Multilingual MoKi, or: how to manage multilingual ontologies in a Wiki

Mauro Dragoni[1], Chiara Ghidini[1], Alessio Bosca[2]

[1] FBK-irst, Via Sommarive 18 Povo, I-38123, Trento, Italy
[2] Celi s.r.l., Via S.Quintino 31, I-10131, Torino, Italy

Abstract. In this paper we describe an extension of the MoKi tool able to support the management of multilingual ontologies. The multilingual features of MoKi are based on an integration with Dictionary Based Translation ad Machine Translation technologies. Also the collaborative features of MoKi are used to support the interaction between domain experts in order to discuss and agree on the translations of the terms to be used in each language.

1 Research Background

The construction of multilingual ontologies has become an important objective for organizations working in multilingual environments. As described in [3], obtaining multilingual ontologies is a complex activity which requires to tackle a number of problems spanning from the translation of the labels and description associated to a given ontology entity to the adaptation of the ontology to a concrete language and cultural community.

In this paper we describe an extension of the MoKi tool [4] able to support: (i) an automatic translation of labels and description associated to a given ontology entity, and (ii) the collaboration between domain experts in order to reach an agreement on the terms to be used in each language. This extension of MoKi is produced in the context of the Organic.Lingua EU project[3] where a multilingual version of an Organic Agriculture ontology is produced, starting from an English version, in order to support content classification and information retrieval in at least 16 different languages (Estonian, Slovenian, Romanian, Turkish, English, Spanish, Greek, German, Hungarian, Russian, Bulgarian, Hindi, French, Armenian, Dutch, Norwegian).

The first part of the extension is based on the integration of MoKi with a translation service able to provide translations for both individual labels and longer textual descriptions of a given entity. The second part of the extension takes advantage of the collaborative, wiki-style, characteristics of MoKi. In the following we better elaborate on these two aspects.

Language technologies and ontology translation The translation of a domain specific ontology such as the Organic Agriculture ontology used in the Organic.Lingua project requires the usage of different translation techniques. In fact, the translation of such ontologies concerns both the translation of labels, usually composed by (single or multi)

[3] http://www.organic-lingua.eu/

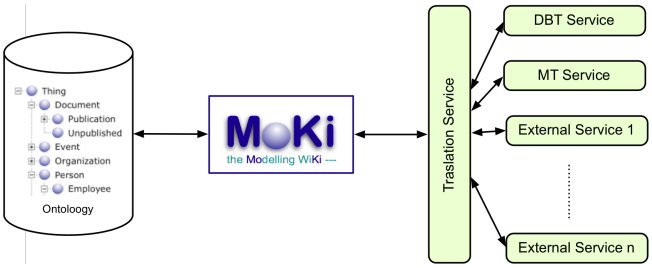

Fig. 1: MoKi and the translation services.

words, and the translation of longer textual descriptions that can be associated to a given entity. To take into account the specificity of the domain (e.g., the agro-ecology domain), an adaptation of the translation models should also be adopted. For this reason MoKi is integrated with a translation service able to automatically re-direct a translation request to three different translation sub-services (see Figure 1): (i) a dictionary based translation (DBT) sub-service provided by Celi s.r.l. used to translate the ontology labels; (ii) a machine translation (MT) sub-service provided by Xerox for full-text translation and adapted to the organic agriculture domain for 8 language pairs; and (iii) a connection to external translation sub-services (such as Google translate) for the pairs of languages not supported in (ii).

Collaborative agreement on term translation Given the complexity of translating domain specific ontologies, translations often need to be checked and agreed upon by a community of experts. This is especially true when ontologies are used to represent terminological standards which need to be carefully discussed and evaluated. To support this collaborative activity we foresee the usages of the wiki-style features of MoKi, expanded with the possibility of assigning specific translations of ontology entities to specific experts who need to monitor, check, and approve the suggested translations.

2 The multilingual MoKi Tool

MoKi[4] is a collaborative MediaWiki-based [8] tool for modeling ontological and procedural knowledge. The main idea behind MoKi is to associate a wiki page, containing both unstructured and structured information, to each entity of the ontology and process model to support on-line collaboration between members of the modeling team, including collaboration between domain experts and knowledge engineers. An extensive description of the MoKi tool and of its main features can be found in [5, 4].

In the context of the Organic.Lingua EU-Project, MoKi enables users (domain experts) to manage the editing of both the Organic.Lingua ontology and the Learning Object Metadata (LOM) ontology, as well as their translations. Here we briefly describe the MoKi main features. with a particular emphasis on the multilingual components and the technologies used to implement the DBT and MT translation services.

[4] See also `http://moki.fbk.eu`

2.1 The MoKi Features

The version of MoKi presented in this paper is equipped with five groups of functionalities, described below, that can be accessed via a wikis style menu bar.

Ontology Import and Export This set of functionalities support the automatic import and export of knowledge in and from MoKi (see [4]). Here we have added a filtering criterion that enables the export of the ontology in a specific (set of) language(s).

Ontology Editor This set of functionalities enables the management of the ontology (see [4]) and of its translation.

Interface Translator This set of functionalities is used to add a new language to the MoKi interface and manage the translation of the labels that define the interface. It is based on the multi-lingual features of MediaWiki, extended with the translation features connected with MoKi.

Ontology Translator This set of functionalities manage the translation operations required by MoKi. When a translation is requested (of an entity label or description), these functionalities contact the Translation service in order to obtain one. Also, when a new language is added to the ontology, the Ontology Translator retrieves, for each entity described in the ontology, the translations of their labels and descriptions. Then, new assertions, representing the relationships between the entities and their new translations, are added into the ontology.

Approval and Discussion This set is composed of two parts: a first part that monitors the activities that are performed on the entity pages, and a second part that manages the discussions associated with each entity page. The module implemented in MoKi permits to manage the approval requests that are automatically generated when changes are carried out on entities. Indeed, when a user updates the translation of a label or a description, MoKi generates an approval request that is automatically received by the users in charge to review the translations of a particular language. Besides the possibility of approving the change, they are able to use the discussion facility of MoKi for discussing about the correctness of the new translation.

2.2 The Translation Features

The version of MoKi presented in this paper is connected, via the Translation Service, to two main types of translation features: a DBT service and a MT service.

The DBT service This consist of a language identifier, morphological analyzers and a dictionary translation lookup components; a short overview of these components is presented in the followings.

Language Identifier A Language Identifier is a software module that is necessary whenever the language of a given textual resource (i.e. queries, metadata) is not explicitly known. The approach used here implements three different algorithms: a Character N-gram based strategy, a Word frequency based strategy, and a Function word based strategy [1].

Morphological Analyzer A Morphological Analyzer is a software module capable of associating morphological features (grammatical category, tense, number, gender, etc.) to terms as well as performing tokenization, lemmatization, decompounding, multi-word detection and POS (part of speech) tagging. The service is based on proprietary and open source solutions (see [2, 9]).

Translation Dictionaries Translation Dictionaries are mappings between terms in different languages. The functionality provided by this software component consists in retrieving all the translations in a target language available for a given term. The translations are performed by using the lemmas of terms. If no suitable translations are found in the dictionary, the module searches for translations of the term's lemma via a predefined bridge language.

The MT service This service is tailored for full-text translation as opposed to the translation of keywords or terms. The underlying translation model is a Phrase-based Statistical Translation model [7]. The service is currently available for the following 8 language pairs: English from/into German, French, Italian and Spanish. The underlying decoder (translator) used in the context of the Organic.Lingua project is based on Moses [6]. There is are a number of additional pre- and post- processing steps done with in-house (Xerox) tools which interact with the decoder such as de-compounding (for German), Named Entities Recognition, Brackets Management, Truecasing (recovering the case of the lowercased translations produced by the decoder), and so on. The translation models are adapted to the Organic.Lingua domain using a combination of in-domain (agriculture) and out-of-domain (Europarl, Wikipedia) parallel and monolingual texts. In-domain resources include: multilingual abstracts and titles extracted from bibliographical records on agricultural science and technology provided by FAO and INRA, sub-corpora of JRC-Aquis[5] comprising documents annotated with domain-related EUROVOC categories (agriculture, forestry and fisheries, agri-foodstuffs), as well as terminological data from the agricultural multilingual thesaurus AGROVOC[6].

3 System Demonstration

In the live demonstration we will present the MoKi features that are used to manage multilingual aspects of ontology in a collaborative way. We will first show how pages describing entities in a multilingual way look like, and we will describe the editing functionalities used to manage the translations of each entity label as well as their descriptions.

Then, we will demonstrate how the collaboration part may be used, by the domain experts, for discussing and approving the changes carried out on the entities. As regards the visualization functionalities, we will demonstrate how to get different overviews of the translated models, in particular we will show:

- how to display (and edit) in a tree-view the taxonomy or partonomy hierarchy of the element of the domain model in different languages;

[5] http://langtech.jrc.it/JRC-Acquis.html
[6] http://aims.fao.org/website/AGROVOC-Thesaurus/sub

– how to change the language used for visualizing the ontology in each view;
– how to compare different translations of the ontology and to perform quick edit operations.

We will show how it is possible to add a new language to the ontology and how to initialize the translations by invoking the machine translation service. Then, we will demonstrate how it is possible to add/update the languages used in the MoKi interface.

Finally, we will demonstrate how to export the ontology deployed in MoKi with the possibility to filter the exported axioms through language criteria.

4 Conclusions

In this demonstration we have presented a version of MoKi which supports the management of multilingual ontologies in a collaborative way. We are currently improving the tool in several directions, which span from including services for exposing the ontology, improving support for evolving the ontology, and implementing a module for the alignment between the content described in MoKi pages and external ontologies.

Acknowledgments. Organic.Lingua is funded under the ICT Support Program of the EU Commission (Grant Agreement Number 270999).

References

1. A. Bosca and L. Dini. Language identification strategies for cross language information retrieval. Clef Working Notes, 2010.
2. A. Bosca, L. Dini, M. Kouylekov, and M. Trevisan. Linguagrid: a network of linguistic and semantic services for the italian language. To be published in LREC 2012 conference, 2012.
3. Mauricio Espinoza, Asunción Gómez-Pérez, and Eduardo Mena. Enriching an ontology with multilingual information. In *Proceedings of the 5th European semantic web conference (ESWC'08)*, ESWC'08, pages 333–347. Springer, 2008.
4. Chiara Ghidini, Marco Rospocher, and Luciano Serafini. Moki: a wiki-based conceptual modeling tool. In *ISWC 2010 Posters & Demonstrations Track: Collected Abstracts*, volume 658 of *CEUR Workshop Proceedings (CEUR-WS.org)*, pages 77– 80, Shanghai, China, 2010.
5. Chiara Ghidini, Marco Rospocher, and Luciano Serafini. Conceptual modeling in wikis: a reference architecture and a tool. In *The Fourth International Conference on Information, Process, and Knowledge Management (eKNOW2012)*, 2012.
6. Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: open source toolkit for statistical machine translation. In *ACL '07: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics, 2007.
7. Philipp Koehn, Franz J. Och, and Daniel Marcu. Statistical phrase based translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT/NAACL)*, 2003.
8. Wikimedia Foundation. Mediawiki. http://www.mediawiki.org.
9. H. Schmid. Probabilistic part-of-speech tagging using decision trees. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK, 1994.