Procesamiento del Lenguaje Natural, núm. 43 (2009), pp. 377-378 recibido 1-05-2009; aceptado 5-06-2009

KNOW: Developing large-scale multilingual technologies for language understanding*

KNOW: Desarrollo de tecnologías multilingües a gran escala para la comprensión del lenguaje

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

Irene Castellón
Laura Alonso
UB
GRIAL
icastellon@ub.edu
alemany@famaf.unc.edu.ar

Lluis Padró
Montse Cuadros
UPC
TALP
padro@lsi.upc.edu
cuadros@lsi.upc.edu

Salvador Climent
Marta Coll-Florit
UOC
GRIAL
scliment@uoc.edu
mcollfl@uoc.edu

Resumen: El proyecto KNOW pretende añadir significado, conocimiento y razonamiento a las tecnologías actuales de Procesamiento del Lenguaje Natural.
Palabras clave: Procesamiento del Lenguaje Natural, Análisis Sintáctico, Interpretación Semántica, Adquisición de Conocimiento, Recuperación de Información

Abstract: The KNOW project aims to add meaning, knowledge and reasoning to current Natural Language Processing technologies.
Keywords: Natural Language Processing, Syntactic Analysis, Semantic Interpretation, Knowledge Acquisition, Information Retrieval

1. General description

KNOW aims to add meaning, knowledge and reasoning to current interface technologies. Specifically, KNOW is providing novel natural language interpretation and reasoning capabilities to current multilingual computer applications: full syntactic parsers and semantic interpreters (including word sense disambiguation systems and semantic role labelers) for the languages involved in the project, a common conceptual structure (the Multilingual Central Repository or MCR), and both automatic reasoning and analogy-based inference based on the MCR. KNOW has opened the way for a new generation of broad-coverage unrestricted-domain concept-based language understanding applications. KNOW is demmonstrating the feasibility of these technologies in two applications linked to the project EPOs: Cross Lingual Information Retrieval and Question/Answering on two multi-modal databases. The project started on 2006 and will finish in October 2009.

The project has the following goals:

1. Improving current syntactic analysis of Spanish, Catalan and Basque, combining machine learning techniques, hand-written rules, and integrating the deep semantic information acquired below.
2. Automatic acquisition of deep knowledge on verbal models, including selectional preferences and semantic roles.
3. Integration of large-scale semantic knowledge in the MCR.
4. Advancing in the use of Machine Learning techniques for the resolution of semantic analysis and language understanding tasks, namely, word sense disambiguation, and semantic role labeling, approached either as independent tasks, or in a simultaneous way.
5. Developing efficient deduction and automatic reasoning techniques able to generate new knowledge from the existing in MCR, and using it as an additional source of knowledge to include in the MCR.
6. Proving the viability of these applications, and the advantages that the use of advanced semantic knowledge represents, in two demonstrators for cross-lingual information retrieval and cross-lingual question answering.

2. Main Results

Up-to-date, the main achievements of the project are:

* TIN2006-15049-C03

Linguistic Processors: The project has gathered, adapted and enriched the basic tools and linguistic resources available for all the tasks in the project in English, Spanish, Catalan and Basque, including: tokenization and sentence boundary detection, morphological analysis and treatment of referential expressions, named entity recognition and syntactic analysis. We have also developed grammars for deep syntactic analysis of the languages in the project. There are three demos available for the Basque surface parser, the Basque full parser, and the Spanish, Catalan and English parsers. Finally, we have also developed a consistent annotation of the nominal part of the EuroWordNet with the Top Concept Ontology.

Knowledge Integration: KNOW has managed and maintained the MCR, which is used for uploading and porting the knowledge acquired across languages and resources and maintaining the compatibility among them. The MCR includes modules for: uploading the data acquired from one language to the MCR, porting the knowledge stored in the MCR to the local wordnets, checking the consistency of the TCO and exporting their content using the standard LMF. The content of the MCR can be consulted following the links in the project website. In addition, we have enriched by semi-automatic means the reasoning rules of Top Concept Ontology, which have been used by automatic theorem provers and inferencing tools in the reasoning tasks.

Acquisition of semantic knowledge: We have designed and applied automatic acquisition techniques, both supervised and unsupervised, to widely representative corpora, to infer lexical selectional preferences, with special attention to verbal preferences for semantic roles. They are available for English. We have also acquired an extensive amount of new semantic relations, forming what we call KNOWNETs for the project languages, which are publicly available. In addition, we have mined the MCR for Base Level Concepts, which offer different generalization levels. Base Level Concepts are also freely available. The acquired knowledge is being used to improve word sense disambiguation, semantic role labeling and to improve parsing, with positive results on both tasks for English.

Semantic Processing: A knowledge-based Word Sense Disambiguation (WSD) system based on the MCR has been developed. This system is capable of disambiguating words for any language with a WordNet and the results surpass the state-of-the-art. The system is available as open source. A demo is also available. In parallel, we have continued to develop supervised WSD, with new developments on the use of untagged corpora. In the case of Basque, we have completed the annotation of the Basque SemCor (which can be consulted online) and developed a supervised WSD based on this annotated corpora. For Spanish, we have initiated the annotation of the Sensem Corpus with senses of WordNet. Regarding Semantic Role Labeling, we have developed a system which use the selectional preferences acquired, and a joint-learning system which has an online demo.

Reasoning: KNOW has produced a layer of inference on top of the knowledge acquired and integrated in the MCR, using the a FOL theorem provers working with formal rules based on SUMO and graph-based algorithms working with all the semantic information in the MCR.

Evaluation and demonstration: We have evaluated the quality and accuracy of the developed software and the acquired data. Objective measures on significant samples of the data and software results have been taken, and we have also participated in public evaluation exercises on both parsing, information retrieval, semantic role labeling and word sense disambiguation, with high ranking positions on all.

We also have demonstrated the feasibility of integrating the project results in Information Retrieval and Question Answering. Two demos are available, although the ArgazkiPress demo is only available at request.

3. Future work

The main objective for KNOW in the last part of the project is to further improve current results on the applications using the knowledge and tools developed during the two previous years, with special attention to semantic technologies.

Additional information, including a longer version of this document, plus links to demos, resources, tools and publications, are available at the project website[1].

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