Facilitating Non-expert Users of the KYOTO Platform: the TMEKO Editing Protocol for Synset to Ontology Mappings

Roxane Segers[1], Piek Vossen[2]

[1] VU University Amsterdam, FEW, Department of Computer Science, De Boelelaan 1081A, 1081 HV Amsterdam, The Netherlands

[2] VU University Amsterdam, Department of Humanities, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands

E-mail: rh.segers@cs.vu.nl, p.vossen@let.vu.nl

Abstract

This paper presents the general architecture of the TMEKO protocol (Tutoring Methodology for Enriching the Kyoto Ontology) that guides non-expert users through the process of creating mappings from domain wordnet synsets to a shared ontology by answering natural language questions. TMEKO will be part of a Wiki-like community platform currently developed in the Kyoto project (http://www.kyoto-project.eu). The platform provides the architecture for ontology based fact mining to enable knowledge sharing across languages and cultures. A central part of the platform is the Wikyoto editing environment in which users can create their own domain wordnet for seven different languages and define relations to the central and shared ontology based on DOLCE. A substantial part of the mappings will involve important processes and qualities associated with the concept. Therefore, the TMEKO protocol provides specific interviews for creating complex mappings that go beyond subclass and equivalence relations. The Kyoto platform and the TMEKO protocol are developed and applied to the environment domain for seven different languages (English, Dutch, Italian, Spanish, Basque, Japanese and Chinese), but can easily be extended and adapted to other languages and domains.

1. Introduction

Experts have tremendous knowledge about their domain concepts but not the means for modeling this knowledge in a way that ensures reusability across languages and cultures. The need for reusing and exchanging knowledge is especially pressing within the environment domain since specific environmental issues are seldom restricted to single countries; a decrease in a migration bird population in a certain area might be related to changed hunting regulations in another part of the world.

Knowledge about environmental issues is stored in large amounts of document collections that are now only partially accessible through keyword search. The drawbacks of this approach are obvious: search results for a decrease in species in a certain area are limited, since only a few fragments with the specific keywords will be shown. Also, the keywords might appear in different places in a text and not be related at all. Relevant information in other languages is not easily accessible either. To ease the search for relevant information and to enhance information sharing between different languages, the Kyoto project is developing a community platform for modeling knowledge and finding facts across languages and cultures (Vossen et al., 2008). The platform operates as a Wiki and establishes semantic interoperability across languages for the environment domain by creating domain wordnets for seven languages that are interlinked through a shared DOLCE based ontology (Masolo et al., 2003).

Users are able to upload documents that will be processed in the Kyoto language processing pipeline that includes lemmatizing, syntactic parsing, creation of dependency trees, word sense disambiguation, named entity recognition and ontology tagging on word sense level. Each of these modules adds specific layers to the Kyoto Annotation Format (KAF) (Bosma et al., 2009), a LAF based text annotation format (Ide & Romary, 2003). The KAF annotated documents are the source for the term extractor (Tybot) that extracts relevant concepts from the texts in hierarchical structures (Bosma et al., 2010). The resulting termdatabases for each language are part of the Kyoto knowledge base that also comprises a SKOS converted Species2000 database [1], the seven generic wordnets, and the shared ontology. All these components (except the ontology) are presented to the users of the Wikyoto editing platform as an input for creating a domain wordnet.

The terms extracted from the documents and the species in the Species2000 database are disambiguated and partially aligned with synsets in the generic wordnets. In this way, each new synset in the domain wordnets hierarchy is linked to a synset in the generic wordnets by traversing the hierarchy. By this alignment, the existing mappings from the generic wordnets to the ontology can be used to apply the ontological distinctions to the domain terms. As such, the platform allows for continuous updating and modeling of the vocabulary by the people in the community, while their domain wordnets remain anchored to the generic wordnets.

Knowledge is added to the documents by creating domain wordnets and additional mappings from the domain synsets to the central ontology. Domain specific terms can then be recognized and annotated with their synset ID and according ontological information in the KAF. Ontological patterns that express domain knowledge on an abstract level can then be applied to the processed texts and find relevant information that can be lexicalized in various ways in documents from different sources or written in different languages. These patterns (Kybots) will mine facts from the documents and store these facts in a fact database. This fact database allows for semantic search and directs the user to the actual document where this information was found. These different components of the Kyoto architecture are presented in figure 1.

[1] http://www.sp2000.org