Projets de recherche

Dernière mise à jour : 20-06-2014

PROJETS EN COURS

LIMAH

Comin Labs,
2014-2017

Partenaires : TEXMEX INRIA Rennes (Coordinateur), Le Laboratoire Informatique de Nantes-Atlantique, Telecom Bretagne, Univ. Rennes 2.

Objectifs :

TERMITH

ANR Contint,
2012-2014 www.atilf.fr/ressources/termith

Partenaires : ATILF (Coordinateur), Le Laboratoire Informatique de Nantes-Atlantique, INIST, LILIDEM, INRIA Nancy et Saclay.

Objectifs : TermITH s'intéresse à l'accès à l'information des documents numériques par le biais d'une indexation fondée sur les termes qu'ils contiennent, ce qui suppose reconnaissance, désambiguïsation et analyse des termes.
Sur le plan expérimental, TermITH s'intéresse à un champ scientifique très ambigu entre langue terminologique de spécialité et langue générale : les sciences humaines et sociales. La méthodologie, sera validée sur l'archéologie, la psychologie (psychanalyse, psychologie sociale et sciences cognitives), les sciences de l'information, et la chimie.

CRISTAL

ANR Contint,
2012-2014 www.projet-cristal.org

Partenaires : Le Laboratoire Informatique de Nantes-Atlantique (Coordinateur), ERSS, Université de Genève, Lingua et Machina.

Objectifs : L'objectif du projet CRISTAL est de développer une technologie d'extraction de contextes riches en connaissances (CRC) permettant de produire de nouveaux dictionnaires. Ces derniers présenteront, pour chaque terme et ses traductions éventuelles, une fiche terminologique listant les CRC et explicitant les connaissances qu'ils contiennent. L'extraction de tels contextes nécessite la conception d'algorithmes avancés d'extraction de connaissances, capables d'analyser automatiquement et de façon robuste le contenu linguistique et sémantique de textes bruts et non-structurés et ce, quelle qu'en soit la langue.

ANCIENS PROJETS TTC- Terminology Extraction, Translation tools and Comparable corpora
FP7 - Information Society and Media - ICT 2009.2.2: Language-based interaction

2010-2012 en savoir plus

Le
projet TTC vise à exploiter les possibilités offertes par les corpus
comparables pour améliorer les performances des outils informatiques de
traduction sur des domaines techniques et dans un contexte massivement

multilingue où il est nécessaire de traduire un même document dans plusieurs langues.

Les corpus comparables sont composés de documents ayant des traits communs (le genre, la période, le domaine, les thèmes abordés) sans être des traductions. Dans toutes les langues et pour tous les domaines techniques, il existe des documents publiés sur le web. Wikipedia est un exemple type de corpus comparable où des définitions réalisés par des locuteurs natifs existent pour de nombreuses langues et portent sur des domaines précis.

Le projet TTC vise à construire automatiquement des terminologies bilingues à partir de corpus comparables dans cinq langues européennes : anglais, français, allemand, espagnol et une langue peu dotée, letton, ainsi qu'en chinois et russe, et pour douze couples de langues :

- Chinois-Français et Français-Chinois
- Anglais-Français, Anglais-Allemand, Anglais-Russe, Anglais-Letton, Anglais-Espagnol
- Français-Allemand, Français-Russe, Français-Espagnol
- Allemand-Espagnol
- Letton-Russe

Les corpus comparables et les terminologies bilingues produits dans le projet TTC relèvent des deux spécialisés : les énergies renouvelables et la téléphonie mobile.

Le logiciel TermSuite issu du projet TTC est un logiciel libre, disponible ici

DEPART (Documents Ecrits et Paroles – Reconnaissance et Traduction) Projet Région Pays de Loire

2009-2012

- Partenaires : Le Laboratoire Informatique de Nantes-Atlantique (Coordinateur), le Laboratoire Informatique de l'Université du Maine (LIUM) et l'Institut de Recherche en Communications et Cybernétique de Nantes (IRCCyN).
- Objectifs : Ce projet vise la constitution au niveau de la région des Pays de la Loire d'un pôle de compétences unique en France associant analyse du signal audio et manuscrit au traitement automatique des langues. Il s'intéresse plus particulièrement à la résolution de problèmes scientifiques et technologiques difficiles mettant en jeu des données multimodales et multilingues.
- Pour en savoir plus : www.projet-depart.org METRICC

ANR Contint, www.metricc.com
2008-2010

Partenaires : Le Laboratoire Informatique de Nantes-Atlantique (Coordinateur), le Laboratoire d'Informatique de Grenoble (LIG), les entreprises Lingua et Machina, Sinequa et Syllabs ainsi que le Laboratoire de Recherche en Informatique de l'Université de Bretagne Sud (VALORIA).
Objectifs : Le projet MeTRICC vise à utiliser des corpus comparables en vue de l'extraction de lexiques multilingues dans le cadre des mémoires de traduction et de la recherche d'informations interlingue.

C-MANTIC

ANR Masse de données et connaissances ambiantes 2007-2010
- Partenaires : ERTIM, LIMSI, LINA, INSERM
- Objectifs :

Le projet C-Mantic vise à élaborer une méthodologie inédite de détection de l'information et d'organisation des masses de données documentaires dans une perspective multilingue. La méthode proposée est fondée sur une analyse sémiotique et linguistique approfondie prenant en considération l'ensemble des critères textuels et non pas seulement les mots-clés.

MILES - Axe multimédia
Région Pays de Loire, 2007-2009
- Partenaires : entre autres le LINA et le LIUM
- Objectifs :

Le projet MILES participe à la création d'un pôle européen de recherche en STIC dans l'Ouest de la France. Interdisciplinaire, MILES est porté par la fédération de recherche AtlanSTIC. L'équipe TALN du LINA intervient dans l'axe gestion de données multimédia du projet et en particulier sur la problématique de l'analyse conjointe de données multimédia telle que la reconnaissance des locuteurs dans des documents audio (langage naturel et traitement de la parole).
- Pour en savoir plus : Réalisations du LINA

PITHIIE
ANR Technologies logicielles, 2006-2008

- Partenaires : Advestigo, LIA, LINA, Synequa, Sylabs
- Objectifs :

Le projet Piithie s'inscrit dans un mouvement de plus en plus important de maîtrise de l'information diffusée. Il vise premièrement la détection de plagiats de textes. Les techniques de traitement automatique des langues (TAL), devraient permettrent d'améliorer les performances et d'accroître le potentiel de recherche des outils d'Advestigo et de Sinequa. Le deuxième objectif concerne le suivi d'impact : les diffuseurs d'information sont très intéressés par la possibilité d'évaluer l'mpact de leur production. Aujourd'hui cette évaluation est faite par une étude manuelle alors que des méthodes automatiques sont possibles.

BLOGOSCOPIE
ANR Technologies logicielles, 2006-2008

- Partenaires : LINA, Over-Blog, Synequa, Sylabs
- Objectifs :

Les blogs sont aujourd'hui au coeur de l'actualité : ils prennent une importance de plus en plus grande. Lus par une population de blogers qui représente assez bien la population globale, ils couvrent toute l'étendue de la production de l'information. Il s'agit d'un nouveau pouvoir informationnel qui est capable d'influencer les opinions des gens. Le but de ce projet est de développer des outils de surveillance des blogs qui permettent d'effectuer, automatiquement, deux tâches. La première est l'étude d'image. Elle veut créer une photographie de ce

que pense le public d'une certaine personne, d'un organisme, etc. à un moment donné. La deuxième tâche est l'étude de tendance qui veut déterminer, par exemple, les sujets émergents, l'état d'humeur d'une certaine population, l'évolution des sentiments vis à vis d'une certaine personne, organisme, etc.

- Pour en savoir plus : Réalisations du LINA

TCAN DECO
Projet DECO du programme interdisciplinaire du CNRS TCAN 2004-2006