KNOW2: Language understanding technologies for multilingual domain-oriented information access*

KNOW2: Tecnologías de comprensión del lenguaje para el acceso multilingüe a la información orientada a dominios

Eneko Agirre
German Rigau
EHU, IxA taldea
e.agirre@ehu.es
german.rigau@ehu.es

Irene Castellón
UB, GRIAL
icastellon@ub.edu

Salvador Climent
UOC, GRIAL
scliment@uoc.edu

Jordi Turmo
Lluis Padró
UPC, TALP
turmo@lsi.upc.edu
padro@lsi.upc.edu

Resumen: El objetivo de KNOW2 es avanzar en el desarrollo de un entorno integrado que permita la implantación a bajo coste de portales verticales de acceso a la información para dominios concretos. El proyecto tiene una duración de tres años y acaba de comenzar en enero del 2010.
Palabras clave: Procesamiento del Lenguaje Natural, Análisis Sintáctico, Interpretación Semántica, Adquisición de Conocimiento, Extracción de Información, Recuperación de Información

Abstract: The goal of the project is to explore integrated environments allowing the cost-effective deployment of vertical information access portals for specific domains. The project started in January 2010, and will last three years.
Keywords: Natural Language Processing, Syntactic Analysis, Semantic Interpretation, Knowledge Acquisition, Information Extraction, Information Retrieval

1. General description

New forms of (multilingual) information access (MLIA, IA) based on Natural Language Processing (NLP, specially featuring semantic information) are being adopted by strong companies such as Google, Microsoft or Yahoo: Question Answering has been deployed (PowerSet -now part of Microsoft-, Yahoo Answers, Google), IA centered on entities is being explored (Spock, Yahoo, Silobreaker) alongside new navigation strategies (MMexplorer), and cross-lingual IA has been deployed by major search engines (Google).

KNOW2[1] is a coordinated project which just started in January 2010, and will last for three years. It involves researchers from four universities (EHU, UoC, UB and UPC).

The project is based on the idea that automatic text processing, specially in the semantic layer, is already enabling a new generation of MLIA systems. In order to acquire the required knowledge and process free-running text accurately, our strategy has three interconnected threads: **(1)** Thee need to focus on specific domains, and thus apply text mining and domain adaptation techniques to improve NLP tools and resources, including inference and reasoning capabilities. **(2)** The need to include users and domain experts in the loop, via collaborative interfaces to the acquired knowledge. **(3)** The acquired knowledge should allow to build cost-effectively vertical IA portals for domains.

2. Relation to other projects

KNOW2 builds on the results of KYOTO and KNOW. KYOTO[2] is a three year European project which proposes a system that allows people in communities to define the meaning of their words and terms in a shared Wiki platform so that it becomes anchored across languages and cultures but also so that a computer can use this knowledge to detect knowledge and facts in text. We plan to use and further develop the software and expertise gathered in KYOTO.

KNOW[3] is the predecessor of KNOW2, and it already enhanced Cross Lingual IA and Question Answering technology with

* The project is funded by the Ministerio de Ciencia e Innovación TIN2009-14715-C04

[1] http://ixa.si.ehu.es/know2

[2] `http://www.kyoto-project.eu`

[3] `http://ixa.si.ehu.es/know`

improved NLP technologies for the open-domain. With respect to KNOW, KNOW2 aims to obtain better performance by using two main strategies: (i) moving from general to specific domains and (ii) incorporating text-mining and collaborative interfaces.

3. Project coordination

The ambitious goals on the project can only be achieved gathering a critical mass of researchers. For this reason KNOW2 has been designed as a coordinated project integrating the research and the multilingual abilities of three groups, which are structured in three subprojects with well-defined goals:

Subproject 1 (EHU) focuses in management and design, development of collaborative interfaces, reasoning and inference, layers, linguistic processors for Basque, question answering, extraction of multilingual lexical knowledge, adaptation of linguistic processors to the domain, integration of the knowledge gathered in the rest of subprojects and evaluation.

Subproject 2 (UPC) focuses on the study, evaluation and comparison of advanced text mining techniques to support the building of domain ontologies; this goal involves enhancement of machine learning techniques and improvements in sintactic-semantic processors and knowledge acquisition for text classification, information extraction, question answering and textual entailment.

Subproject 3 (UOC-UB) focuses in linguistic research for developing semantic processors and in building lexical-semantic knowledge bases (WordNets) for Spanish and Catalan using Machine-Translation and Computer-Assisted Translation techniques.

4. Specific objectives

The main objective is to improve current MLIA systems with research that enables the construction of an integrated environment allowing the cost-effective deployment of vertical IA portals for domains, which comes down to the following specific objectives:

- Adoption of current standards for the representation of linguistic annotations, both of documents and of semantic resources. This adoption will enable easier interoperability and an easier adoption of KNOW2 technology by the industry. In addition, KNOW2 will support free software licenses of all developed tools and resources.
- Development of robust linguistic processors, including semantic processing, for Basque, Catalan and Spanish; procedures to adapt those processors, and English ones, to the target domain; analysis of discourse structure.
- Development of knowledge mining techniques, which will mine domain texts and enrich (and adapt to the domain) current multilingual knowledge bases with concepts, relations and factual events. The acquisition will be driven by automatically captured document collections.
- Development of a collaborative interface to the domain knowledge. This wiki-style interface will allow the user community to manage the whole process, including the edition of the acquired concepts, domain ontologies, and the extraction rules.
- Integration of all acquired knowledge in a single Multilingual Central Repository. Development of a semantic engine which will include new techniques for automatic reasoning and inference, and which will be adapted to the domain.
- Development of prototypes for the monolingual and multilingual IA to the documents and factual information extracted from them. It will include Information Retrieval, Cross-Lingual IA and Question Answering demonstrators.
- Resources, tools, and applications will be evaluated in international benchmarks and competitions whenever possible.

5. Defining cases of use in real scenarios

KNOW2 will produce demonstrators and prototypes on different cases of use in real scenarios related to specific domains, such as environment, European parliament, geographic text and/or popular science and technology (including public portals like zientzia.net and BasqueResearch, part of AlphaGalileo). We are currently working on the definition of such set of cases of use in collaboration with collaborating companies (EPOs). In this sense, we are opened to any kind of suggestions from interested companies.

KNOW2 wants to apply state-of-the-art research to real scenarios. The adoption of recent representation standards and free software licenses should facilitate technology transfer to industrial environments.