Proceedings of the International Conference on
Information Technologies (InfoTech-2017)
20-21 September 2017, Bulgaria

MULTILINGUAL ONTOLOGIES AND ENGLISH-BULGARIAN ONTOLOGY DEVELOPMENT

Tatyana Ivanova Ivanova

College of Energy and Electronics, TU-Sofia,
tiv72@abv.bg
Bulgaria

Abstract: In this paper we make a short survey of the approaches for development of multilingual ontologies. Our main goal is to find appropriate approach for development of multilingual ontologies, including Bulgarian language terminology. We propose a collaborative methodology for development of English-Bulgarian bilingual ontologies by usage of information extraction from e-learning textual content, linguistic resources and services.
Key words: bilingual ontology, multilingual ontology development, ontology translation methodology.

1. INTRODUCTION

Recently, the needs of ontologies in Bulgarian and multilingual ontologies have been increased because of their possible use into web-based applications (as search engines, RSS feeders, web services), query and answer systems, e-learning, etc. Unfortunately, it is almost impossible to find good ontology, containing Bulgarian-language terminology.

Most of the accessible in the internet ontologies are developed using English language terminology. Specialized search engines for ontologies as Watson (http://watson.kmi.open.ac.uk/WatsonWUI/), Swoogle (http://swoogle.umbc.edu/2006/), OntoSearch (http://www.ontosearch.com/) return only small number (less than 10) of ontologies, having labels or comments in Bulgarian. Big ontology repositories as Bioportal (http://bioportal.bioontology.org/), OBOfoundry (http://www.obofoundry .org/ontology/po.html), Protege ontology library (https://protegewiki.stanford.

edu/wiki/Protege_Ontology_Library#OWL_ontologies) also does not store ontologies, having labels in Bulgarian.

Most of the ontologies, developed by Bulgarian authors and covering specific Bulgarian domains also use English language terminology. For example, [1] presents ontological model of the Bulgarian folklore knowledge, including the semantics of the phenomena of Bulgarian traditional culture, but all the labels in this ontology are written only in English language. So, this Bulgarian folklore domain ontology is also using lexical items only in English and will not be usable in applications, that works with Bulgarian terminology.

The aim of this paper is to analyze existing approaches for development of bilingual or multilingual ontologies and answer the question "How we can ensure or stimulate inclusion of Bulgarian language terminology as an obligatory element of ontology development, evolution and maintenance process?".

The rest of the paper is organized as follows:

Section 2 describes related works. Section 3 discusses available semantic resources and possibilities for usage of e-learning materials and learner participation in ontology translation to Bulgarian language. Section 4 describes collaborative methodology for translation of English-language ontologies to Bulgarian, using linguistic resources and e-learning textual content. Section 5 gives some concluding remarks.

2. RELATED WORK

After comprehensive literature analysis we classify approaches for building bilingual or multilingual ontologies in the following categories (Fig. 1): Manual approaches; semi-automatic and automatic approaches. As Manual development requires many human efforts, and automated development has insufficient quality, most of well-working approaches are semi-automatic. They use some methods or technologies for automated Multilanguage label generation or suggestion followed by checking and acceptance by human developers. According to the used methods these approaches are based on Information Retrieval, Machine translation, ontology mapping, Controlled Natural Languages (CNL) (see Fig. 1). According to the involved humans the approaches are based on expert involvement or voluntary user collaboration.

2.1. Manual approaches

Some efforts are done recently for manual translation of ontology labels and /or comments in several languages, as multilingual ontologies are needed in many application areas. The "Language Technology for eLearning" (LT4EL) [2] project for example integrates multilingual semantic knowledge in a Learning Management

System to enhance the management, distribution and especially the cross-lingual retrieval of learning material. In this project a language-independent domain-ontology with lexicons of eight languages linked to it is developed and used. Some Learning objects of the English, German, Dutch, Polish, Portuguese, Romanian, Bulgarian and Czech languages have been annotated with concepts from this ontology. In this project manual approach for ontology development is used. Computer science concepts were formalised and appropriate classes and properties in the domain ontology ware created. Then each concept was mapped to synsets in the OntoWordNet version of Princeton WordNet. The mapping was performed via the two main relations of equivalence and hypernymy/hyponymy. Every lexical unit can be mapped to several concepts, and in this case its ontology representation is a disjunction of these concepts. Ontology concept can be mapped to a lexical unit or to a phrase.

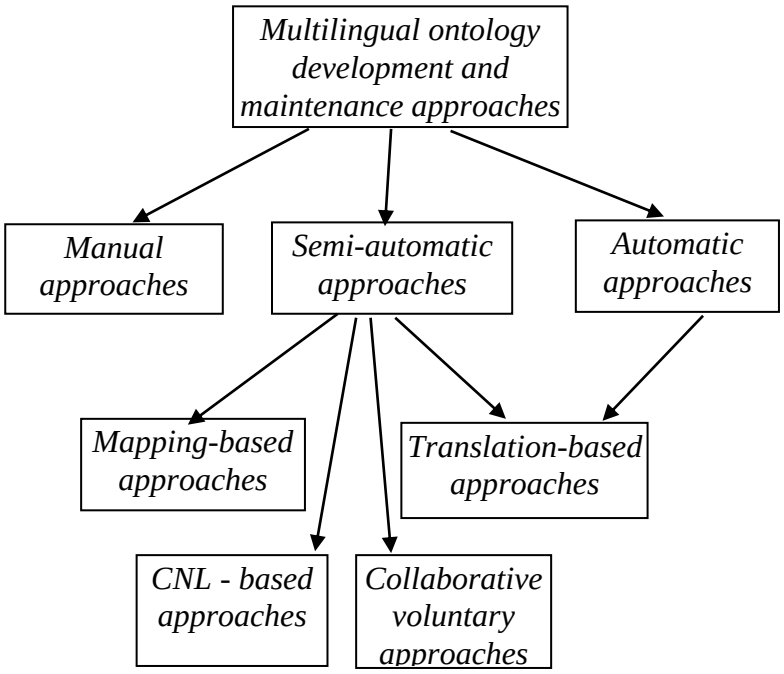

Fig. 1. Classification of the multilingual ontology development and maintenance approaches

Manual development can ensure high-quality of the developed ontology, but requires many human experts' efforts and this is very expensive. Moreover most of ontologies are related to the projects, for which they are developed, or represent context –dependent knowledge and are usable in other projects or contexts only after valuable changes.

2.2. Collaborative approach

The collaborative approach includes ontology component definitions based on common agreement. Easy-to use tools are of the great importance for involving many users in collaborative ontology-maintenance process.

Workbench platform for collaborative multilingual ontological knowledge is proposed in [3]. It supports construction and maintenance of the knowledge by end-user community. This platform includes three main ontological knowledge authoring tools: two ontology extraction tools, the Automatic Thai Ontology Construction and Maintenance tools (ATOM), the Semi-automatic Computational Lexicon Construction (KULEX), and one ontology integration tool. The ontology extraction uses three types of corpora: unstructured, structured, and semi-structured textual resources. The ontology-learning technique, both for taxonomic and non-taxonomic relations is also used.

A collaborative platform that facilitates the management of diversity in language and knowledge across cultures, via development of localized domain ontologies is proposed in [4]. It develops and uses multi-language resources. In this project a neat separation between natural language and formal ontological language is made and this is a fundamental feature to be able to manage diversity in knowledge, related to different languages. The platform provides an interactive and user-friendly web environment that allows geographically distributed linguistic and domain experts to contribute in a collaborative manner, based on following a collaborative development methodology, that uses the notions of collaborative objects and collaborative workflows, which specifies the development processes, user roles, and access rights. The platform supports collaboration both in the development and validation phases and uses linguistic resources as a set of collaborative objects.

Semantic Wiki tools combine wiki systems with Semantic Web technologies like RDF and OWL. Ontology development Semantic Wiki based systems become familiar to many domain experts, and in such a way both experts and users can be involved in the development process. An Example of such popular tool is OntoWiki [5], a knowledge base which provides visual representation of domain ontologies as information maps. These maps then are represented as web accessible pages, interlinked to related digital resources.

Possible conflicts are one of the main problems in collaborative development. In CofficientMakna [6] the collaborative development is augmented with the use of an argumentation ontology that formalizes the arguments exchanged between participants (issues, ideas and discussions). CofficientMakna provides a reasoning mechanism that can be used collaboratively by all participants in the development process.

Semantic wiki-based tools do not propose rich ontology-building and evaluation capabilities, but they are collaborative, accessible to many users (including domain

experts), and can combine ontology development and semi-structured textual multilingual linked data and knowledge. Usage of automated ontology-development methods in background is very important for supporting collaborative development.

2.3. Controlled natural languages (CNL) and multilingual ontologies

A controlled natural language (CNL) is a strictly defined fragment of a natural language [7]. The aim of Semantic web-oriented controlled natural languages is to equip domain specialists with an expressive easy to learn knowledge representation languages that are close to natural languages and are fully processable by a computer. These languages support easy ontology development in a way, close to knowledge modeling, using natural languages.

One of the most frequently-used CNLs is RABBIT [8]. It is developed and used for the communication between domain experts and ontology engineers to create ontologies. Research on CNL is turning its attention towards multilingual controlled languages. ACE CNL [9] is recently developed to support easy ontology development for several European languages. It is a subset of Standard English designed for knowledge representation and technical specifications, and constrained to be unambiguously machine readable

Grammatical Framework (GF) is a framework for multilingual grammar. It is developed for supporting Machine Translation of controlled natural languages. The GF libraries now contain grammars for around 20 natural languages, including Bulgarian. The GF 3.2 distribution comes with large coverage lexicons for Bulgarian, English, Finnish, Swedish and Turkish [10]. Combined with the corresponding resource grammars these lexicons make it possible to analyze almost free text.

2.4. Automated ontology label translation approach

While CNLs are mainly for ontology development, and collaborative approach can be used both in development and label translation, there are also approaches for direct translation of ontology labels or comments. Ontology localization consists of adapting ontology to a concrete language and cultural community. A system, named LabelTranslator, that automatically localizes ontologies, is proposed in [11]. LabelTranslator takes as input an ontology whose labels are described in a source natural language and automatically obtains the most probable translation into some target natural languages of each ontology label. To do this, the system uses a translation service from/into English, German, or Spanish based on linguistic resources such as lexical databases, bilingual or multilingual dictionaries and thesauruses (as EuroWordNet GoogleTranslate, Wiktionary) and terminologies. The system takes as input a list of words, discovers their semantics in run-time and obtains a list of senses, deals with the possible semantic overlapping among senses.

[11] Also proposes Compositional Method to Translate Compound Labels. Translation precision and recall highly depends on the used resources, domain and language and are between 30% and 80%. Bulgarian is not supported by this system.

One of the main reasons for the low translation quality is the natural language ambiguity. Many domain-specific lexical elements can be ambiguous as language elements, but are not ambiguous in fixed domain. For example, the term "ontology" is ambiguous, as it is used in philosophy and computer science, but in every one of these domains it is clearly defined. There are also terms, that are ambiguous in it's specific domain. Examples of ambiguous terms in computer science domain are the terms header, word, port, point. They have more than one meaning in this domain.

The main problems, related to the usage of lexicons are related to:

- Higt polysemy, which induce problems in automation of ontology learning and mapping.
- Low coverage. Some domain-specific words do not exist or some domain-specific senses (as ontology learning) are not presented in lexicons

2.5. Mapping – based approach

There are two main models for multilingual ontology development: one, aimed to enrichment of the one ontology with multilingual information (by adding multilingual labels and comments), and the other, aiming to map ontologies, having the same formal model and domain, and using labels in different languages. Net project [12] (EuroWordNet) for example consists in building language-specific WordNets independently from each other, and trying in a second phase to find correspondences between them. As adding multilingual information makes ontologies bigger and it's querying - slower, some resent multilingual ontology – development approaches are oriented to develop small multilingual ontologies (covering specific domains) and create mappings between these and upper level ontologies, containing more general terminology [13].

2.6. Combined approaches

Combined usage of several automated translation approaches and collaborative ones, supported by appropriate tools can improve the quality of the translation results. For example, [14] describes a semantic wiki system with an underlying controlled natural language grammar implemented in Grammatical Framework (GF). The wiki content is restricted to a well-defined subset of Attempto Controlled English (ACE), and facilitates a precise bidirectional automatic translation between ACE and language fragments of a number of other natural languages, making the wiki content accessible multilingually. The developed wiki environment allows users to build collaboratively, edit, query and view OWL knowledge bases via a userfriendly multilingual natural language interface.

3. AVAILABLE SEMANTIC RESOURCES AND POSSIBILITIES FOR USAGE OF E-LEARNING MATERIALS AND LEARNER PARTICIPATION IN ONTOLOGY TRANSLATION TO BULGARIAN

After our analysis of the approaches for multilingual ontology development and possibilities to use them for including Bulgarian language terminology we have made several remarks about available resources and prerequisites, related to Bulgarian languages.

Semantic, linguistic and human resources for translation into Bulgarian:

There are web-based and freely available lexical databases and linguistic resources (as WordNet, FrameNet, BabelNet, MultiWordNet and EuroWordNet), but only small number of them support Bulgarian language. BabelNet [15] is a multilingual semantic network (supporting Bulgarian language) that integrates lexicographic and encyclopedic knowledge from WordNet and Wikipedia. It was automatically constructed using machine translations. BabelNet provides a graphical user interface, known as BabelNetXplorer [16] that allows the users to visually explore the knowledge repository. BabelNet can be downloaded or used online. Several Bulgarian lexicons and grammatical resources have been developed, but most of them are not freely available in the internet. Quality of Goggle and other web service-based translation services is low in the scientific domains fields. Wiki-based scientific content in Bulgarian is also restricted. So, semantic and linguistic resources that can be used for mashie translation to Bulgarian are restricted (especially in the close scientific domains fields)

Many CNLs are English – based and additional efforts are needed for embedding them in collaborative graphical interfaces.

There are not so many Bulgarian knowledge experts and most of them prefer to develop ontologies, using English terminology because of the low popularity of the Bulgarian language in the Semantic Web area.

E-learning textual resources are good for terminology and knowledge extractions as they are well structured, contain clear definitions, explanations, examples. In many learning domains (for example electronic, computer science, informatics) important concepts are presented both in Bulgarian and English languages.

Our main idea is to implement English-Bulgarian ontology development or translation tools as part of the collaborative learning process to ensure it’s usage by learners. Learners’ knowledge can be classified as general domain and science-independent, General science-related, general domain – related, and closely-related to the specific learning domain. Similar classification and organization in layers according to the knowledge specificity is used in the ontology development area. Therefore, we will discuss the idea to develop mapped system of bilingual ontologies every of which representing related knowledge, having different specificity.

4. BILINGUAL ENGLISH – BULGARIAN ONTOLOGY DEVELOPMENT METHODOLOGY

Following the idea for development of the layered system of mapped multilingual ontologies, discussed above, we propose there layered system of mapped ontologies (Fig. 2) as a basis for bilingual ontology development in the context of collaborative e-learning system. This type of layered system is typical not only for human knowledge, but also in ontology development and usage domains. In the Semantic Web and ontology domains usually general knowledge and domain-specific knowledge are stored in different ontologies to support ontology reuse and faster working of ontological systems. And in wide domain ontology usually some terms, specific for close subdomains are missing. Ontologies, representing common general and scientific general knowledge, needed for understanding learning content also should be translated and represent the upper layer in our system.

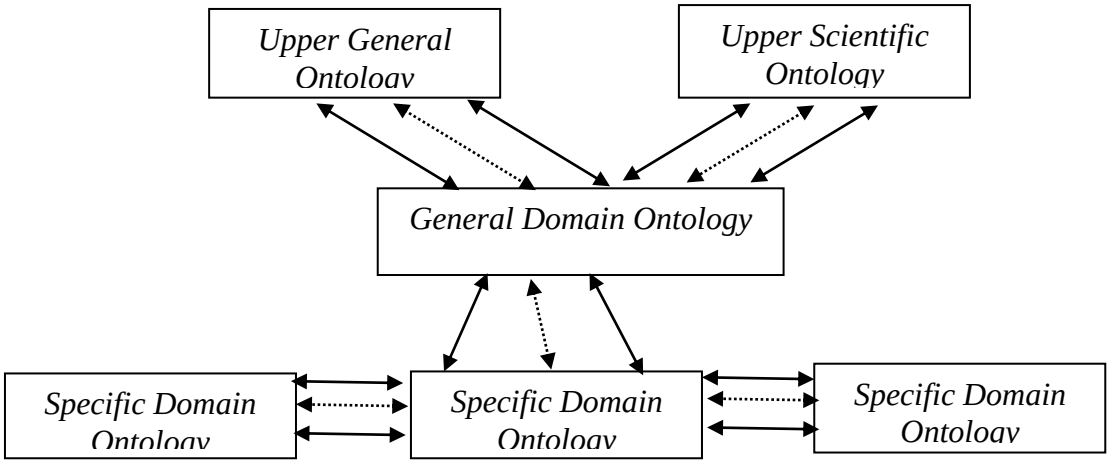

Fig.2. The system of mapped ontologies in for translation in e-learning

So we propose methodology for enriching and translating systems of mapped ontologies (not independent ontologies) as such system is a good structure for knowledge representation maintenance, evolution and reuse. During the translation process mappings between ontologies can be used to represent the context of mapped terms and in such a way they can support resolving of some ambiguity problems. Translated ontologies can be used independently one by another, or as parts of complex ontological system.

Our methodology for translating mapped system of English language ontologies into Bulgarian is intended for item label translation and include following main steps:

1. Searching and downloading from the internet English language terminology ontologies, related to the learning domain on the levels, shown in fig. 2.

2. Discovering initial mappings between founded ontologies and building initial version of the mapped ontologies of the layered mapped system in fig. 2.

These mappings can be performed manually or by usage of ontology mapping tools.

3. Selecting needed linguistic resources for ontology translation (English-Bulgarian thesauruses, lexicons, web services). Various sources can be used for various domains.

4. Integrating all the resources listed in steps 1, 2, 3 in the e-learning system.

5. Development of the ontology translation tool that uses the selected resources to generate suggestions for adding some changes or translation into Bulgarian of the English language labels in the mapped ontologies. These suggestions should be accepted or rejected by learners in collaborative environment. Experts also may participate in the ontology translation process.

6. Incremental evolution of the layered system of mapped ontologies is performed to enrich or support actual versions of the mapped ontologies in correspondence with the learning content or its possible changes.

7. Periodic consistency checking of the evolved ontological system.

Step 6 includes not only translation, but ontology changes and possible addition of new mappings. The new ontology items can be added in correspondence with the terminology, represented in the e-learning content. This process requires the usage of natural (Bulgarian) language text analysis and terminology extraction tools. The language resources, used in the translation process are also good resources for ontology enrichment. Mappings and ontologies in the system can be used as a context for term disambiguation. And e-learning system users can be asked to accept or reject proposed ontology changes. They can also be asked to solve possible logical contradictions during ontology management process.

5. CONCLUSION

Ontologies, containing Bulgarian language labels are needed to support many tacks in Bulgarian context in a way that ontologies, having English labels are used in tacks having English language context. Automatic translation of ontology labels has very low quality because of insufficient linguistic resources, natural language ambiguities and translation tool's immaturity. There are no adequate language and knowledge experts, or they are not motivated to do manual ontology translations to Bulgarian. So, we believe that collaborative ontology translation methods, based on adequate automation and user-friendly interfaces are needed.

In this paper we propose a methodology for collaborative translation of ontological systems that can work as part of e-learning system. This type translation (as other automated translations approaches) requires good linguistic and natural language processing resources and tools. The main strength of our approach is learner involvement in the ontology translation process during the learning process.

Such system will support learning and at the same time produce good ontology enrichment and translation results.

REFERENCES

[1] Luchev, D. et al. (2008). Use of knowledge technologies for presentation of bulgarian folklore heritage semantics. *International Journal "Information Technologies and Knowledge" Vol.2 / 2008, pp.307-313.*

[2] Lemnitzer, et al. (2008). Using a domain-ontology and semantic search in an e-learning environment. *Innovative techniques in instruction technology, e-learning, e-assessment, and education,* pp. 279-284.

[3] Suktarachan, M., et al. (2008). Workbench with Authoring Tools for Collaborative Multi-lingual Ontological Knowledge Construction and Maintenance. In LREC.

[4] Tawfik, A., et al. (2014) Collaborative Platform for Multilingual Ontology Development. *International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering* Vol:8, No:12, 2014

[5] Auer, S., et al. (2006). OntoWiki–a tool for social, semantic collaboration. In The Semantic Web-ISWC 2006 (pp. 736-749). Springer Berlin Heidelberg.

[6] Tempich, C., et al. (2007). Argumentation-based ontology engineering. IEEE Intelligent Systems, 22(6), 52-59.

[7] Angelov K. The Mechanics of the Grammatical Framework. PhD thesis, Chalmers University of Technology, 2011.

[8] Denaux et.Al.Rabbit to OWL:Ontology Authoring with a CNL Based Tool With a NL based tool CNL (2009), LNCS, vol.5972, pp.246 -264.

[9] Schwitter, R. (2010, August). Controlled natural languages for knowledge representation. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters (pp. 1113-1121). Association for Computational Linguistics.

[10] Angelov, K. (2011). The Mechanics of the Grammatical Framework. Chalmers University of Technology.

[11] Espinoza, M., et al. (2008). Enriching an ontology with multilingual information. The Semantic Web: Research and Applications, 333-347.

[12] Vossen, P. (1998) Introduction to EuroWordNet. Computers and the Humanities 32 (2-3) pp. 73–89

[13] Dragoni, M. (2015, October). Multilingual ontology mapping in practice: a support system for domain experts. In International Semantic Web Conference (pp. 169-185). Springer International Publishing.

[14] Kaljurand, K., T. Kuhn, (2013). A multilingual semantic wiki based on Attempto Controlled English and Grammatical Framework. In Extended Semantic Web Conference (pp. 427-441). Springer Berlin Heidelberg.

[15] Navigli, R., SP. Ponzetto, (2010) BabelNet: Building a very large multilingual semantic network. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010, pp. 216–225.

[16] Navigli, R., SP. Ponzetto, (2012) BabelNetXplorer: A Platform for Multilingual Lexical Knowledge Base Access and Exploration. Proceedings of International World Wide Web Conference (IW3C2), Lyon,France, 16-20 April 2012.