Procesamiento del Lenguaje Natural, núm. 43 (2009), pp. 389-390 recibido 1-05-2009; aceptado 5-06-2009

KYOTO Project*

Eneko Aguirre, Arantza Casillas, Arantza Díaz de Ilarraza, Ainara Estarrona, Kike Fernández, Koldo Gojenola, Egoitz Laparra, German Rigau, Aitor Soroa
IXA Taldea. UPV-EHU
german.rigau@ehu.es

Resumen: El proyecto Kyoto construye un sistema de información independiente del lenguaje para un dominio específico (medio ambiente, ecología y diversidad) basado en una ontología independiente del lenguaje que estará enlazada a Wordnets en siete idiomas.

Palabras clave: Extracción de Información

Abstract: The KYOTO project will construct a language-independent information system for a specific domain (environment, ecology and biodiversity) anchored in a language-independent ontology that is linked to wordnets in seven languages.

Keywords: Information Extraction

The KYOTO project[1] (ICT-211423) stands for Knowledge Yielding Ontologies for Transition-based Organization. KYOTO is a co-funded by the European Union[2] and by national funding of Taiwan and Japan. The project started in March 2008 and will end in March 2011.

KYOTO will construct a language-independent information system for a specific domain (environment, ecology and biodiversity) anchored in a language-independent ontology that is linked to wordnets in seven languages. For each language, information extraction and identification of lexicalized concepts with ontological entries will be carried out by text miners (Kybots). The mapping of language-specific lexemes to the ontology allows for cross-linguistic identification and translation of equivalent terms. KYOTO is developing a wiki infrastructure for enabling long-range knowledge sharing and transfer across many languages and cultures, addressing the need for global and uniform transition of knowledge beyond the specific domains adressed in the project.

Semantic interoperability in KYOTO is achieved by defining the words and expressions in each language through a shared ontology. The KYOTO ontology will be formal language-independent representation of entities that will be used for inferencing and reasoning. The Wiki environment will help the users to agree on the meaning of the concepts of interest, to share their knowledge and to relate the terms and expressions in their language to this knowledge. This process is guided by automatic acquisition of terms and meanings from the textual documents provided by the users, and through automatic definition extraction techniques which will provide glosses for the acquired terms. The collaborative system will help the users review and edit all acquired information, with a special focus on achieving consensus but also for different views and interpretations across languages and cultures. The users can maintain their own system over time and work towards interoperability by fine-tuning their specifications or adding linguistically and culturally diverse groups.

The Wiki environment also generates formal knowledge representations from the conceptual modeling. These representations are not shown to the user directly; computer software will extract detailed information and

[*] This work has been supported by the KYOTO project which is funded under project number 211423 in the 7^{th} Frame Work in the area of Digital Libraries: FP7-ICT-2007-1, Objective ICT-2007.4.2: Intelligent Content and Semantics

[1] http://www.kyoto-project.eu/

[2] Co-funded by EU - FP7 ICT Work Programme 2007 under Challenge 4 - Digital libraries and Content, Objective ICT-2007.4.2 (ICT-2007.4.4): Intelligent Content and Semantics (challenge 4.2).

ISSN: 1135-5948

facts from the document collection in the group. The extraction process will use the agreed-upon ontological patterns and their relation to the words and expressions in each language so that the information can be interpreted in the same way across these languages and cultures. Likewise, the KYOTO system functions as an information and knowledge sharing platform. The system aims to establish cross-linguistic and cross-cultural communication and to support building and maintaining the system by groups of people in a shared domain and area of interest.

Currently, we completed the specification and design phase and we are integrating the first versions of the system components. In the project, we will be working on a restricted set of languages: English, Dutch, Italian, Spanish, Basque, Simplified Mandarin Chinese and Japanese. We also will apply the system to the domain of the environment and specifically to the topic of ecosystem services, a global phenomenon with different linguistic and cultural interpretations. Nevertheless, the system is designed in such a way that it can be used for any language and can be applied to any domain.

Most domain acquisition systems in the semantic web community model each domain separately and restrict the system to a single language or a limited set of languages. They also require knowledge engineers and language-technology experts to do the modeling. The KYOTO system, by contrast, is specifically designed to build global and cross-cultural consensus about the meaning and interpretation of domain-specific language. It tries furthermore to overcome the technology gap between users and system builders. The users are given control over the engineering task on a level that they can understand and that can be directly implemented for their community. As such it is an open system that can be extended and maintained by the users themselves without requiring skills in knowledge engineering or language technology. The main challenges of the project are:

- Automatic term and concept mining techniques should be of sufficient quality and have sufficient semantic depth, so that the data are useful for experts in the domain who are not trained in knowledge engineering and language technology;
- The users should be able to relate the terms across languages and cultures so as to agree on definitions and share them;
- Terms and concepts should be anchored to generic language databases and ontologies to provide interoperability and sharing to people outside the domain but in the same language communities;
- The term databases and their definitions in the ontology should enable extraction of sufficiently useful information and facts from text repositories for all the related languages, while at the same time the information should be of sufficient quality and depth;
- The interpretation of the information and facts should be the same across the different languages and cultures;
- The users should be able to specify the information and facts of their interest without having to access the complex underlying knowledge structure through a handful of textual examples from which the system abstracts the relevant underlying patterns;

The project just completed the design and specification phase. Currently, the first prototypes are being developed. An early version will become available in early 2009. The current website includes, among others, papers, deliverables, presentations and demos. For instance, an early baseline retrieval system that allows one to search in over 15,000 documents in 4 different languages: English, Dutch, Italian and Spanish. The current search system does not yet exploit the results of KYOTO but carries out a standards keyword based search. It allows one to fill in a complex environmental issue and to try to compile an answer through retrieval actions. All the searches and their success are logged, as are the answer that is compiled.