IN PARTNERSHIP WITH:
CNRS

Université de Lorraine

Activity Report 2011

Project-Team TALARIS

Natural Language Processing: representation, inference and semantics

IN COLLABORATION WITH: Laboratoire lorrain de recherche en informatique et ses applications (LORIA)

RESEARCH CENTER
Nancy - Grand Est

THEME
Audio, Speech, and Language Processing

Table of contents

Project-Team TALARIS

Keywords: Inference, Knowledge Representation, Natural Language, Semantics, Multimedia

1. Members

Research Scientists

Claire Gardent [Team Leader, Senior Research, CNRS (SHS Department), HdR]
Carlos Areces [Junior Researcher, INRIA (On sabatical leave 2010-04-01 / 2011-04-01, on leave since 2011-04-01)]

Faculty Members

Lotfi Bellalem [PRAG, ESIAL, UHP Nancy 1]
Nadia Bellalem [Associate Professor, IUT Nancy-Charlemagne, University of Nancy 2]
Samuel Cruz-Lara [Associate Professor, IUT Nancy-Charlemagne, University of Nancy 2]
Christine Fay-Varnier [Associate Professor, School of Geology, INPL]
Jean-Charles Lamirel [Associate Professor, IUT Robert Schuman, University of Strasbourg, HdR]
Fabienne Venant [Associate Professor, IUT Nancy-Charlemagne, University of Nancy 2. On parental leave from 2010-06-01]

Technical Staff

Emmanuel Didiot [INRIA Engineer on CPER MISN/TALC, from 2010-12-01.]
German Kruszewski [Engineer on INTERREG IV A Allegro project, from 2011-09-15.]
Treveur Bretaudiere [INRIA Engineer on ADT MoViTAL, from 2010-12-01]

PhD Students

Corinna Anderson [Yale University, USA]
Ingrid Falk [University of Nancy 2, SemBySem and Allegro project, from 2008-10-01 until 2011-09-30]
Laura Perez [University of Nancy 2, Ministry grant, from 2009-10-01 until 2012-10-01]
Alejandra Lorenzo [UHP Nancy 1, Région/ALLEGRO Interreg IV A project, from 2010-10-15]
Shashi Narayan [UHP Nancy 1, Ministry Grant, from 2011-10-01]

Post-Doctoral Fellows

Marilisa Amoia [Postdoc on INTERREG IV A Allegro project, from 2011-01-01 till 2011-08-31.]
Alexandre Denis [Postdoc on ANR CCCP Project, from 2008-03-01 till 2012-03-15]
Lina-Maria Rojas Barahona [Postdoc on EUROSTAR Emospeech project, from 2010-12-01.]

Administrative Assistant

Isabelle Blanchard

2. Overall Objectives

2.1. Background

TALARIS stands for *Traitment Automatique des Langues: Representation, Inference, et Semantique*. As this name suggests, the aim of the TALARIS team is to investigate semantic phenomena (broadly constructed) in natural language from a computational perspective. More concretely, TALARIS's goal is to develop grammars (with a special emphasis on French) with a semantic dimension, to explore the linguistic and computational issues involved in such areas as natural language generation, dialog modeling and multilinguality; and to investigate the interplay between representation and inference in computational semantics for natural language.

2.2. Organization

The work of the TALARIS team can be subdivided into four overlapping and mutually supporting categories.

Computational Semantics. This theme is devoted to the theoretical and computational issues involved in building semantic representations for natural language. Special emphasis is placed on developing large scale semantic coverage for the French Language.

Discourse, Dialogue and Pragmatics. This theme is devoted to developing theoretical and computational models of discourse and dialogue processing, and investigating the inferential impact of pragmatic factors (that is, the factors affecting how human beings actually use language).

Logics for Natural Language and Knowledge Representation. The theme is devoted to theoretical and computational tools for working with logics suitable for natural language inference and knowledge representation. Special emphasis is place on hybrid logic, higher order logic, and discourse representation theory (DRT).

Multilinguality for Multimedia. This theme is devoted to creating generic ISO-based mechanisms for representing and dealing with multilingual textual information. The center of this activity is the MLIF (Multi Lingual Information Framework) specification platform for elementary multilingual units.

2.3. Overall Objectives

The major long term computational goals of the TALARIS team are:

- The design and implementation of powerful clustering techniques which support both the incremental classification of large amount of heterogeneous textual data and a detailed, supervised and unsupervised, evaluation of the output clusters.
- The creation of a large scale computational semantics framework for French that supports deep semantic analysis and sentence generation.
- The integration and use of this framework in systems interfacing 3D worlds and Natural Language Processing (NLP) technologies e.g., extending a serious game with dialog capabilities or exploiting Natural Language Generation to automate the production of language learning exercices in a 3D setting.
- The creation of efficient inference systems for logics that are capable of representing natural language content and the background knowledge required to support reasoning.
- The integration of language technology and semantic resources into multimedia applications.
- The development of standards for representing multilingual textual information and its interaction with various media.

These computational goals are pursued in the context of theoretical investigations that rigorously map out the required scientific and mathematical context.

2.4. Highlights

2.4.1. Generating Instructions in a Virtual Environment.

Talaris participated in the preparation of the international GIVE 2.5 challenge on the Generation of Instructions in a Virtual Environment. This challenge brought together researchers from six universities and evaluated the participating systems on their ability to generate instructions in a dynamic 3D setting [40]. One of the two systems developed by TALARIS [26] won the second place both in terms of objective and of subjective metrics.

2.4.2. Generating from Knowledge Bases.

Talaris' work on data-to-text generation has attracted increasing interest this year. The GenI sentence generator developed by Talaris was licensed to be used by Stanford Research International (SRI) in the large scale AURA (Automated User-centered Reasoning and Acquisition System) project whose aim is to provide an Intelligent Electronic Textbook that could be used by teachers and students; negociations are currently underway with SeeReason Partners LLC for another commercial license; and a two day invited visit to Bolzano lead to the launch of a new collaboration which will focus on integrating GenI in the Quelo system developed by the KRDB group to verbalise queries on knowledge bases. In parallel, Claire Gardent gave an invited talk on data-to-text generation at the French NLP conference (TALN, Traitement Automatique des Langues Naturelles) and an invited tutorial on Generation for the Semantic Web at the K-CAP (Knowledge Catpure) conference in Banff, Canada.

3. Scientific Foundations

3.1. Computational Linguistics and Computational Logic

The scientific foundations of TALARIS's work boil down to the motto: *computational linguistics* meets *computational logic* and *knowledge representation*.

From computational linguistics we take the large linguistic and lexical semantics resources, the parsing and generation algorithms, and the insight that (whenever possible) statistical information should be employed to cope with ambiguity. From computational logic and knowledge representation we take the various languages and methodologies that have been developed for handling different forms of information (such as temporal information), the computational tools (such as theorem provers, model builders, model checkers, sat-solvers and planners) that have been devised for working with them, together with the insight that, whenever possible, it is better to work with inference tools that have been tuned for particular problems, and moreover that, whenever possible, it is best to devote as little computational energy to inference as possible.

This picture is somewhat idealized. For example, for many languages (and French is one of them) the large scale linguistic resources (lexicons, grammars, WordNet, FrameNet, PropBank, etc.) that exist for English are not yet available. In addition, the syntax/semantics interface often cannot be taken for granted, and existing inference tools often need to be adapted to cope with the logics that arise in natural language applications (for example, existing provers for Description Logic, though excellent, do not cope with temporal reasoning). Thus we are not simply talking about bringing together known tools, and investigating how they work once they are combined — often a great deal of research, background work and development is needed. Nonetheless, the ideal of bringing together the best tools and ideas from computational linguistics, knowledge representation and computational logic and putting them to work in coordination is the guiding line.

3.2. Semantics and Inference

Over the next decade, progress in natural language semantics is likely to depend on obtaining a deeper understanding of the role played by inference. One of the simplest levels at which inference enters natural language is as a disambiguation mechanism. Utterances in natural language are typically highly ambiguous: inference allows human beings to (seemingly effortlessly) eliminate the irrelevant possibilities and isolate the intended meaning. But inference can be used in many other processes, for example, in the integration of new information into a known context. This is important when generating natural language utterances. For this task we need to be sure that the utterance we generate is suitable for the person being addressed. That is, we need to be sure that the generated representations fit in well with the recipient's knowledge and expectations of the world, and it is inference which guides us in achieving this.

Much recent semantic research actively addresses such problems by systematically integrating inference as a key element. This is an interesting development, as such work redefines the boundary between semantics and pragmatics. For example, van der Sandt's algorithm for presupposition resolution (a classic problem of pragmatics) uses inference to guarantee that new information is integrated in a coherent way with the old information.

The TALARIS team investigates such semantic/pragmatic problems from various angles (for example, from generation and discourse analysis perspectives) and tries to combine the insights offered by different approaches. For example, for some applications (e.g., the textual entailment recognition task) shallow syntactic parsing combined with fast inference in description logic may be the most suitable approach. In other cases, deep analysis of utterances or sentences and the use of a first-order inference engine may be better. Our aim is to explore these approaches and their limitations.

3.3. Linguistic Resources

In an ideal world, computational semanticists would not have to worry overly much about linguistic resources. Large scale lexica, treebanks, and wide coverage grammars (supported by fast parsers and offering a flexible syntax semantics interface) would be freely available and easy to combine and use. The semanticist could then focus on modeling semantic phenomena and their interactions.

Needless to say, in reality matters are not nearly so straightforward. For a start, for many languages (including French) there are no large-scale resources of the sort that exist for English. Furthermore even in the case of English, the idealized situation just sketched does not obtain. For example, the syntax/semantics interface cannot be regarded as a solved problem: phenomena such as gapping and VP-ellipsis (where a verb, or verb phrase, in a coordinated sentence is missing and has to be somehow "reconstructed" from the previous context) still offer challenging problems for semantic construction.

Thus a team like TALARIS simply cannot focus exclusively on semantic issues: it must also have competence in developing and maintaining a number of different lexical resources (and in particular, resources for French).

TALARIS is involved in such aspects in a number of ways. For example, it participates in the development of an open source syntactic and synonymic lexicon for French, in an attempt to lay the ground for a French version of FrameNet; and it also works on developing a large scale, reversible (i.e., usable both for parsing and for generation) Tree Adjoining Grammar for French.

3.4. Logic Engineering

Once again, in the ideal world, not only would computational semanticists not have to worry about the linguistic resources at their disposal, but they would not have to worry about the inference tools available either. These could be taken for granted, applied as needed, and the semanticist could concentrate on developing linguistically inspired inference architectures. But in spite of the spectacular progress made in automated theorem proving (both for very expressive logics like predicate logics, and for weak logics like description logics) over the last decade, we are not yet in the ideal world. The tools currently offered by the automated reasoning community still have a number of drawbacks when it comes to natural language applications.

For a start, most of the efforts of the first-order automated reasoning community have been devoted to theorem proving; model building, which is also a useful technology for natural language processing, is nowhere nearly as well developed, and far fewer systems are available. Secondly, the first-order reasoning community has adopted a resolutely 'classical' approach to inference problems: their provers focus exclusively on the satisfiability problem. The description logic community has been much more flexible, offering architectures and optimisations which allow a greater range of problems to be handled more directly. One reason for this has been that, historically, not all description logics offered full Boolean expressivity. So there is a long tradition in description logic of treating a variety of inference problems directly, rather than via reduction to satisfiability. Thirdly, many of the logics for which optimised provers exists do not directly offer the kinds of expressivity required for natural language applications. For example, it is hard to encode temporal inference problems in

implemented versions of description logics. Fourth, for very strong logics (notably higher-order logics) few implementations exists and their performance is currently inadequate.

These problems are not insurmountable, and TALARIS members are actively investigating ways of overcoming them. For a start, logics such as higher-order logic, description logic and hybrid logic are nowadays thought of as various fragments of (or theories expressed in) first-order logic. That is, first-order logic provides a unifying framework that often allows transfer of tools or testing methodologies to a wide range of logics. For example, the hybrid logics used in TALARIS (which can be thought of as more expressive versions of description logics) make heavy use of optimization techniques from first-order theorem proving.

3.5. Empirical Studies

The role of empirical methods (model learning, data extraction from corpora, evaluation) has greatly increased in importance in both linguistics and computer science over the last fifteen years. TALARIS members have been working for many years on the creation, management and dissemination of linguistic resources reusable by the scientific community, both in the context of implementation of data servers, and in the definition of standardized representation formats like TAG-ML. In addition, they have also worked on the applications of linguistic ideas in multimodal settings and multimedia.

Such work is important to our scientific goals. As we said above, one of the most important points that needs to be understood about logical inference is how its use can be minimized and intelligently guided. Ultimately, such minimization and guidance must be based on empirical observations concerning the kinds of problems that arise repeatedly in natural language applications.

Finally, is should be remarked that the emphasis on empirical studies lends another dimension to what is meant by inference. While much of TALARIS's focus is on symbolic approaches to inference, statistical and probabilistic methods, either on their own or blended with symbolic approaches, are likely to play an increasingly important role in the future. TALARIS researchers are well aware of the importance of such approaches and are interested in exploring their strengths and weaknesses, and where relevant, intend to integrate them into their work.

4. Application Domains

4.1. Grammar building and Linguistic Analysis

Developing large scale computational grammars permits a precise documentation and analysis of natural language phenomena. In collaboration with Calligramme, Talaris has developed a grammar compiler (XMG, Extended Metagrammar) which supports the computational specification of large scale, multi-dimensional tree grammars[1]. One long term application pursued by Talaris in the domain of computational linguistics is the development of a large scale Feature Based Lexicalised Tree Adjoining Grammar describing both the syntax and the semantics of French.

4.2. Surface Realization

As mentioned above, the tree adjoining grammars developed by Talaris associate with each natural language expression not only a syntactic tree but also a semantic representation. In addition, because these grammars are unification based, they can be used either to derive a semantic representation from a sentence (analysis) or to generate a sentence from a semantic representation (generation). We are actively exploring how the grammars we develop, can be used to support data-to-text generation. After having developed several sentence generation algorithms (GenI, RTGen and D-RTGen) [2], we are currently investigating: how to further optimise them; how to use them to verbalise knowledge bases and queries on knowledge bases; and how to evaluate their output.

[1] [46], [48], [47], [49], [50], [51], [52], [62], [59], [60], [61], [55], [58]
[2] [57], [54], [56], [53]

4.3. Hybrid Automated Deduction

TALARIS's main contribution in this topic has been the design of resolution and tableaux calculi for hybrid logics, calculi that were then implemented in the HyLoRes and HTab theorem provers. For example, TALARIS members have proved that the resolution calculus for hybrid logics can be enhanced with optimisations of order and selection functions without losing completeness. Moreover, a number of 'effective' (i.e., directly implementable) termination proofs for the hybrid logic $\mathcal{H}(@)$ has been established, for both resolution and tableaux based approaches, and the techniques are being extended to more expressive languages. Current work includes adding a temporal reasoning component to the provers, extending the architecture to allow querying against a background theory without having to explore again the theory with each new query, and testing the hybrid provers performance against dedicated state-of-the-art provers from other domains (firs-order logic, description logics) using suitable translations.

Moreover, we are interested in providing a range of inference services beyond satisfiability checking. For example, the current version of HyLoRes and HTab includes model generation (i.e., the provers can generate a model when the input formula is satisfiable).

We have also started to explore other decision methods (e.g., game based decision methods) which are useful for non-standard semantics like topological semantics. The prover HyLoBan is an example of this work.

4.4. Multimedia

MLIF (Multi Lingual Information Framework) is intended to be a generic ISO-based mechanism for representing and dealing with multilingual textual information. A preliminary version of MLIF has been associated with digital media within the ISO/IEC MPEG context and dealing with subtitling of video content, dialogue prompts, menus in interactive TV, and descriptive information for multimedia scenes. MLIF comprises a flexible specification platform for elementary multilingual units that may be either embedded in other types of multimedia content or used autonomously to localise existing content.

4.5. Interfacing Virtual Worlds and Natural Language Processing

In 2010, Talaris addressed a new application domain namely, the integration of deep natural language processing (NLP) techniques with 3D worlds and games. A first foray into that theme has been the submission of two systems to the international GIVE (Giving instructions in a virtual environment). Two recently accepted EU funded projects (Interreg project Allegro and Eurostar project Emo-Speech) on that theme will permit a fully blown exploration of the research issues and of the technological problems arising in this area. This new theme builds on the tools and techniques developped by Talaris over the last 5 years for deep NLP and in particular, on the availability of an expressive grammar writing environnement (XMG), of wide coverage deep grammars for French and English (SemTAG and SemXTAG), of a grammar based surface realiser (GenI) and of parsers (LLP2, SemConst) using these grammars.

5. Software

5.1. GenI

Participants: Claire Gardent [correspondent], Eric Kow [developer], Carlos Areces [developer].

GenI is a surface realiser that generates sentences from first order logical formulae. It is implemented in Haskell and uses the Glasgow Haskell compiler to obtain executable code for Windows, Solaris, Linux and Mac OS X. GenI is compatible with both a grammar for French (SemTag) and for English (SemXTag), both grammars beeing produced using the XMG MetaGrammar Compiler. SemTag covers the basic syntactic structures of French as described in Anne Abeillé's book "An Electronic Grammar for French". SemXTag has a coverage similar to that of XTAG, the TAG grammar for English developped by the University of Pennsylvannia . GenI is under GPL License. See also the web page http://talc.loria.fr/GenI-un-realisateur-de-surface.html.

- Version: 0.20.1

5.2. Web Service for the Multilingual-Assisted Chat Interface

Participant: Samuel Cruz-Lara [correspondent].

The Web Service for the Multilingual-Assisted Chat Interface program (WSMACI) is a linguistic assistant for virtual worlds. Its first version is dedicated to English assistance in such worlds. It has been developed in the context of the Metaverse1 project. It provide the end-users with MLIF-based provision of sentence analysis and word information (synonyms, definitions, translations) based on Google Translate, WordNet and the Brown Corpus.

- Version: 0.2

5.3. Emotion detection from textual information

Participant: Samuel Cruz-Lara [correspondent].

The 4 Layers Emotion Detection program (4LED) is an emotion detection tool. The emotions are extracted from texts in particular, from chat interfaces in virtual worlds. It has been developed in the context of the Metaverse1 project. The emotion detection process is based on SMILEY detection using WordNet-Domains and Tree-Tagger-based rules, WordNet-Affect, and keywords. http://talc.loria.fr/~metaverse/web_test/emotions/filterDetection/corpusCreation.php.

- Version: 0.2

5.4. Second Life Magic Carpet

Participant: Samuel Cruz-Lara [correspondent].

The Second Life Magic Carpet program (SLMC) is an assistant whose role is to guide people through virtual worlds with textual instructions. It has been developed in the context of the Metaverse1 project. It analyses the instructions of the visitors in order to find where they want to go, using web services for the analysis, for synonyms retrieving and for path finding.

- Version: 0.2

5.5. WikiAnalyzer

Participant: Alexandre Denis [correspondent].

The WikiAnalyzer is a tool developed in the CCCP-Prosodie project that aims to describe participants of Wikipedia projects. It provide a range of linguistic and structural analyses of Wikipedia discussion pages. The tool performs pages retrieval and automatic annotation of markers to build interactive profiles of participants. These profiles include information such as their level of expertise in the domain at hand, the use of subjective elements in their contributions, the connotation of the terms they use and enable to describe participants relative to their *degree of conflictuality* in the discussion. The structural analyses are parallel analyses on the structure of messages, enabling to categorize participants with regards to the type of contribution (starting a thread, participants they answer to, etc.). The tool has been developed in Java, and will be released as an online web application to the other members of the CCCP-Prosodie project.

- Version: 0.8

5.6. Emospeech Dialogue Toolkit

Participant: Lina-Maria Rojas Barahona [correspondent].

The Emospeech Dialogue Toolkit is a multi agent architecture for developing man/machine dialog systems in the context of a video game. It includes the following agents;

- Midiki Dialogue Manager: We extended and improved the open source MIDIKI (MITRE Dialogue Toolkit) software to support the multi-agent architecture and the configuration from a relational database.
- Wizard of Oz: We implemented two Wizard of OZ interfaces which allow a human to interact with other agents in the dialogue architecture. *The free-wizard* acts as a dialogue manager and permits a chat between two humans the player and the Wizard while simultaneously storing all interactions in a database. In contrast, *The semi-automatic wizard*, connects the Wizard with Midiki, whereby the Wizard interprets and adjusts Midiki generation.
- Answer Selection: We trained a classifier with Conditional Random Fields that chooses the most plausible response to a player utterance.

In addition, we trained a Logistic Regression Classifier for the interpretation agent that communicates with MIDIKI.

The dialogue agents communicate with the Game Agent, Speech Recognition and/or Chatbox agents developed by the Parole team. The Wizard of Oz, in which a human simulates a dialogue system, is used to collect dialogue data which can be used for training the interpreter and/or the Answer Selection Classifier. Moreover, a Dialogue Configuration Tool has been implemented for the configuration of several dialogues for different game scenarios by configuring the characters and goals in the game and the goals to be discussed in each dialogue. (See http://talc.loria.fr:8081/EmoDial).

- Version: 1.0

5.7. IGNG-Fv2

Participant: Jean-Charles Lamirel [correspondent].

The IGNG-Fv2 program implements a new incremental clustering algorithm whose main domain of application is the statistical analysis of continuous flow of evolving textual data, as well as the one of static textual data. It has been developed in the context of the CPER TALC (McFiiD action). It is based on a generic adaptation of the classical neural-based clustering approaches relying on gas of neurons with free topology. The IGNG-Fv2 approach exploits a combination of distance based and cluster data feature maximization criteria. This approach has been proved more efficient than the usual techniques for the analysis all kinds of static textual datasets. Considering its incremental character, it can also provide the information analysts with precise online detection of topic changes in the curse of a textual information flow.

5.8. C-Quality

Participant: Jean-Charles Lamirel [correspondent].

The C-Quality toolkit provides method-independent clustering quality measures and cluster labeling techniques specifically adapted to the interpretation of data analysis performed on textual data. The toolkit relies on an evaluation approach based on the exploitation of the maximized features of the data associated to each cluster after the clustering process without prior consideration of clusters profiles. The toolkit basic role is to act as an overall clustering quality evaluation tool. In a complementary way toolkit's clusters labeling functionalities can be used altogether for visualizing or synthesizing clustering results, for optimizing learning of a clustering method, for validating cluster content and act as efficient variable selection methods in the framework of supervised or semi-supervised learning tasks.

5.9. tl_dv2_ladl, a subcategorisation lexicon for French verbs.

Participant: Ingrid Falk [correspondent].

tl_dv2_ladl is a subcategorisation lexicon for French verbs produced by merging three lexicons which were built or validated manually: Dicovalence (version 2), TreeLex and the LADL tables. tl_dv2_ladl lists subcategorisation frames for 5918 French verbs. An entry in the lexicon consists of a verb and an associated subcategorisation frame whereby each subcategorisation frame describes a set of syntactic arguments with each argument being described by a grammatical function and a syntactic category. Each entry also gives the original lexical resource the information was extracted from. tl_dv2_ladl can be downloaded from http://talc.loria.fr/tl_dv2_ladl-a-subcategorisation.html.

- Version: 0.1

6. New Results

6.1. MLIF

TALARIS contributes to ISO TC 37 committee "Terminologies and other Language Resources", and more specifically to the activities of its SC3 "Computer Applications in Terminology", and SC4 "Linguistic Resources Management". Within TC37/SC4, TALARIS is currently contributing, as project leader, to the definition and specification of the Multi Lingual Information Framework (MLIF) [ISO FDIS 24616]. MLIF is being designed with the objective of providing a common abstract model being able to generate several formats used in the framework of translation and localization. MLIF has been released as FDIS (Final Draft International Standard) and it should finally be published as an official ISO standard soon. MLIF has been extensively used within the ITEA2 METAVERSE1 project. [42], [43], [12].

6.2. TEXT CLASSIFICATION

Neural clustering algorithms show high performance in the general context of the analysis of homogeneous textual datasets. We have recently proposed a new incremental growing neural gas algorithm using the cluster label maximization (IGNGF) [44] [34]. In this strategy the use of a standard distance measure for determining a winner is completely suppressed by considering the label maximization approach as the main winner selection process. One if its important advantage is that it provides the method with an efficient incremental character as it becomes independent of parameters. Although it performs better than the standard clustering methods on textual data, we have shown this year than the obtained results are not as efficient as expected whenever an analysis of very complex heterogeneous textual datasets is performed [33]. We have thus explored several variations of IGNG-F approach based on combination of distance based criteria and cluster label maximization. Our new results on all kinds of datasets, especially on the most complex heterogeneous textual datasets, clearly reflect the advantages of our new algorithm as compared to other existing algorithms and to our former adaptations [29]. Cluster quality evaluation represents a key process for all kinds of data analysis tasks, and more especially for textual data. We have recently presented different variations of unsupervised Recall/Precision and F-measures measures that cope with the defects of classical indexes, like inertia-based indexes. Our new indexes directly exploit the maximized features of the data associated to each cluster after the clustering process without prior consideration of clusters profiles. As compared to classical indexes, their main advantage is thus to be independent of the clustering methods and of their operating mode. They thus altogether permit the objective comparison of clustering methods and represent a sound technique for efficient cluster labeling. We have more especially worked this year on the large scale validation our indexes using reference labeled textual datasets [35].

We are also currently investigating to set up a platform for efficiently assisting the patents experts in the process of patents validation. Reaching such a goal has implied to develop new semi-supervised classification methods or propose in-deep adaptation of existing ones in order to establish relevant relationships between hierarchical patents classification and bibliographical references describing research covering the fields related to the different patents classes. In this context, we have successfully explored this year new classification techniques based on taboo search [14].

To cope with the current defects of existing incremental clustering methods, an alternative approach for analyzing information evolving over time consists in performing diachronic analysis. We have thus explored this year different an original technique based on this approach on texts by the use of the combination of cluster labeling with unsupervised Bayesian reasoning between cluster labels extracted from clustering model issued from different time periods. Based on a reference dataset issued from the IST-PROMTECH project, we have clearly shown that these new techniques, whilst providing a new framework for automatizing such kind of analysis, outperformed existing ones [32] [31] [30].

6.3. DIALOG

Within the Emospeech project, we developed the Emospeech Dialogue Toolkit (cf. Software section); used the Wizard of Oz infrastructure it includes to collect dialog data; and trained an interpreter and a dialog manager. The collected data comprises 591 dialogues in French collected within the context of the Mission Plastechnology serious game, 4874 utterances, 77854 words and 1321 player utterances containing 12901 word tokens and 1427 word types. We collected in average 50 conversations for each sub-dialogue in the game. Dialog length varies from 78 to 142 with an average length of 106 utterances per dialog.

6.4. GIVE

For the Generation of Instruction in Virtual Environment challenge edition 2.5 (GIVE), we developed two systems. The first system is the successor of the system that we presented to the GIVE 2 challenge (2010). We solved two known problems of this system, namely the indefinite presupposition problem and the ambiguity arising from underspecified referring expressions [26]. The GIVE 2.5 challenge proved that these improvements were efficient, and showed an increase of 21% in terms of task success (47% in GIVE 2, 68% in GIVE 2.5). The second system, developed in collaboration with the University of Cordoba is the first to our knowledge that uses a human-human corpus to provide whole utterances thanks to plan matching techniques [22], [21]. The system ranked fifth in terms of task success (58%), but second in terms of referring targets identification. The naturalness of instructions and the simplified system development makes it an interesting research track to follow.

6.5. Verb Classification

To help computer systems in the task of understanding and representing the full meaning of a text, verb classifications have been proposed which group together verbs with similar syntactic and semantic behaviour. For English verbs, VerbNet provides such a large scale classification but there are no similar French resource available. We investigated different ways both of automatically constructing such a resource; and of evaluating it.

Using Formal Concept Analysis (FCA), we developed a method for classifying verbs based on their (syntactic) subcategorisation information extracted from existing French lexical resources; and by translating the English Verbnet, we showed how to associate the obtained classes with semantic information represented by Verbnet's thematic role sets ([27]). As a result, a VerbNet like classification for French verbs can be constructed fully automatically.

The FCA approach we pursued, first builds a classification based on verbs and verb features and second filters this classification using various metrics (e.g., concept probability, concept stability). We are currently comparing this approach with a clustering approach which makes use of detailed evaluation metrics [44] and uses probabilistic information to guide classification. First results are promising and outperform the state of the art methods in this domain [63].

One important difference between the clustering and the FCA approach we experimented with is that only the second, allows a verb to belong to several classes. Since verbs are highly ambiguous, this is an important difference. To evaluate the impact of this difference on the usability of the classifications built by each of the methods, we are currently conducting a task-based, extrinsic evaluation of both classifications by analysing their impact when used in a Semantic Role Labeling task on a French corpus.

6.6. I-FLEG

Within the Allegro project, we developed the I-FLEG game [16], [15], a virtual world in which the learner exercises French by clicking on objects and answering the questions raised by the system. The language learning exercises produced by I-FLEG are automatically generated using the GenI sentence generator from a knowledge base describing the virtual world. A preliminary evaluation of I-FLEG with school children [17] suggests that the "game" aspect increases learner motivation and that spoken output is essential in maintaining learner interest.

7. Partnerships and Cooperations

7.1. McFIID

Program: CPER MISN/TALC

Project acronym: McFIID

Project title: Clustering; Statistical Analysis; Textual data; Time-evolving data; Distributed data

Duration: 2007-01-01 / 2011-12-31

Coordinator: Jean-Charles Lamirel

Other partners: INIST

Abstract: The McFIID project is a CPER project continuing the CPER CLASSIF project. It concerns the development of incremental multi-clustering techniques for managing distributed and evolving flows of textual data. New approach of diachronic analysis based on the use of multiple viewpoints combined with unsupervised bayesian reasoning, as well as new online incremental clustering techniques based on non standard similarity measures, are tested in the curse of these project.

7.2. National Initiatives

7.2.1. CCCP-Prosodie

Program: ANR CONTINT

Project acronym: CCCP Prosodie

Project title:

Duration: 2008-01-12 / 2011-31-06

Coordinator:

Other partners: Institut Télécom, UTC Compiégne, UNSA (Univ. Nice Sophia-Antipolis), Univ. de Versailles St-Quentin

Abstract: The goal of CCCP-Prosodie is to empirically investigate the functioning of online communities such as Wikipedia, and particular to link their activities and their use of language (as recorded in such corpora as email exchanges, for example). The TALARIS team is involved in this project for three reasons: to provide Natural language processing tools, to design an annotation scheme capable of dealing with information from both the social sciences (sociology and economics) and the humanities (psychology and ergonomics), and to provide help with inference technology.

See also: http://recherche.telecom-bretagne.eu/labo_communicant/cccp-prosodie/

7.2.2. PORT-MEDIA

Program: ANR CONTINT

Project acronym: PORT-MEDIA

Project title:

Duration: 2009-03-01 / 2012-03-01

Coordinator:

Other partners: ELDA, LIG/GETALP, LIA, LIUM, LORIA

Abstract: The PORT-MEDIA project is an ANR project that aims to collect linguistic data for multiple domains and to investigate the use of a high-level semantic representation for annotating dialogue corpora. TALARIS contributed to the high-level semantics specification for annotating the MEDIA corpus and to the development of tools for the manual annotation (e.g., ATOOL and SRL-Web Annotation) as well as to the development of the blackboard architecture for the automatic annotation of the MEDIA corpus. Additionnally, Talaris provided the automatic annotation of the whole corpus and its evaluation.

See also: http://www.port-media.org/doku.php?id=start

7.2.3. SYFRAP

Program: PEPS INS2I-INSHS

Project acronym: SYFRAP

Project title: Analyse syntaxique du français parlé

Duration: 2011-06-01 / 2013-06-01

Coordinator: Claire Gardent, LORIA

Other partners: ATILF Nancy, LLF Paris 7

Abstract: SYFRA is a exploratory interdisciplinary project (PEPS INS2I-INSHS) funded by the CNRS. It gathers researchers from LORIA (Nancy), LLF (Paris 7) and from ATILF (Nancy); and aims to develop resources (annotated corpora) and tools for the syntactic analysis of spoken French.

See also: http://talc.loria.fr/-SYFRAP,71-.html

7.3. Collaborations in European Programs, except FP7

7.3.1. Allegro

Program: INTERREG IV A

Project acronym: Allegro

Project title:

Duration: 01.2010 - 12.2012

Coordinator: U. Saarbrücken (Germany)

Other partners: Supelec Metz, INRIA Nancy Grand Est

Abstract: The Allegro project aims to develop NLP techniques that support language teaching for French and German.

7.3.2. Emospeech

Program: Eurostars

Project acronym: Emospeech

Project title:

Duration: 2010-09-01 / 2013-08-31

Coordinator: Artefacto, Rennes

Other partners: Acapella, INRIA Nancy Grand Est

Abstract: The EMOSPEECH project aims to augment serious games with natural language (spoken and written dialog) and emotional abilities (gesture, intonation, facial expressions).

7.3.3. Metaverse

Program: ITEA2

Project acronym: Metaverse

Project title:

Duration: 2009-01-01 / 2011-04-31

Coordinator:

Other partners: Belgian partners: Alcatel-Lucent Bell N.V., Nazooka, IBBT-SMIT; French partners: Alcatel-Lucent France, Orange Labs, CEA List, Artefacto; Greek partners: Forthnew S.A., Ellinogermanki Agogi; Dutch partners: Philips Research, Philips I-Lab, DevLab, Technical University Eindhoven, University of Twente, Stg. EPN, VU Economics & BA, VU CAMeRA; Spanish partners: Innovalia, Ceeda, VirtualWare, CBT, Nextel, Corsa, Avantalia, I&IMS, VicomTECH, E-PYME, CIC Tour Game, UPF-MTH; Israeli partners: Metaverse Labs.

Abstract: Metaverse is an exciting project whose goal is to provide a standardized global framework enabling the interoperability between virtual worlds (for example Second Life, World of Warcraft, IMVU, Active Worlds, Google Earth and many others) and the Real world (sensors, actuators, vision and rendering, social and welfare systems, banking, insurance, travel, real estate and many others).

7.4. International Initiatives

7.4.1. INRIA Associate Teams

7.4.1.1. INTOHYLO

Title: Inference Tools for Hybrid Logics and Applications for Natural Language Processing

INRIA principal investigator: Carlos Areces

International Partner:

Institution: Universidad de Buenos Aires (Argentina)

Laboratory: Universidad de Buenos Aires, GLyC

Duration: 2009 - 2011

See also: http://led.loria.fr/dokuwiki/doku.php?id=intohylo_-_inria_equipes_associees

The main aim of the InToHyLo project is to investigate inference methods for hybrid logics, to develop highly optimized inference tools based on these methods, and to use these tools in natural language applications. Talaris and GLyC are currently leaders in automated theorem proving for hybrid logics, and they are the developers of the two provers HyLoRes (based on resolution) and HTab (based on tableaux). With the InToHyLo project we want to investigate how to combine resolution and tableaux algorithms to allow our provers to collaborate and share partial results. We will integrate our tools in a platform suitable for inference in NLP applications (focusing on Dialogue Systems and Textual Entailment). This platform will include not only tools for satisfiability testing, but also for model building, model checking, bisimulation checking, and knowledge maintenance and retrieval. Finally, we want to develop parallel inference algorithms to improve performance, and distributed testing to speed up developing.

7.4.2. Visits of International Scientists

- Kristina Striegnitz, Union College, Schenectady, NY, 1 week in January 2011

- Eva Banik, Computational Linguistics Ltd, 1 week in May 2011

7.4.3. Participation In International Programs

7.4.3.1. GIVE challenge organisation

Talaris co-organized the 2.5 edition of the Generation of Instructions in Virtual Environment challenge. This challenge brought together six universities or laboratories working on natural language generation: University of Aberdeen, University of Bremen, University of Cordoba, University of Postdam, University of Twente, and the LORIA. The challenge was available online for players to test the different systems. Eight systems were participating to the campaign. Over two months, we collected 536 games, which is lower than last year. We assume that the summer break which coincided with the challenge played a role. Our participation in the organisation of the campaign involved rewriting the network layer, a complete change of the visibility algorithm, and advertising the challenge.

8. Dissemination

8.1. Animation of the scientific community

- Nadia Bellalem
  - Member of the International Program Committee of ICEIS 2011 (http://www.iceis.org).
  - Gestion des stages au département Informatique de l'IUT Nancy-Charlemagne, Univ. Nancy2
- Samuel Cruz-Lara
  - Project leader of MLIF FDIS 24616 ISO TC37 / SC4 / WG3 *Language Resources Management*
  - Member of the SYnchronized Multimedia group (SYMM) of he World Wide Web Consortium
  - Co-editor of the Journal of Virtual Worlds Research, Volume 4, Number 3. Theme: MPEG-V and Other Virtual Worlds Standards
  - Member of the scientific committee of IEEE RITA *Revista Iberoamericana de Tecnologías del Aprendizaje*
  - French coordinator for the Computer Science area of the MEXPROTEC program (France - Mexico)
  - Invited talk at *3D3C Worlds Meeting*, March 2011, Shefayim Kibbutz, Israel
  - Invited talk at *International Conference on Engineering and Computer Education, New Trends in Engineering Education Workshop* September 2011, Guimarães, Portugal
- Alexandre Denis
  - PC member for the *13th European Workshop on Natural Language Generation* (ENLG), Nancy, France.
- Christine Fay-Varnier
  - Second Vice president of the Council of studies and university life at the INPL Nancy
  - Manager of the service for Information and Communication Technology for Education of Nancy University (NUTICE)
  - Representative of Nancy University in 6 of the French thematic virtual university (UNT : http://www.universites-numeriques.fr/fr)

  - Organizer of 7th Conference Information and Communication Tools for learning and training Nancy, December 6 - 8 2010 (http://www.tice2010.nancy-universite.fr)
  - NOOJ Software tutoring at the NOOJ seminar
  - Manager of computer science course, 1st and 2nd year in National School of Geology

- Claire Gardent
  - Invited Talk at *Traitement Automatique des Langues Naturelles (TALN)*, June 2011, Montpellier (France)
  - Invited Tutorial at *Knowledge Capture (K-CAP)*, June 2011, Banff (Canada).
  - Invited Seminar at the *KRDB Research Center for Knowledge and Data*, November 2011, Bolzano (Italy).
  - PC Chair for the *13th European Workshop on Natural Language Generation* (ENLG), Nancy, France.
  - Co-editor of the Computational and Mathematical section for the Language and Linguistics Compass journal
  - PC member *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (ACL-HLT), Porland, Oregon.
  - PC member *TextInfer 2011 Workshop on Textual Entailment*, Ediburgh, Scotland.
  - PC member *Production of Referring Expressions* (PRe-CogSci), Boston, Massachusetts.
  - Member of the LORIA steering committee
  - Member of the Bureau de Formation Doctorale
  - Local organiser for NaTAL 2011
  - Local organiser for the *13th European Workshop on Natural Language Generation* (ENLG), Nancy, France.
  - Coordinator of the TALC theme (Computational Linguistics and Computational Approaches to Knowledge) for the MISN CPER (National and Regional Research Funding).
  - Organiser of the LORIA TALC seminar http://talc.loria.fr/NATAL-10-16-18-juin-2010-LORIA,178.html
  - Local organiser for the NaTAL 2011 workshop http://talc.loria.fr/Programme,251.html

- Jean Charles Lamirel
  - Member of editorial board of the new international journal "COLLNET Journal of Scientometrics and Information Management", Taru Publications, New Delhi, India
  - Member of the Algerian-French Doctoral School in Linguistics and Didactics (EDAF)
  - Permanent member of the COLLNET scientific network
  - Scientific supervisor in the framework of the QUAERO project for the intelligent patents management subtask
  - Program chair of the 12th COLLNET meeting (Global interdisciplinary research network "Col"laboration in Science and in Technology)
  - Member of the Program committee for the Twenty-fourth International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE 2011)
  - Member of the Program committee for the 23rd IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2011)

  - Co-organizer of the Workshop “Clustering incrémental et méthodes de détection de nouveauté et leur application à l’analyse intelligente d’information évoluant au cours du temps”, organisé à la 12ème conférence francophone "Extraction et Gestion de Connaissances" (EGC 2012), Bordeaux, France, January 2012
  - Organizer of Special Session on Incremental clustering and novelty detection techniques and their application to intelligent analysis of time varying information in the framework of IEA/IAE International Conference, Syracuse, NY, USA, June 2011

- Lina-Maria Rojas Barahona
  - Member of *The Young Researchers’ Roundtable on Spoken Dialogue Systems* (YRRSDS), Organizing Committee. Portland Oregon 2011. http://www.yrrsds.org/

8.2. Teaching

Licence: Découverte de l’informatique, 160 h, L1, UHP Nancy, France

Licence: Conception et développement, 120h, L1/L2, Ecole Superieur D’informatique Appliquée De Lorraine (ESIAL), Nancy, France

Licence: Conception de systèmes d’informations, 120h, Ecole Superieur D’informatique Appliquée De Lorraine (ESIAL), Nancy, France

Licence: Système de Gestion de Base de données, 70 h, L1, IUT Nancy-Charlemagne, Nancy, France.

Licence: Java Language Programming. 60 hours, DUT Informatique, IUT Nancy-Charlemagne, University of Lorraine, France.

Licence: C Language Programming. 30 hours, DUT Informatique, IUT Nancy-Charlemagne, University of Lorraine, France.

Licence: Web-based Mobile Applications (HTML5 and CSS3). 30 hours, LP CISII, IUT Nancy-Charlemagne, University of Lorraine, France.

Master: Cognitive Sciences and Digital Media Technologies. 22 hours, Cognitive Sciences, University of Lorraine, France.

Master: Declarative Languages and Multimedia Applications. 40 hours, Cognitive Sciences, University of Lorraine, France.

Master: Video: Streaming and Captioning. 12 hours, Cognitive Sciences, University of Lorraine, France.

Master: An Introduction to Virtual Worlds and Natural Language Processing. 12 hours, Cognitive Sciences, University of Lorraine, France.

Master: Données semi-structurées, 30 h, M1, UHP Nancy, France.

Master: Text Mining techniques applied to Linguistics, 20 h, M2, University of Alger, Algeria

Research project tutoring, Erasmus Mundus Master Computer of sciences UHP

Master: Software Project for NLP Applications, 10 hours, M2, Université Nancy 2, France.

Master: Conférence de Rentrée, 3 hours, M1 et M2, ENS Cachan.

PhD in progress : Ingrid Falk, “Lexicon and Ontology”. PhD in Computer Science. Université de Nancy 2 (France). From 15.10.2008. Supervisors: Claire Gardent and Samuel Cruz-Lara

PhD in progress : Laura Perez-Beltrachini. “Large scale surface realisation”. Université Henri Poincaré, Nancy. From 15.10.2009. Supervisor: Claire Gardent

PhD in progress : Alejandra Lorenzo, “Improving interpretation robustness and increasing system adaptivity in dialogue based, language learning systems”. Université Henri Poincaré, Nancy. From 15.10.2010. Supervisors: Claire Gardent and Christophe Cerisara

PhD in progress : Shashi Narayan, “Adaptive Generation”. Université Henri Poincaré, Nancy. From 15.10.2011. Supervisor: Claire Gardent.

9. Bibliography

Major publications by the team in recent years

[1] P. BLACKBURN, J. VAN BENTHEM, F. WOLTER (editors). *Handbook of Modal Logic*, Elsevier, 2007.

[2] C. ARECES, S. FIGUEIRA. *Which Semantics for Neighbourhood Semantics?*, in "Proceedigns of IJCAI 09", Pasadena, California, USA, 2009, p. 671–676.

[3] C. ARECES, B. TEN CATE. *Hybrid Logics*, in "Handbook of Modal Logics", P. BLACKBURN, F. WOLTER, J. VAN BENTHEM (editors), Elsevier, 2006.

[4] L. BENOTTI. *Incomplete Knowledge and Tacit Action: Enlightened Update in a Dialogue Game*, in "DECALOG 2007 Workshop on the Semantics and Pragmatics of Dialogue", Rovereto, Italy, 2007.

[5] P. BLACKBURN, B. TEN CATE. *Pure Extensions, Proof Rules, and Hybrid Axiomatics*, in "Studia Logica", 2006, n^{o} 84, p. 277–322.

[6] T. BOLANDER, P. BLACKBURN. *Termination for Hybrid Tableaus*, in "Journal of Logic and Computation", 2007, n^{o} 17, p. 517–554.

[7] D. C. A. BULTERMAN, A. J. JANSEN, P. CESAR, S. CRUZ-LARA. *An efficient, streamable text format for multimedia captions and subtitles*, in "DocEng '07: Proceedings of the 2007 ACM symposium on Document engineering", New York, NY, USA, ACM, 2007, p. 101–110, http://doi.acm.org/10.1145/1284420.1284451.

[8] A. DENIS, G. PITEL, M. QUIGNARD, P. BLACKBURN. *Incorporating Asymmetric and Asychronous Evidence of Understanding in a Grounding Model*, in "DECALOG 2007 Workshop on the Semantics and Pragmatics of Dialogue", Rovereto, Italy, 2007.

[9] C. GARDENT, E. KOW. *A symbolic approach to near-deterministic surface realisation using Tree Adjoining Grammar*, in "Proceedings of ACL", Prague, 2007.

[10] C. GARDENT, H. MANUÉLIAN. *Création d'un corpus annoté pour le traitement des descriptions définies*, in "Traitement Automatique des Langues", 2005, vol. 46, n^{o} 1.

[11] C. GARDENT, K. STRIEGNITZ. *Generating Bridging Definite Descriptions*, in "Computing Meaning", H. BUNT, R. MUSKENS (editors), Studies in Linguistics and Philosophy Series, Kluwer Academic Publishers, 2007, vol. 3.

Publications of the year

Articles in International Peer-Reviewed Journal

[12] S. CRUZ-LARA, T. OSSWALD, J. GUINAUD, N. BELLALEM, L. BELLALEM. *A Chat Interface Using Standards for Communication and E-learning in Virtual Worlds*, in "Lecture Notes in Business Information Processing", 2011, vol. 73, n° Part 6, p. 541-554 [*DOI :* 10.1007/978-3-642-19802-1_37], http://hal.inria.fr/inria-00580198/en.

[13] C. GARDENT, B. GOTTESMAN, L. PEREZ-BELTRACHINI. *Using Regular Tree Grammars to enhance Sentence Realisation*, in "Natural Language Engineering", January 2011, vol. 17, n° 2, p. 185-201 [*DOI :* 10.1017/S1351324911000076], http://hal.inria.fr/inria-00537219/en.

[14] M. ZENNAKI, A. ECH-CHERIF, J.-C. LAMIREL. *Learning Solution Quality of Hard Combinatorial Problems by Kernel Methods*, in "International Journal of Computer Applications (IJCA)", August 2011, http://hal.inria.fr/hal-00645400/en.

International Conferences with Proceedings

[15] M. AMOIA. *I-FLEG: A 3D-Game for Learning French.*, in "The 9th International Conference on Education and Information Systems, Technologies and Applications - EISTA 2011", Orlando, États-Unis, July 2011, http://hal.inria.fr/hal-00643959/en.

[16] M. AMOIA, C. GARDENT, L. PEREZ-BELTRACHINI. *A Serious Game for Second Language Acquisition*, in "Third International Conference on Computer Supported Education - CSEDU 2011", Noordwijkerout, Pays-Bas, A. VERBRAECK, M. HELFERT, J. CORDEIRO, B. SHISHKOV (editors), SciTePress, April 2011, p. 394–397, ISBN : 978-989-8425-49-2, http://hal.inria.fr/hal-00643955/en.

[17] M. AMOIA, C. GARDENT, L. PEREZ-BELTRACHINI. *Learning a Second Language with a Videogame*, in "ICT for Language Learning", Firenze, Italie, Pixel Associazione, October 2011, http://hal.inria.fr/hal-00643953/en.

[18] C. ANDERSON, C. CERISARA, C. GARDENT. *Vers la détection des dislocations à gauche dans les transcriptions automatiques du Français parlé / Towards automatic recognition of left dislocation in transcriptions of Spoken French*, in "Traitement Automatique des Langues Naturelles - TALN'2011", Montpellier, France, June 2011, 6, http://hal.inria.fr/hal-00600510/en.

[19] C. ARECES, P. FONTAINE. *Combining theories: the Ackerman and Guarded Fragments*, in "8th International Symposium Frontiers of Combining Systems - FroCoS 2011", Saarbrücken, Allemagne, C. TINELLI, V. SOFRONIE-STOKKERMANS (editors), Lecture Notes in Computer Science, Springer Verlag, 2011, vol. 6989, p. 40–54 [*DOI :* 10.1007/978-3-642-24364-6_4], http://hal.inria.fr/hal-00642529/en.

[20] P. BEDARIDE, C. GARDENT. *Deep Semantics for Dependency Structures*, in "12th International Conference on Computational Linguistics and Intelligent Text Processing - CICLing 2011", Tokyo, Japon, A. F. GELBUKH (editor), Lecture Notes in Computer Science, Springer Verlag, February 2011, vol. 6608, p. 277–288, The original publication is available at www.springerlink.com [*DOI :* 10.1007/978-3-642-19400-9_22], http://hal.inria.fr/hal-00639825/en.

[21] L. BENOTTI, A. DENIS. *CL system: Giving instructions by corpus based selection*, in "13th European Workshop on Natural Language Generation", Nancy, France, September 2011, http://hal.inria.fr/inria-00636474/en.

[22] L. BENOTTI, A. DENIS. *Giving instructions in virtual environments by corpus based selection*, in "SIGdial Meeting on Discourse and Dialogue", Portland, USA, June 2011, http://hal.inria.fr/inria-00636303/en.

[23] L. BENOTTI, A. DENIS. *Prototyping virtual instructors from human-human corpora*, in "Association for Computational Linguistics: Human Language Technologies", Portland, USA, June 2011, http://hal.inria.fr/inria-00636300/en.

[24] C. CERISARA, C. GARDENT. *The JSafran platform for semi-automatic speech processing*, in "12th Annual Conference of the International Speech Communication Association - Interspeech 2011", Florence, Italie, August 2011, 4, http://hal.inria.fr/hal-00600520/en.

[25] C. CERISARA, P. KRAL, C. GARDENT. *Commas recovery with syntactic features in French and in Czech*, in "12thAnnual Conference of the International Speech Communication Association - Interspeech 2011", Florence, Italie, August 2011, 4, http://hal.inria.fr/hal-00600528/en.

[26] A. DENIS. *The Loria Instruction Generation System L in GIVE 2.5*, in "13th European Workshop on Natural Language Generation", Nancy, France, September 2011, http://hal.inria.fr/inria-00636479/en.

[27] I. FALK, C. GARDENT. *Combining Formal Concept Analysis and Translation to Assign Frames and Thematic Role Sets to French Verbs*, in "Concept Lattices and Their Applications", Nancy, France, A. NAPOLI, V. VYCHODIL (editors), October 2011, http://hal.inria.fr/inria-00634270/en.

[28] C. GARDENT, Y. PARMENTIER, G. PERRIER, S. SCHMITZ. *Lexical Disambiguation in LTAG using Left Context*, in "5th Language & Technology Conference - LTC'11", Poznan, Pologne, November 2011, 4, http://hal.inria.fr/hal-00629902/en.

[29] J.-C. LAMIREL, S. AL SHEHABI, G. SAFI. *A new label maximization based incremental neural clustering approach: application to text clustering*, in "8th Workshop on Self-Organizing Maps (WSOM 2011)", Espoo, Finlande, June 2011, http://hal.inria.fr/hal-00645394/en.

[30] J.-C. LAMIREL. *A new diachronic methodology for automatizing the analysis of research topics dynamics : an example of application on optoelectronics research*, in "7th International Conference on Webometrics, Informetrics and Scientometrics and 12th COLLNET Meeting", Istanbul, Turquie, September 2011, http://hal.inria.fr/hal-00645389/en.

[31] J.-C. LAMIREL. *A new method for automatically analyzing the research dynamics: application on optoelectronics research*, in "13th ISSI 2011 Conference", Durban, Afrique Du Sud, July 2011, http://hal.inria.fr/hal-00645391/en.

[32] J.-C. LAMIREL. *Une nouvelle méthodologie diachronique pour automatiser l'analyse de la dynamique des domaines de recherche*, in "1st. International Symposium ISKO-Maghreb", Hammamet, Tunisie, May 2011, http://hal.inria.fr/hal-00645397/en.

[33] J.-C. LAMIREL, R. MALL, M. AHMAD. *Comparative behaviour of recent incremental and non-incremental clustering methods on text: an extended study*, in "The Twenty-fourth International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE 2011)", Syracuse, USA, June 2011, http://hal.inria.fr/hal-00645393/en.

[34] J.-C. LAMIREL, R. MALL, M. AHMAD. *Comportement comparatif des méthodes de clustering incrémentales et non incrémentales sur les données textuelles hétérogènes*, in "11th International Francophone Conference on Knowledge Extraction and Management (EGC 2011)", Brest, France, January 2011, http://hal.inria.fr/hal-00645398/en.

[35] J.-C. LAMIREL, R. MALL, P. CUXAC, G. SAFI. *A new efficient and unbiased approach for clustering quality evaluation*, in "PAKDD 2010 2nd International Workshop on Quality Issues, Measures of Interestingness and Evaluation of Data Mining Models (QIMIE)", Shenzen, Chine, May 2011, http://hal.inria.fr/hal-00645395/en.

[36] J.-C. LAMIREL, R. MALL, P. CUXAC, G. SAFI. *Variations to incremental growing neural gas algorithm based on label maximization*, in "International Joint Conference on Neural Networks", San Jose, USA, July 2011, http://hal.inria.fr/hal-00645390/en.

[37] J.-C. LAMIREL, R. MALL, P. CUXAC, G. SAFI. *Variations to incremental growing neural gas algorithm based on label maximization*, in "International joint conference on neural networks - IJCNN 2011", San Jose, États-Unis, 2011, http://hal.inria.fr/inria-00624191/en.

[38] L. M. ROJAS BARAHONA, T. BAZILLON, M. QUIGNARD, F. LEFEVRE. *Using MMIL for the High Level Semantic Annotation of the French MEDIA Dialogue Corpus*, in "Ninth International Conference on Computational Semantics - IWCS 2011", London, Royaume-Uni, J. BOS, S. PULMAN (editors), ACL, January 2011, http://hal.inria.fr/inria-00638000/en.

[39] L. M. ROJAS BARAHONA, M. QUIGNARD. *An Incremental Architecture for the Semantic Annotation of Dialogue Corpora with High-Level Structures. A case study for the MEDIA corpus*, in "12th annual SIGdial Meeting on Discourse and Dialogue - SIGDIAL 2011", Portland Oregon, USA, J. MOORE, D. TRAUM (editors), ACL, 2011, p. 332-334, http://hal.inria.fr/inria-00637991/en.

[40] K. STRIEGNITZ, A. DENIS, A. GARGETT, K. GAROUFI, A. KOLLER, M. THEUNE. *Report on the Second Second Challenge on Generating Instructions in Virtual Environments (GIVE-2.5)*, in "13th European Workshop on Natural Language Generation", Nancy, France, September 2011, http://hal.inria.fr/inria-00636498/en.

Scientific Books (or Scientific Book chapters)

[41] P. BEDARIDE, C. GARDENT. *Non Compositional Semantics Using Rewriting*, in "Human Language Technology. Challenges for Computer Science and Linguistics", Z. VETULANI (editor), Lecture Notes in Computer Science, Springer Verlag, February 2011, vol. 6562, p. 257-267 [*DOI :* 10.1007/978-3-642-20095-3_24], http://hal.inria.fr/hal-00639843/en.

[42] S. CRUZ-LARA, J. M. CABELLO, T. OSSWALD, A. COLLADO, J. M. FRANCO, S. BARRERA. *Tourism in Virtual Worlds: Means, Goals and Needs*, in "Handbook of Research on Practices and Outcomes in Virtual Worlds", H. H. YANG, S. C.-Y. YUEN (editors), IGI Globals, 2011, http://hal.inria.fr/inria-00580241/en.

[43] S. CRUZ-LARA, T. OSSWALD, J.-P. CAMAL, N. BELLALEM, L. BELLALEM, J. GUINAUD. *Enabling Multilingual Social Interactions and Fostering Language Learning in Virtual Worlds*, in "Handbook of Research on Practices and Outcomes in Virtual Worlds", H. H. YANG, S. C.-Y. YUEN (editors), IGI Globals, 2011, http://hal.inria.fr/inria-00580219/en.

[44] J.-C. LAMIREL, P. CUXAC, C. FRANÇOIS. *Recherche des évolutions technologiques à partir de bases de données bibliographiques : apport de la classification incrémentale*, in "Technologies de la Connaissance

et Recherche d'Information en Contexte", S. COLLECTION (editor), Hermes Science Publishing Ltd, 2011, http://hal.inria.fr/inria-00535972/en.

Other Publications

[45] G. JACQUET, J.-L. MANGUIN, F. VENANT, B. VICTORRI. *Déterminer le sens d'un verbe dans son cadre prédicatif*, JACQUET G., MANGUIN J.L., VENANT F., VICTORRI B., Construire le sens d'un verbe dans son cadre prédicatif, in F. Neveu, P. Blumenthal, N. Le Querler (Eds.), Au commencement était le verbe. Syntaxe, Sémantique et Cognition, Peter Lang, 2011, à paraître, http://hal.inria.fr/hal-00611665/en.

References in notes

[46] B. CRABBÉ. *Alternations, monotonicity and the lexicon: an application to factorising information in a Tree Adjoining Grammar*, in "Proc. of the European Summer School of Logic Language and Computation, Student Session - ESSLLI'03, Vienna, Austria", Aug 2003, p. 69-80.

[47] B. CRABBÉ. *Lexical Classes for structuring the lexicon of a TAG*, in "Lorraine-Saarland Workshop series: prospects and advances in the syntax semantics interface, Nancy, France", Oct 2003.

[48] B. CRABBÉ, B. GAIFFE, A. ROUSSANALY. *Reprsentation et gestion de grammaires d'arbres adjoints lexicalises*, in "Traitement Automatique des Langues", Dec 2003, vol. 44, n^{o} 3, p. 67-91.

[49] B. CRABBÉ, B. GAIFFE, A. ROUSSANALY. *Une plateforme de conception et d'exploitation de grammaire d'arbres adjoints lexicaliss*, in "Traitement Automatique du Langage Naturel 2003 - TALN 2003, Batz-sur-Mer, France", Jun 2003.

[50] B. CRABBÉ, D. DUCHIER. *Metagrammar Redux*, in "Proc. of the International Workshop on Constraint Solving and Language Processing - CSLP 2004, Copenhagen, Norway", Sep 2004.

[51] D. DUCHIER, J. LE ROUX, Y. PARMENTIER. *The MetaGrammar Compiler: An NLP Application with a Multi-paradigm Architecture*, in "Proc. of the 2nd International Mozart/Oz Conference - MOZ 2004, Charleroi, Belgium", Oct 2004.

[52] B. GAIFFE, B. CRABBÉ, A. ROUSSANALY. *A New Metagrammar Compiler*, in "Proc. of the 6th International Workshop on Tree Adjoining Grammars and Related Frameworks - TAG+6, Venice, Italy", May 2002.

[53] C. GARDENT, B. GOTTESMAN, L. PEREZ-BELTRACHINI. *Using Regular Tree Grammar to enhance Surface Realisation*, in "Natural Language Engineering", 2011, vol. 17, p. 185 - 201, Special Issue on Finite State Methods and Models in Natural Language Processing.

[54] C. GARDENT, E. KOW. *Generating and selecting paraphrases*, in "Proc. of the 10th European Workshop on Natural Language Generation - ENLG 05, Aberdeen, Scotland", Aug 2005, p. 191-196.

[55] C. GARDENT, Y. PARMENTIER. *Large scale semantic construction for Tree Adjoining Grammars*, in "Proc. of Logical Aspects of Computational Linguistics - LACL'05, Bordeaux, France", P. BLACHE, E. STABLER, J. BUSQUETS, R. MOOT (editors), Lecture Notes in Computer Science, Springer, Apr 2005, vol. 3492, p. 131-146.

[56] C. GARDENT, L. PEREZ-BELTRACHINI. *RTG-based Surface Realisation for TAG*, in "COLING 2010", Beijing, China, August 2010, p. 367–375.

[57] E. KOW. *Adapting polarised disambiguation to surface realisation*, in "Proc. of the 17th European Summer School in Logic, Language and Information - ESSLLI'05, Edinburgh, United Kingdom", Aug 2005.

[58] Y. PARMENTIER, J. LE ROUX. *XMG: a Multi-formalism Metagrammatical Framework*, in "Proc. of the 17th European Summer School in Logic, Language and Information - ESSLLI '05, Edinburgh, United Kingdom", Aug 2005.

[59] D. SEDDAH, B. GAIFFE. *Des arbres de dérivation aux forêts de dépendance : un chemin via les forêts partagées*, in "Traitement automatique des langues Naturelles - TALN'05, Dourdan, France", Jun 2005.

[60] D. SEDDAH, B. GAIFFE. *How to Build Argumental graphs Using TAG Shared Forest: a view from control verbs problematic*, in "Proc. of the 5th International Conference on the Logical Aspect of Computional Linguistic - LACL'05, Bordeaux, France", Apr 2005.

[61] D. SEDDAH, B. GAIFFE. *Using both Derivation tree and Derived tree to get dependency graph in derivation forest*, in "Proc. of the 6th International Workshop on Computational Semantics - IWCS-6, Tilburg, The Netherlands", Jan 2005.

[62] D. SEDDAH. *Synchronisation des connaissances syntaxiques et sémantiques pour l'analyse d'énoncés en langage naturel à l'aide des grammaires d'arbres adjoints lexicalisées*, Université Henri Poincaré - Nancy 1, Nov 2004.

[63] L. SUN, A. KORHONEN, T. POIBEAU, C. MESSIANT. *Investigating the cross-linguistic potential of verbnet: style classification*, in "Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10)", USA, Association for Computational Linguistics, 2010, p. 1056–1064.