en Automatique (INRIA), Universität Stuttgart (Germany)
- Starting: September 2004, ending: August 2006

Embedded WiSeNts aims to increase the awareness and to find out a vision as well as a research roadmap towards wireless sensors networks of cooperating embedded systems, within the academic community and, most importantly, within the manufacturers of proper technologies as well as potential users community.

7.1.2. NoE Resist

- Title: Resilence and Survability for IST
- Head: LAAS
- Starting: beginning of 2006

The NoE ReSIST (Resilence and Survability for IST) will focus on the following four objectives in addressing the scalability of dependability and security via resilience:

- Integration of teams of researchers so that the fundamental topics concerning scalably resilient ubiquitous systems are addressed by a critical mass of co-operative, multi-disciplinary research.
- Identification, in an international context, of the key research directions induced on the supporting ubiquitous systems by the requirement for trust and confidence in AmI.
- Production of significant research results that pave the way for scalably resilient ubiquitous systems.
- Promotion and propagation of a resilience culture in university curricula and in engineering best practices.

- Certaines stuctures régulières (les listes) revêtent une importance particulière : ORG, NNPS*

INRIA

Extraction de schémas

- Étant donné un nom d'organisme, on construit tous les schémas contenant jusqu'à 5 syntagmes à droite ou à gauche
- Seuls les "bons" schémas seront ensuite conservés en fonction de leur performance lors d'une phase d'apprentissage

Extraction de schémas

- Étant donné un nom d'organisme, on construit tous les schémas contenant jusqu'à 5 syntagmes à droite ou à gauche
- Seuls les "bons" schémas seront ensuite conservés en fonction de leur performance lors d'une phase d'apprentissage

Méthode

- Un ensemble de départ : L
- Un ensemble d'apprentissage : A
- Un ensemble de test : B
- Ces ensembles de documents sont annotés à la main
- Essais sur différentes combinaisons

Jeu de test 1	L	A	B	total
Documents	4	8	8	20
Entités différentes	74	238	144	456
Occurrences d'entités	238	418	271	827

Méthode

- Un ensemble de départ : L
- Un ensemble d'apprentissage : A
- Un ensemble de test : B
- Ces ensembles de documents sont annotés à la main
- Essais sur différentes combinaisons

Jeu de test 1	L	A	B	total
Documents	4	8	8	20
Entités différentes	74	238	144	456
Occurrences d'entités	238	418	271	827

Méthode

- Un ensemble de départ : L
- Un ensemble d'apprentissage : A
- Un ensemble de test : B
- Ces ensembles de documents sont annotés à la main
- Essais sur différentes combinaisons

Jeu de test 1	L	A	B	total
Documents	4	8	8	20
Entités différentes	74	238	144	456
Occurrences d'entités	238	418	271	827

Méthode

- Un ensemble de départ : L
- Un ensemble d'apprentissage : A
- Un ensemble de test : B
- Ces ensembles de documents sont annotés à la main
- Essais sur différentes combinaisons

Jeu de test 1	L	A	B	total
Documents	4	8	8	20
Entités différentes	74	238	144	456
Occurrences d'entités	238	418	271	827

Méthode

- Un ensemble de départ : L
- Un ensemble d'apprentissage : A
- Un ensemble de test : B
- Ces ensembles de documents sont annotés à la main
- Essais sur différentes combinaisons

Jeu de test 1	L	A	B	total
Documents	4	8	8	20
Entités différentes	74	238	144	456
Occurrences d'entités	238	418	271	827

Algorithme

- Soit les organismes qui sont annotés dans L
- Construire la liste des schémas de A+L qui les contiennent
- Appliquer ces schémas sur A ; supprimer ceux qui génèrent peu de résultats corrects ou trop d'incorrects ; classer par performance (nb de corrects/nb d'incorrects)
- Apprentissage sur A ; les schémas sont appliqués un à un dans l'ordre ; arrêt quand :
  présision < seuil (la précision est initialement à 1)
  rappel > seuil
- Test sur B (contrôle de la précision et du rappel)
- On applique sur tout les documents (éventuellement calcul de la précision)
- Uliser le résultat "nettoyé" comme base de départ l'année suivante

Algorithme

- Soit les organismes qui sont annotés dans L
- Construire la liste des schémas de A+L qui les contiennent
- Appliquer ces schémas sur A ; supprimer ceux qui génèrent peu de résultats corrects ou trop d'incorrects ; classer par performance (nb de corrects/nb d'incorrects)
- Apprentissage sur A ; les schémas sont appliqués un à un dans l'ordre ; arrêt quand :
  présision < seuil (la précision est initialement à 1)
  rappel > seuil
- Test sur B (contrôle de la précision et du rappel)
- On applique sur tout les documents (éventuellement calcul de la précision)
- Uliser le résultat "nettoyé" comme base de départ l'année suivante

Algorithme

- Soit les organismes qui sont annotés dans L
- Construire la liste des schémas de A+L qui les contiennent
- Appliquer ces schémas sur A ; supprimer ceux qui génèrent peu de résultats corrects ou trop d'incorrects ; classer par performance (nb de corrects/nb d'incorrects)
- Apprentissage sur A ; les schémas sont appliqués un à un dans l'ordre ; arrêt quand :
  présision < seuil (la précision est initialement à 1)
  rappel > seuil
- Test sur B (contrôle de la précision et du rappel)
- On applique sur tout les documents (éventuellement calcul de la précision)
- Uliser le résultat "nettoyé" comme base de départ l'année suivante

Algorithme

- Soit les organismes qui sont annotés dans L
- Construire la liste des schémas de A+L qui les contiennent
- Appliquer ces schémas sur A ; supprimer ceux qui génèrent peu de résultats corrects ou trop d'incorrects ; classer par performance (nb de corrects/nb d'incorrects)
- Apprentissage sur A ; les schémas sont appliqués un à un dans l'ordre ; arrêt quand :
  présision < seuil (la précision est initialement à 1)
  rappel > seuil
- Test sur B (contrôle de la précision et du rappel)
- On applique sur tout les documents (éventuellement calcul de la précision)
- Uliser le résultat "nettoyé" comme base de départ l'année suivante

Algorithme

- Soit les organismes qui sont annotés dans L
- Construire la liste des schémas de A+L qui les contiennent
- Appliquer ces schémas sur A ; supprimer ceux qui génèrent peu de résultats corrects ou trop d'incorrects ; classer par performance (nb de corrects/nb d'incorrects)
- Apprentissage sur A ; les schémas sont appliqués un à un dans l'ordre ; arrêt quand :
  présision < seuil (la précision est initialement à 1)
  rappel > seuil
- Test sur B (contrôle de la précision et du rappel)
- On applique sur tout les documents (éventuellement calcul de la précision)
- Uliser le résultat "nettoyé" comme base de départ l'année suivante

Algorithme

- Soit les organismes qui sont annotés dans L
- Construire la liste des schémas de A+L qui les contiennent
- Appliquer ces schémas sur A ; supprimer ceux qui génèrent peu de résultats corrects ou trop d'incorrects ; classer par performance (nb de corrects/nb d'incorrects)
- Apprentissage sur A ; les schémas sont appliqués un à un dans l'ordre ; arrêt quand :
  présision < seuil (la précision est initialement à 1)
  rappel > seuil
- Test sur B (contrôle de la précision et du rappel)
- On applique sur tout les documents (éventuellement calcul de la précision)
- Uliser le résultat "nettoyé" comme base de départ l'année suivante

Algorithme

- Soit les organismes qui sont annotés dans L
- Construire la liste des schémas de A+L qui les contiennent
- Appliquer ces schémas sur A ; supprimer ceux qui génèrent peu de résultats corrects ou trop d'incorrects ; classer par performance (nb de corrects/nb d'incorrects)
- Apprentissage sur A ; les schémas sont appliqués un à un dans l'ordre ; arrêt quand :
  présision < seuil (la précision est initialement à 1)
  rappel > seuil
- Test sur B (contrôle de la précision et du rappel)
- On applique sur tout les documents (éventuellement calcul de la précision)
- Uliser le résultat "nettoyé" comme base de départ l'année suivante

Résultats

Seuil 0,6	Jeu de test 1	Jeu de test 2	Jeu de test 3
Entités dans L	74	93	114
Entités de L trouvées dans A	26	10	23
Occurrences d'entités de L trouvées dans A	116 (R=0,17 ; P=0,92) (MR=0,26 ; MP=0,94)	44 (R=0,17 ; P=0,96) (MR=0,22 ; MP=0,89)	51 (R=0,21 ; P=1) (MR=0,25 ; MP=0,96)
Schémas retenus	325	250	126
Entités dans A à la fin de l'apprentissage	247 (R=0,56 ; P=0,60) (MR=0,67 ; MP=0,63)	183 (R=0,56 ; P=0,60) (MR=0,67 ; MP=0,62)	126 (R=0,61 ; P=0,90) (MR=0,64 ; MP=0,91)
Entités de L trouvées dans B	12	24	17
Entités extraites de B	140	330	240
Comptage simple	R=0,31 ; P=0,46	R=0,45 ;P=0,41	R=0,36 ; P=0,47
Comptage multiple	R=0,35 ; P=0,44	R=0,53 ;P=0,41	R=0,42 ; P=0,49

INRIA

Résultats

- Expériences avec différents seuils pour l'apprentissage (la précision est prépondérente : pourquoi un seuil sur le rappel ?) ; si on demande plus de précision, on extrait moins d'entitées
- Expériences en séparant les parties "contrats" et "collaborations" : résultats moins bons

Résultats

- Expériences avec différents seuils pour l'apprentissage (la précision est prépondérente : pourquoi un seuil sur le rappel ?) ; si on demande plus de précision, on extrait moins d'entitées
- Expériences en séparant les parties "contrats" et "collaborations" : résultats moins bons

Conclusion

- Problème difficile
- Ne marche pas très bien car les schémas sont trop génériques
- Aucune utilisation d'une connaissance de la langue

Conclusion

- Problème difficile
- Ne marche pas très bien car les schémas sont trop génériques
- Aucune utilisation d'une connaissance de la langue

Conclusion

- Problème difficile
- Ne marche pas très bien car les schémas sont trop génériques
- Aucune utilisation d'une connaissance de la langue

Perspectives

- Être moins générique
- Utiliser des propriétes lexicales (majuscules...)
- Utiliser des “valeurs” de syntagmes (partners, University of...) et des schémas prédéfinis
- Sélectionner plus localement les schémas (style d’écriture)
- Utiliser plus de connaissance linguistique
- ...

Perspectives

- Être moins générique
- Utiliser des propriétes lexicales (majuscules...)
- Utiliser des “valeurs” de syntagmes (partners, University of...) et des schémas prédéfinis
- Sélectionner plus localement les schémas (style d’écriture)
- Utiliser plus de connaissance linguistique
- ...

Perspectives

- Être moins générique
- Utiliser des propriétes lexicales (majuscules...)
- Utiliser des “valeurs” de syntagmes (partners, University of...) et des schémas prédéfinis
- Sélectionner plus localement les schémas (style d’écriture)
- Utiliser plus de connaissance linguistique
- ...

Perspectives

- Être moins générique
- Utiliser des propriétes lexicales (majuscules...)
- Utiliser des “valeurs” de syntagmes (partners, University of...) et des schémas prédéfinis
- Sélectionner plus localement les schémas (style d’écriture)
- Utiliser plus de connaissance linguistique
- ...

Perspectives

- Être moins générique
- Utiliser des propriétes lexicales (majuscules...)
- Utiliser des “valeurs” de syntagmes (partners, University of...) et des schémas prédéfinis
- Sélectionner plus localement les schémas (style d’écriture)
- Utiliser plus de connaissance linguistique
- ...

Perspectives

- Être moins générique
- Utiliser des propriétes lexicales (majuscules...)
- Utiliser des “valeurs” de syntagmes (partners, University of...) et des schémas prédéfinis
- Sélectionner plus localement les schémas (style d’écriture)
- Utiliser plus de connaissance linguistique
- ...

Fin

- Merci...