Research fundings

Last Updated vendredi, 20 juin 2014

CURRENT PROJECTS

LIMAH

Comin Labs,
2014-2017

Participant

Consortium: TEXMEX INRIA Rennes (Coordinator), Laboratoire Informatique de Nantes-Atlantique (LINA), Telecom Bretagne, Univ. Rennes 2.

Description: Available multimedia content is rapidly increasing in scale and diversity, yet today, multimedia data remain mostly unconnected, i.e., with no explicit links between related fragments. The project Linking Media in Acceptable Hypergraphs (LIMAH) aims at exploring hypergraph structures for multimedia collections, instantiating actual links between fragments of multimedia documents, where links reflect particular content-based proximity—similar content, thematic proximity, opinion expressed, answer to a question, etc. Exploiting and developing further techniques targeting pairwise comparison of multimedia contents, LIMAH addresses two key issues of content-based graph-oriented multimedia collection structuring: How to automatically build from a collection of documents an hypergraph, i.e., graph combining edges of different natures, which provides exploitable links in selected use cases? How collections with explicit links modify usage of multimedia data in all aspects, from a technology point of view as well as from a user point of view? LIMAH will study hypergraph authoring and acceptability in two distinct complementary use-cases, namely, navigation in news data and learning with online courses.

TERMITH

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

Participant

Consortium: ATILF (Analysis and Natural Language Processing of French Language), INIST (National Institute of Scientific and Technical Information), LINA (Laboratory of Computer Science from Nantes), LIDILEM (Laboratory of Linguistics and Applied Linguistics of native and second languages from Grenoble) and two INRIA Centers (National Institute of research in Computer Science and Automatics), INRIA Nancy Grand-Est and INRIA Saclay.

Presentation: TERMITH deals with information access to textual documents via a full-text indexing which is based on terms which are detected, disambiguated and analyzed. This issue is well-known: the digital age is characterized by a very large quantity of information that has to be indexed to allow access to it, by the growing diversity of the areas and disciplines which entails a more and more frequent interdisciplinary. Text indexing based on terms occurring still is a hot research topic though different approaches have recently provided some good results. These approaches use either occurrences of terms which are detected on the basis of their textual form (projection of controlled vocabularies or structured terminologies using pattern matching, inflection rules, syntagmatic variations like for instance FASTR), or term candidates which result from some automatic terms detection components. All these methodologies require expensive human verification: (1) for indexing: manual checking of the automatically defined indexes or even, complete analysis of documents in order to define the good indexes of these documents, (2) for the automatic terms detection: classification of the very large amount of terms candidates, (3) for the projection of controlled vocabularies or structured terminology: updating of the terminological resources.

CRISTAL

ANR Contint,
2012-2014 www.projet-cristal.org
Member
Consortium: Laboratoire Informatique de Nantes-Atlantique (Coordinator), ERSS, Université de Genève, Lingua et Machina.

Description: The globalization of commercial trade has made it necessary for all companies to communicate with their partners, clients, and employees in their native language. Often, employees must also be capable of easily communicating in a foreign language, particularly in their field of expertise. Terminology competency (either multilingual or monolingual) is then essential, as company documents must be created in a coherent and time-efficient manner. However, linguistic resources adapted to companies' specific technical fields are exceedingly rare. Furthermore, technical knowledge, and the terms used to describe it, evolve very rapidly, which makes the manual creation of such resources cost-heavy and susceptible to being quickly obsolete. Computer-assisted translation tools, as well as other innovative terminology management tools, address this problem, by meeting the needs of private companies and public bodies for resources to aid and manage multilingual and monolingual writings. Currently, these tools are able to extract a veritable dictionary of the company, which gives information about, translates, and gives access to terms that are unique to the company as well as technical terms that are related to the business sector in question.
The goal of the CRISTAL project is to develop a method of extracting Knowledge-rich Contexts (KRC) in order to create new and innovative types of dictionaries. These will feature, for each term and its possible translations, a terminological index listing the KRCs and the knowledge that they contain. The extraction of such contexts requires the creation of advanced and robust algorithms that are able to extract knowledge and to automatically and efficiently analyze linguistic and semantic content in unstructured plain texts, no matter the language. Our goal is to develop a new generation of tools to assist translation and terminology management, which will facilitate multilingual communication.

PAST PROJECTSTTC- Terminology Extraction, Translation tools and Comparable corporaFP7 - Information Society and Media - ICT 2009.2.2: Language-based interaction
2010-2012 Coordinator www.ttc-project.eu

The FP7 research project TTC (Terminology Extraction, Translation Tools and Comparable Corpora) project leveraged machine translation (MT) systems, computer-assisted translation (CAT) tools and multilingual content (corpora and terminology) management tools by developing methods and tools that allow users to generate terminologies automatically from comparable (non-parallel) corpora in seven languages: five European languages (English, French, German, Spanish, Latvian) as well as Chinese and Russian. The tools were tested on twelve translation directions: Chinese-English, Chinese-French, English-French, English-German, English-Russian, English-Latvian, English-Spanish, French-German, French-Russian, French-Spanish, German-Spanish, Latvian-Russian. The TTC project has developed generic methods and tools for the automatic extraction and alignment of terminologies, in order to break the lexical acquisition bottleneck in both statistical and rule-based MT. Thereby, it contributed to domain adaptation for SMT. It has also developed and adapted tools for gathering and managing comparable corpora, collected from the web, and managing terminologies. In particular, a topical web crawler and open terminology platform (MyEuroTermBank3) have been developed. The key output of the project is the TTC web platform4. It allows users to create thematic corpora
given some clues (such as terms or documents on a specific domain), to expand a given corpus, to create a comparable corpus from seeds in two languages, to choose the tools to apply for terminology extraction, to extract monolingual terminology from such corpora, to translate bilingual terminologies, and to export monolingual or bilingual terminologies in order to use them easily in automatic and semi-automatic translation tools.
For generating bilingual terminologies automatically from comparable corpora innovative approaches have been researched, implemented and evaluated that constituted the specificities of the TTC approaches: (1) topical web crawling which will gather comparable corpora from domainspecific Web portals or using query-based crawling technologies with several types of conditional analysis; (2) for monolingual term extraction, different techniques, a knowledge-rich and a knowledge-poor approaches were followed; a massive use of morphological knowledge to handle morphologically complex lexical items; (3) for bilingual term extraction, an unified treatment for single word term and multi-word term was designed as well as an hybrid method that used both the internal structure and the context information of the term.
We measured the impact of the TTC bilingual glossaries on MT output and found that they considerably improve human perception of MT quality and speed up production of professional quality translations. The utility of TTC tools improved also the productivity of translators during the execution of a localization project when the TTC-generated glossaries are integrated in commercial

tools, the translation of terminology alone is of 40%.

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

www.projet-depart.org 2009-2013 member

METRICC

ANR Contint, www.metricc.com
2008-2012 coordinator

Nominated at "Digital Technologies ANR Awards" April 2013 ANR

The MeTRICC project addresses the issue of comparable corpora from two perspectives: the gathering of texts and their operational condition of use. Collecting texts is done with a focused web crawler which starts from a small set of seed words of a domain and downloads texts from the Internet which are relevant to the domain. In order to measure the comparability of the crawled texts in two languages, new measures are proposed which guarantee the quality of the bilingual term-related resources which will be built from these texts. New methods are requested to identify inside comparable corpora pieces of translations that could be entries of bilingual dictionaries. State-of-the-art methods that rely on the simple observation that a word and its translation tend to appear in the same contexts will be improved. Furthermore, the bilingual lexicon that will be built are evaluated in several industrial applications, such as computer-assisted translation tools, and professional search engine in order to perform multilingual search.

C-MANTIC

ANR Masse de données et connaissances ambiantes, www.c-mantic.org
2007-2009 participant

P { margin-bottom: 0.21cm; direction: ltr; color: rgb(0, 0, 0); text-align: justify; widows: 2; orphans: 2; }P.western { font-family: "Palatino Linotype",serif; }P.cjk { }P.ctl { font-size: 12pt; }A:link { color: rgb(0, 0, 255); }A.sdfootnotesym-cjk { }A.sdfootnotesym-ctl { }

C-Mantic was a research project supervised by INaLCO. The work was done with
LINA, LIMSI and INSERM. Secondary partners are “Alliance pour le
tabac” and SYLED. The project kicked off in January 2008 and lasted
36 months.

P { margin-bottom: 0.21cm; direction: ltr; color: rgb(0, 0, 0); text-align: justify; widows: 2; orphans: 2; }P.western { font-family: "Palatino Linotype",serif; }P.cjk { }P.ctl { font-size: 12pt; }A:link { color: rgb(0, 0, 255); }A.sdfootnotesym-cjk { }A.sdfootnotesym-ctl { When
information is mined from text (filtering, knowledge extraction,
opinion mining…), the methods used are mostly mathematical.
Classification algorithms (machine learning) make up the bulk of the
methods used. Linguistics is only called on indirectly, through

Natural Language Processing, in relation to such tasks as standardization, lemmatization or syntactic analysis. The expertise of linguists as applied to text semantic analysis is hardly made use of, although meaning is the core concern of the applications developed. The C-Mantic project's goal was to elicit, model and develop software tools based on semantic expertise. The final result is a software platform aiming at (i) building corpora from massive heterogeneous document collections, (ii) the textual analysis of such corpora, (iii) the development of filtering and categorization criteria. The methods used have been tested successfully on health and social issues, namely the categorisation of web documents related to smoking. The methods can be ported to new application fields.

MILES - Multimedia theme

Région Pays de Loire, www.paysdelaloire.fr
2007-2009   co-coordinator of the Multimedia theme with M. Gelgon, LINA

PITHIIE

ANR Technologies logicielles, www.piithie.com

2006-2008   participant

BLOGOSCOPIE

ANR Technologies logicielles, www.blogoscopie.org

2006-2008   coordinator

TCAN DECO

DECO project of the CNRS TCAN interdisciplinary program

2004-2006   coordinator