Expression of Interest

SemTech

Acquiring specific semantic knowledge for an access to textual content

Coordination: Laboratoire d Informatique de Paris-Nord (LIPN, Univ. Paris-Nord & CNRSTIC)

This EoI presents the motivation for a multidisciplinary Network of Excellence. This NoE would focus on the problem of the acquisition of the semantic knowledge that is involved in text content analysis technologies such as IR and IE. It would allow going further towards the development of a unifying methodology of specific knowledge acquisition and therefore to overcome the limitations of approaches based on pre-existing general ontologies and lexical ressources.

The proposal is organised as follows. Section 1 presents the motivation for the EoI and. It addresses the need and relevance of research activities in this area as a NoE. Section 2 details the main issues on which this NoE would have to focus. Section 3 presents the complementary roles of the potential participants. It shows how this NoE would contribute to integrating research activity in this new domain where today the research effort is mono-disciplinary and scattered.

1. Need and relevance

As well as the generalisation of multimedia communication, the volume of textual information is exponentially increasing. Today mere Information Retrieval technologies are unable to meet specific information needs because they provide information at a document collection level. Developing intelligent tools and methods, which can give access to document content, is therefore more than ever a key issue for knowledge and information management. Text content access is a crucial issue as much in the document engineering system of a small firm as in the document management of a whole scientific domain, whichever the source of information is: an Intranet information system or the "semantic web".

As soon as one wants to automate access to the content of texts in electronic form, one needs semantic knowledge to localise and interpret the relevant information. The acquisition of semantic knowledge is a well-known bottleneck for real-world applications, whichever technology is used (Information Extraction, Question/Answering, and more generally document and knowledge engineering). There are two main reasons. Firstly, little semantic knowledge specific to application domains has been available because, until now, effort has been mainly devoted to the definition of formal languages for the representation of ontology and to the acquisition of generic knowledge bases, either lexical databases such as WordNet or EuroWordNet or general ontologies; CYC, for instance. In contrast, almost no community effort has been devoted to the acquisition of specific semantic knowledge that is required for particular applications and to the design of the acquisition methods that could be applied. We claim that no generic knowledge can be used as such and that the required semantic knowledge, must be specifically acquired or tuned according to the application, domain and task that it will be used for. Although the process of acquiring this specific semantic knowledge cannot be fully automatic, methods and tools can be designed to efficiently help its acquisition. Secondly, it is also noticeable that there has been little dialogue between the various disciplines involved in knowledge acquisition and text analysis, although the integration of methods and tools from various disciplines is obviously needed. These disciplines include Information Science, Linguistics, Natural Language Processing, Knowledge Acquisition, Knowledge Representation, Machine Learning, Information Retrieval and Information Extraction. Various experiments in this framework have been conducted which have yielded very encouraging results. For instance, in France, a whole community of researchers have emerged from the TIA working group, which has been working for almost 10 years on methods for the acquisition of terminological and ontological resources from textual data. Different teams, coming from all the research fields mentioned above have been involved. However effort remains scattered. The research must evolve from *ad hoc* experimental methods towards scalable methodological principles and tools.

To go further towards the development of a unifying methodology and of set of tools for the acquisition of semantic knowledge and its exploitation to give access to text content, it is now crucial to structure the research effort in order to combine these approaches.

2. Crucial issues

The section gives a more detailed description of the semantic knowledge acquisition field as we see it, focusing on a few crucial issues.

On the nature and the representation of the knowledge to be acquired and integrated

a) The required semantic knowledge is domain and task specific

If one wants to automate the understanding and exploitation of the information content of texts, semantic knowledge is required, in whatever form it is represented (dictionaries, terminologies, thesauri, thematic maps, key-word lists, linguistic patterns .) and this knowledge must be specific to the domain and the task which are considered. Generic ontologies or lexicons cannot be exploited as such: they must be adapted, tuned, and specialised for every specific application for a better efficiency.

b) Methods for identifying and integrating relevant available heterogeneous knowledge sources

Acquisition should make use of available data and information to produce the semantic knowledge. These available sources are heterogeneous in terms of generality, quality and domain coverage. Until now, corpus-based acquisition methods have been mainly developed to take advantage of textual material for acquisition but non-textual knowledge sources must also be taken into account. Linguistic resources such as dictionaries, lexicon and thesauri provide important morphological, syntactical and semantic knowledge. Generic sources must be tuned to a specific application whereas specialised sources often need to have their coverage enlarged. Human expert knowledge is a third type of knowledge to be incorporated. It is required at the very beginning of the acquisition process for the identification of relevant texts and resources but also in the course and at the end of the acquisition process for control and validation.

The key problems are hence to identify the relevant sources and their domain of validity and to develop methodologies for combining information and knowledge in a homogeneous acquisition process.

c) Text content cannot be reduced to thematic information

Until now, text mining research has focused mainly on thematic information. With the development of more powerful methods, it becomes possible and necessary to take other types of information into account. A measure of information reliability, the understanding of the point of view and the identification of text gender or diffusion status, also constitute valuable document characteristics that may be of interest.

d) Semantic knowledge must be anchored at the linguistic level, the multilinguality question

As soon as the knowledge acquired is to be used for text access, the two levels of the conceptual description and the text must be articulated. This anchoring is easier to guarantee for text-based acquisition methods, but it is crucial to ensure that the textual data used for acquisition is representative of the target corpus for which the access tool is designed.

It is important to note that a semantic resource is specific to a given language. Text content access tools are generally multilingual but the acquisition phase is to be conducted in parallel in the different languages. However, even if the resulting semantic knowledge is not reusable from one language to another, the acquisition method could be reusable. This point should be raised in the near future.

The acquisition process

e) The acquisition / exploitation cycle

The overall process of text content access can be viewed broadly as a two-step one. The first step (acquisition) aims at building all the linguistic and conceptual knowledge bases and the second step (exploitation) then exploits these knowledge bases to analyse text content. Semantic knowledge bases must also be incrementally revised to fit the evolution of the exploitation practice. Actually, the whole process is often a spiral one with an iteration of acquisition-exploitation phases, of increasing richness. The knowledge acquired at a given stage is exploited to acquire new knowledge. For example, concept hierarchies may be used for extracting relations between these concepts from texts.

f) Heterogeneous methods and tools must be integrated

Up until the 1990 s, various specific methods have been developed to acquire pieces of knowledge, such as word semantic classes, generic-specific relations, bilingual translation couples, terminological expressions, syntactical-lexical extraction patterns These methods from the fields of text analysis, machine learning, information retrieval, and data analysis differ in the techniques they rely on, in the core initial knowledge they make use of (generic vs. specific, conceptual vs. textual), in their goals (initial acquisition, updating, tuning) .

There is an urgent need to integrate these methods into a unifying methodology. This is a technological challenge insofar as software toolkits must be developed. This is also a theoretical question since it is necessary to identify the precise scope, ability and limitation of each elementary method and to understand how they can cooperate.

The human/expert/user/knowledge engineer in the loop

g) Semantic knowledge acquisition is a cooperative process between man and machine

The acquisition of application-oriented knowledge is problematic. It is now accepted that fully automatic acquisition is unrealistic. However manual acquisition is difficult and time-consuming. We argue that a middle way is appropriate. Automatic tools can efficiently help the building of specific semantic knowledge and they can take the final application (task and domain) into account. For instance, the automatic analysis of a sample of the target texts (e. g. learning from corpus) can be a useful step in the acquisition process. Automatic tools can bootstrap the acquisition process, check the consistency of the resulting knowledge base, generalise a core knowledge base and tune a pre-existing lexical base to a given corpus. In any case, however, human control is necessary. The remaining question is to identify precisely the role devoted to humans in this acquisition process and to define methods and interfaces enabling a cooperative process.

h) A use-oriented evaluation protocol is to be defined

The lack of relevant and reliable evaluation is one of most critical points of current research work. The knowledge base resulting from acquisition cannot be evaluated *per se*. Even if unitary tests can be made, the quality of the whole process of acquisition can only be appreciated with respect to an enhancement of the quality of text access at the exploitation level.

These requirements call for the development of real-world applications. The experimental feedback will help to define use-oriented evaluation guidelines and this will be of value to the whole community.

3. Integrating and structuring the research effort

The cooperation within each scientific field has been encouraged and supported by the EEC. The technology development in these fields is now at a stage where new progress require the close collaboration of the other technology and human science fields. The emergence of this need is reflected by the organisation of research projects, working groups and several workshops and conferences, which show that efforts have already been made to federate a community but there is need for further and closer co-operation at a European level, in particular with human science. In the past, two third of the participants have been involved in more than 90 European projects that addressed parts of our EoI (in IST as well as HLT, MLIS, COST, etc.). One third are new at a European level, and some of them were involved in national focused projects with other participants of the EoI. Many application fields (hydrology, medicine, genomics, agronomy, glass production, aerospace, CRM, petrochemistry) have been studied in these projects, which yield encouraging but limited results.

Some working groups have been also set up at a national level, (e.g. in France, the working group TIA) and Esprit and IST NoE (MLNet, ILPNet, KDNet, OntoWeb, among others) have contributed to the emergence of the field.

Several workshops and conferences also testified to the emergence of this semantic acquisition problem: among others, EKAW 2000 and 2002, French TIA conferences in 1997, 1999 and 2001, COLING 1998 and COLING 2002 CompuTerm workshops, ECAI'2000, IJCAI'2001 on "Ontology Learning", ECAI 2002 OntoText workshops, several LREC 2002 workshop (Creating and using semantics Information Retrieval and Filtering), text mining workshops at ML and ECML. This question is also raised in specialised conferences (in domains such as healthcare, genomic), which show its practical importance. All these

results and the effort must be integrated now so that the advantages and limitations of the possible approaches are stated and general methodology and design guidelines can be produced. A NoE would contribute to integrate and structure the research activity in this multi-pluridisciplinary domain and to involve new research centres, especially from human science, information science and linguistics and new SME, especially software developers.

Three types of participants are involved in this Expression of Interest:

1) Research groups that develop new methods, concepts, prototypes and platforms and actively contribute to innovation in this field. This NoE would gather new research teams and outstanding actors. The research will benefit from a multi-disciplinary approach from the following scientific domains.

¥˚**KA-KR**- Knowledge Acquisition and Knowledge Representation formal languages will be used to give an adequate representation of the semantic knowledge. Experience in knowledge-based techniques will also help to embody semantic knowledge into text access tools.

1. Equipe CSC, Institut de Recherche en Informatique de Toulouse (IRIT), Toulouse, France
2. Projet ACACIA, INRIA Sophia-Antipolis, France (+ML)
3. Orpailleur Team, INRIA-LORIA, France (+ML, NLP)
4. National Research Council, ISTIC-CNR, Italy
5. Linguistische Informatik/Computerlinguistik, Universitaet Freiburg, Germany (+NLP, ML)
6. Saint Gobain — Centre de Recherche -, AubervilliersFrance (+ IR, end-user manufacture)
7. Information Management Group, University of Manchester, United-Kingdom (+ end-user biology)

- **NLP** - Linguistic background is crucial. Corpus Linguistics should be able to define cues to identify various textual types and thus help to measure the representativity of text samples. Terminology offers the means to analyse sub-languages. Identifying terminological units, which are the relevant semantic units for any specific domain, is often considered as a starting point for the acquisition process. Natural Language Processing techniques are required to automate textual analysis, from the morphological level to the syntactical and semantic ones. Computational terminology can bootstrap knowledge acquisition and give it its textual anchor.
  1. Mission de Recherche en Sciences et Technologies de l'Information M dicale, Assistance Publique - H pitaux de Paris, France (+ML, KA-KR)
  2. University of Sheffield, United-Kingdom (+ML, KA-KR)
  3. Centre de recherche Termisti, Belgium (+KR)
  4. Laboratoire de linguistique informatique (LLI), D partement de linguistique et de traduction, Universit de Montr al, Canada (+ KR)
  5. Xerox Research Center Europe, Grenoble, France (+IR, ML).
  6. Laboratoire d Informatique de Paris-Nord (LIPN), Universit Paris-Nord & CNRSTIC, France (+KA-KR)
  7. quipe LaLICC, CNRS & Universit Paris-Sorbonne, France
  8. Team TexMex, IRISA, INRIA, France (+ML, +end-user biology)
  9. Laboratoire Dyalang, Universit de Rouen, France
  10. CNTS Language Technology Group, University of Antwerp, Belgium (+ML)
  11. Lunds universitet, Sweden (+ Visualization)
  12. School of Computer Science, Queen's University Belfast, Northern Ireland, United-Kingdom (+˚Visualization)
  13. Tor Vegata group, Universita di Roma, Italy

- **ML** - Machine Learning, data mining and data analysis must be involved in acquisition processes, since it helps to identify regularities in data and bring out salient phenomena.
  1. Artificial Intelligence Group, Dept of Computer Science, University of York, United-Kingdom (+NLP)
  2. Math matique, Informatique et G nome (MIG), INRA Versailles, France (+KA, end-user biology)
  3. Computer Science Department, University of Dortmund, Germany (+KA, IR)
  4. Neurol Networks and Machine Learning groups, LEIBNIZ-CNRS, IMAG, France
  5. Laboratoire d Informatique de Paris 6 (LIP6), Universit Paris VI & CNRS, France (+IR)
  6. Royal Holloway, University of London, United-Kingdom
  7. Machine Learning group, Computer Science Department, University of Bristol, United-Kingdom
  8. Institute AIFB, University of Karlsruhe, Germany (+KA-KR)
  9. Laboratoire de Recherche en Informatique (LRI), Universit Paris Sud & CNRS, France (+KA-KR)
  10. Dipartimento di Informatica, Universit degli Studi di Bari, Italy (+IR)
  11. Department of Social Science, Informatics, University of Amsterdam, Netherland (+KA)
  12. Department of Intelligent Systems, Institute "Jozef Stefan", Slovenia (+IR)
  13. Knowledge Discovery Group, Masaryk University in Brno, Faculty of Informatics, Czech Republic (+NLP)

EoI in the 6th FP: Acquiring specific semantic knowledge for an access to textual content (SemTech)

14. quipe Contraintes et Apprentissage, Laboratoire d Informatique Fondamentale d Orl ans (LIFO), Universit d'Orl ans & CNRS, France
15. Department of Information and Knowledge Engineering, University of Economics, Praha, Czech Republic.
16. Department of Cybernetics and Artificial Intelligence, Technical University of Kosice, Slovakia

- **IR** - Information Science is mainly involved in the exploitation phase. Its main contributions will be in the definition of needs and evaluation. Information Retrieval will provide methods for handling and visualizing large set of documents to prepare the acquisition process.
    1. Laboratoire RECODOC, Universit Claude Bernard Lyon 1, France
    2. IZ, Information Centre of Social Science, Bonn, Germany
    3. LemonLabs GmbH, M nich, Germany
    4. quipe SIG, Institut de Recherche en Informatique de Toulouse, Toulouse, France
    5. Artifical Intelligence group, Engineering Mathematics Department, University of Bristol, United-Kingdom (+ML)
    6. School of Computer Science, Queen's University Belfast, Northern Ireland, United-Kingdom
    7. Artificial Intelligence Laboratory, cole Polytechnique F d rale de Lausanne, Switzerland (+NLP)
    8. Helsinki Institute for Information Technology, Finland (+ML)

2) Software editors involved in linguistic technologies and/or knowledge management who are willing to industrialise and market the new tools that will be developed; Companies which provide services in the domain of information access (e. g. knowledge management, CRM)

1. Ontoprise GmbH, Germany (KA, IR)
2. Intelligent Software Components S.A., ISOCO, Spain (KA, ML, IR)
3. Isoft, France (ML, biology)
4. Sword Information and Communication Technology, Fasano, Italy (IR)
5. BT Exact, United-Kingdom (KA, NLP, IR)
6. StreamSage, Inc. USA (IR)
7. AltaVista, USA (NLP, IR)
8. Language & Computing nv, Belgique (NLP, healthcare, biology)
9. Sinequa, France (NLP, IR)

3) End-users, research centres, companies or organisations, who need to automate the access to textual documents and are willing to make use of the new tools for their own specific activity. The application needs are central in our proposal. Among the potential application domains we want to mention some candidates where crucial needs have been identified: life science (e.g. genomics, agronomy, pharmacology, healthcare), chemistry (e.g. petrochemistry), physics (e.g. hydrology), manufacturer (e.g. aerospace, car), communication (press, TV) and finance.

1. Mission de Recherche en Sciences et Technologies de l'Information M dicale, Assistance Publique - H pitaux de Paris, France (healthcare, +NLP, ML, KA-KR)
2. Institut de Recherche et de Coordination Acoustique Musique (IRCAM), France (music, + IR, KA, ML)
3. MRC Mammalian Genetics Unit , United-Kingdom (biology)
4. Unit de Recherche et Innovation (URI) de l'Institut de l'Information, Scientifique et Technique (INIST), CNRS, France (scientific documentation)
5. Laboratoire de G n tique et Physiologie du D veloppement (LGPD), CNRS, France (biology)
6. National Center for Biotechnology (CNB-CSIC), Madrid, Spain (biology, +ML, KA, IR)
7. Swiss Institute of Bioinformatics, Switzerland (biology, +IE)