MSW 2010

Proceedings of the 1st International Workshop on the
Multilingual Semantic Web

Collocated with the 19th International World Wide Web Conference (WWW2010)

Raleigh, North Carolina, USA, April 27, 2010.

Sponsored by:

Endorsed by:

Preface

Although knowledge processing on the Semantic Web is inherently language-independent, human interaction with semantically structured and linked data will remain inherently language-based as this will be done preferably by use of text or speech input, in many different languages. Semantic Web development will therefore be increasingly concerned with knowledge access to and generation in/from multiple languages.

Multilinguality can be therefore considered an emerging challenge to Semantic Web development and to its global acceptance – across language communities around the world. The MSW workshop was concerned with discussion of new infrastructures, architectures, algorithms, etc., whose goal is to enable an easy adaptation of Semantic Web applications to multiple languages, addressing issues in representation, extraction, integration, presentation, and so on. This workshop brought together researchers from several distinct communities, including natural language processing, computational linguistics, human-computer interaction, artificial intelligence and the Semantic Web.

There were 12 submissions to the workshop, from which the program committee accepted 4 as full papers, 2 as short papers, and 2 as position papers. Taking into account only the full and short papers the selection rate amounts to 50%. The accepted papers cover a variety of topics regarding the use of multilingual ontologies with different purposes that range from information extraction and data querying to user profile enrichment, as well as multilingualism modeling issues, controlled languages or multilingual ontology mapping for the future Semantic Web. The MSW Workshop program also included a keynote talk by Professor Sergei Nirenburg entitled “The Ontology-Lexicon Space for Dealing with Meaning in Natural Language(s): Extraction, Manipulation, Acquisition, Uses, Needs, Lessons Learned, Hopes”.

We would like to thank the authors for providing the content of the program. We would like to express our gratitude to the program committee for their work on reviewing papers and providing interesting feedback to authors. We would also like to thank Behrang Qasemizadeh for his technical support. And finally, we kindly acknowledge the European Union and the Science Foundation Ireland for their support of the workshop through research grants for Monnet (FP7-248458) and Lion2 (SFI/08/CE/I1380), and to the European Project FlareNet (ECP-2007-LANG-617001) and the German Project Multipla (DFG-38457858) for their endorsement.

Paul Buitelaar
Philipp Cimiano
Elena Montiel-Ponsoda

April, 2010

Table of Contents

MSW 2010 Organization

Organizing Committee

Paul Buitelaar
Unit for Natural Language Processing, DERI - National University of Ireland, Galway
http://www.paulbuitelaar.net/

Philipp Cimiano
Semantic Computing Group, Cognitive Interaction Technology Excellence Cluster (CITEC)
Bielefeld University, Germany
http://www.cimiano.de

Elena Montiel-Ponsoda
Ontology Engineering Group, Departamento de Inteligencia Artificial
Universidad Politécnica de Madrid, España
emontiel(at)delicias.dia.fi.upm.es

Program Committee

Guadalupe Aguado de Cea, Universidad Politécnica de Madrid - Artificial Intelligence Department, Spain
Nathalie Aussenac-Gilles, IRIT, Toulouse - Knowledge Engineering, Cognition and Cooperation, France
Timothy Baldwin, Univ. of Melbourne - Language Technology Group, Australia
Roberto Basili, Universita Tor Vergata, Rome - Artificial Intelligence group, Italy
Chris Bizer, Freie Universität Berlin - Web-based Systems Group, Germany
Francis Bond, NICT - Language Infrastructure Group, Japan
Christopher Brewster, Aston University - Operations and Information Management Group, UK
Dan Brickley, Vrije Universiteit & FOAF project, the Netherlands
Nicoletta Calzolari, ILC-CNR - Computational Linguistics Institute, Italy
Manuel Tomas Carrasco Benitez, European Commission, Luxembourg
Jeremy Carroll, TopQuadrant Inc., USA
Key-Sun Choi, KAIST - Semantic Web Research Center, South-Korea
Thierry Declerck, DFKI - Language Technology Lab, Germany
Aldo Gangemi, ISTC-CNR - Semantic Technology Laboratory, Italy
Asuncion Gómez Pérez, Universidad Politécnica de Madrid - Artificial Intelligence Department, Spain
Gregory Grefenstette, Exalead, France
Siegfried Handschuh, DERI, Nat. Univ. of Ireland, Galway - Semantic Collaborative Software Unit, Ireland
Michael Hausenblas, DERI, Nat. Univ. of Ireland, Galway - Data Intensive Infrastructures Unit, Ireland
Ivan Herman, CWI & W3C, the Netherlands
Chu-Ren Huang, Hong Kong Polytechnic University - Dept. of Chinese and Bilingual Studies, Hong Kong
Antoine Isaac, Vrije Universiteit - Web and Media Group, the Netherlands
Ernesto William De Luca, Universität Magdeburg - Data and Knowledge Engineering Group, Germany
Paola Monachesi, Universiteit Utrecht - Institute of Linguistics, the Netherlands
Sergei Nirenburg, University of Maryland - Institute for Language and Information Technologies, USA

Alessandro Oltramari, ISTC-CNR - Laboratory for Applied Ontology & Univ. of Padua, Italy
Jacco van Ossenbruggen, CWI - Semantic Media Interfaces & VU - Intelligent Systems, the Netherlands
Wim Peters, University of Sheffield - Natural Language Processing group, UK
Laurette Pretorius, University of South Africa - School of Computing, South-Africa
James Pustejovsky, Brandeis University – CS Dept., Lab for Linguistics and Computation, USA
Maarten de Rijke, Univ. van Amsterdam - Information and Language Processing Systems, the Netherlands
Felix Sasaki, W3C deutsch-österr. Büro & FH Potsdam, Germany
Martin Volk, Universität Zürich - Institute of Computational Linguistics, Switzerland
Piek Vossen, Vrije Universiteit - Dept. of Language, Cognition and Communication, the Netherlands
Yong Yu, Shanghai Jiao Tong University - APEX Data & Knowledge Management Lab, China

Ontologies for Multilingual Extraction

Deryle W. Lonsdale
Linguistics & English Lang.
Brigham Young University
lonz@byu.edu

David W. Embley
Computer Science
Brigham Young University
embley@cs.byu.edu

Stephen W. Liddle
Information Systems
Brigham Young University
liddle@byu.edu

ABSTRACT

In our global society, multilingual barriers sometimes prohibit and often discourage people from accessing a wider variety of goods and services. We propose multilingual extraction ontologies as an approach to resolving these issues. Our ontologies provide a conceptual framework for a narrow domain of interest. Grounding narrow-domain ontologies linguistically enables them to map relevant utterances and text to meaningful concepts in the ontology. Our prior work includes leveraging large-scale lexicons and terminology resources for grounding and augmenting ontological content [14]. Linguistically grounding ontologies in multiple languages enables cross-language communication within the scope of the various ontologies' domains. We quantify the success of linguistically grounded ontologies by measuring precision and recall of extracted concepts, and we can gauge the success of automated cross-linguistic-mapping construction by measuring the speed of creation and the accuracy of generated lexical resources.

1. INTRODUCTION

Though English has so far served as the principal language for Internet use (with currently 28.7% of all users), its relative importance is rapidly diminishing. Chinese users, for example, comprise 21.7% of Internet users and their growth in numbers between 2000 and 2009 has been 1,018.7%; the growth in Spanish users has been 631.3% over the last decade. Since more people want to access web information in more languages, this poses a substantial challenge and opportunity for research and business organizations whose interest is in providing multilingual access to web content.

The BYU Data Extraction research Group (DEG)[1] has worked for years on tools—such as its Ontology Extraction System (OntoES)—to enable access to web content of various types: car advertisements, obituaries, clinical trial data, and biomedical information. The group to date has focused on English web data, while understanding the eventual need to extend OntoES to other languages. This appears to be an opportune time for our group to enter the area of multilingual information extraction and show how the DEG infrastructure is poised to make significant contributions in this area as it has already has in extracting English information.

There are currently a few efforts in the area of multilingual information extraction. Some focus on very narrow domains, such as technical information for oil drilling and exploration in Norwegian and English. Others are more general but involve more than two languages, such as accessing European train system schedules. The U.S. government (NIST TREC), the European Union (7th Framework CLEF), and Japan (NTCIR) all have initiatives to help further the development and evaluation of multilingual information retrieval and data extraction systems. Of course, Google and other companies interested in web content and market share are enabling multilingual access to the Internet.

Almost all of the existing efforts involve a typical scenario that might include: collecting a query in the user's language, translating that query into the language of the web pages to be searched, locating the answers, and then returning the relevant results to the user or to someone who can help the user understand their content. This approach is fraught with problems since machine translation (MT), a core component in the process, is still a developing technology.

For reasons discussed below, we believe that our approach has technical and linguistic merit, and can introduce a fresh perspective on multilingual information extraction. Our ontology-based techniques are ideal for extracting content in various languages without having to rely directly on MT. By carefully developing the knowledge resources necessary, we can extend DEG-type processing to other languages in a modular fashion.

2. THE ONTOLOGY-BASED APPROACH

2.1 Extraction Ontologies

Just over a decade ago, the BYU Data-Extraction research Group (DEG) began its work on information extraction. In a 1999 paper, DEG researchers described an efficacious way to combine ontologies with simple natural language processing [5].[2] The idea is to de-

[1]This work was funded in part by U.S. National Science Foundation grants for the TIDIE (IIS-0083127) and TANGO (IIS-0414644) projects.

WWW2010, April 26-30, 2010, Raleigh, North Carolina.

[2]Recently, others have begun to combine ontologies with

clare a narrow domain ontology for an application of interest and augment its concepts with linguistic recognizers. Coupling recognizers with conceptual modeling turns a conceptual ontology into an extraction ontology. When applied to data-rich semi-structured text, an extraction ontology recognizes linguistic elements that identify concept instances for the object and relationship sets in the ontology's conceptual model. We call our system *OntoES*, *Ontology-based Extraction System*.

Consider, for example, a typical car ad. Its content can be modeled with a conceptual ontology such as that shown in Figure 1. With linguistic recognizers added for concepts such *Make*, *Model*, *Year*, *Price*, and *Mileage*, the domain ontology becomes an extraction ontology.

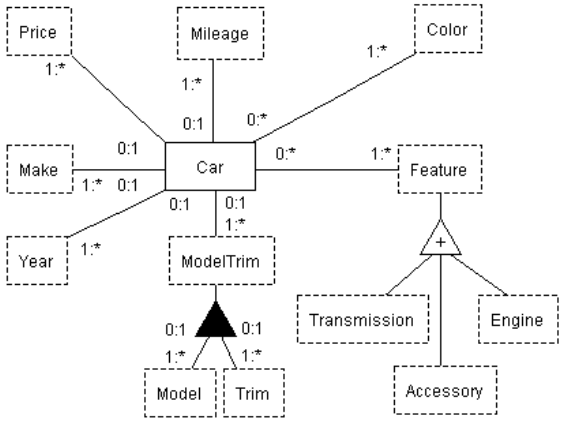

Figure 1: Extraction Ontology for Car Ads.

We have developed a form-based tool [15] that helps users to develop ontologies including declaring recognizers and associating them with ontological concepts. It also permits users to specify regular expressions that recognize traditional value phrases for car prices such as "$15,900", "7,595", and "$9500"—with optional dollar signs and commas. Users can also declare additional recognizers for other expected price expressions such as "15 grand". To help make recognizers more precise, users can declare exception expressions, left and right context expressions, units expressions, and even keyword phrases such as "MSRP" and "our price" to help sort out various prices that might appear. Figure 2 shows snippets from recognizer declarations for car ads data.

Applying the recognizers of all the concepts in a car-ads extraction ontology to a car ad annotates, extracts, and organizes the facts from that ad. The result is a machine-readable cache of facts that users can query or use to perform data analysis or other automated tasks.[3]

To verify that a carefully designed extraction ontology for car ads can indeed annotate, extract, and organize facts for query and analysis, DEG researchers have

natural language processing [11, 2]. The combination has been called "linguistically grounding ontologies."

[3]See http://deg.byu.edu for a working online demonstration of the system.

```
Price
    internal representation: Integer
    external representation: \$[1-9]\d{0,2},?\d{3}
                             | \d?\d [Gg]rand | ...
    context keywords: price|asking|obo|neg(\.|otiable)|...
    ...
    LessThan(p1: Price, p2: Price) returns (Boolean)
    context keywords: (less than|<|under|...)\s*{p2} |...
    ...
Make
    ...
    external representation: CarMake.lexicon
    ...
```

Figure 2: Sample Recognizer Declarations for Car Ads.

conducted experiments with hundreds of car ads from various on-line sources containing thousands of fact instances. In one experiment, when an existing OntoES car ads ontology was hand-tuned on a corpus of 100 development documents and then tested on an unseen corpus of about 110 car ads, the system extracted 1003 attributes with with recall measures of 94% and precision measures nearing 100% [6].

Recently, DEG researchers have experimented with information extraction in Japanese. Figure 3 shows an OntoES extraction ontology that can extract information from Japanese car ads analogous to the English one shown earlier. The concept names are in Japanese as are the regular-expression recognizers. Yen amounts range from 10,000 yen to 9,999,999 yen whereas dollar amounts range from $100 to $99,999. The critical observation is that the structure of the Japanese ontology is identical to the structure of the English ontology.

This type of ontology-based matching across languages at the lexical level indicates a possible strategy for providing a cross-linguistic bridge through concepts rather than relying on traditional means of translation. Similar approaches have been tried in such areas as machine translation (e.g. [4]) and cross-linguistic information retrieval [12].

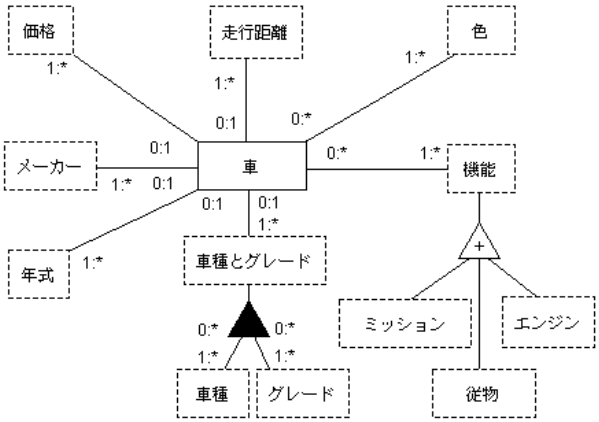

Figure 3: Japanese Extraction Ontology for Car Ads.

As currently implemented, OntoES extraction ontologies can "read" and "write" in any single language. The car-ad examples here are in English and Japanese, but extraction ontologies work the same for all languages. To "read" means to recognize instance values for ontological concepts, to extract them, and to appropriately link related values together based on the associated conceptual relationships and constraints. To "write" means to list the facts recorded in the ontological structure. Having "read" a typical car ad, OntoES might write:

Year: 1984
Make: Dodge
Model: W100
Price: $2,000
Feature: 4x4
Feature: Pickup
Accessory: 12.5x35" mud tires

In addition, based on the constraints, OntoES knows and can write several meta statements about an ontology. Examples: "an *Accessory* is a *Feature*" (white triangles denote hyponym/hypernym is-a constraints); "*Trim* is part of *ModelTrim*" (black triangles denote meronym/holonym is-part-of constraints), "*Car* has at most one *Make*" (the participation constraint 0:1 on *Car* for *Make* denotes that *Car* objects in car ads associate with *Make* names between 0 and 1 times, or "at most once").

As currently implemented, however, OntoES cannot read in one language and write in another. This cross-linguistic ability to read in one language and then translate to and write in another language is the essence of our multilingual-oriented development. For example, we expect to be able to read the price in yen from a Japanese car-ad and write "Price: $24,124" and to read the Kanji symbols for the make and write "Make: Mitsubishi". To assure this level of functionality, we need to encode unit or currency conversion routines for values like *Price* and to encode cross-linguistic lexicons for named entities such as *Make*. In principle, encoding this cross-linguistic mapping is currently possible, but represents a fair amount of manual effort. We are currently finding ways to largely automate this mapping. In addition, we are adding two other capabilities to the system that will similarly enhance extraction and query processing: compound recognizers and patterns.

Compound recognizers allow OntoES to directly recognize ontological relationships beyond simple concepts. For a query like: "Find Nissans for sale with years between 1995 and 2005.", we need to recognize each of the years as well as the *between* constraint that relates them. Our previous work has implemented compound recognizers for operators in free-form queries [1], but we now seek to linguistically ground these types of ontological relationships.

Patterns will allow OntoES to identify and extract from structured text. For example, car ads often appear as a table with *Price* in one column, *Year* in another column, and *Make* and *Model* in a third column. Detecting patterns in documents will allow OntoES to apply specialized extraction rules and likely improve extraction accuracy. By extending our work with table patterns [8], we expect to fully exploit patterns in text.

2.2 Multilingual Mappings

We are extending in a principled way the cross-linguistic effectiveness of our OntoES system by adapting it for use in processing data-rich documents in languages other than English. Though the OntoES system was originally designed to handle English-language documents, it was implemented according to standard web-related software engineering principles and best practices: version control, integrated development enviroments, standardized data markup and encoding (XML, RDF, and OWL), Unicode character representation, and tractability (SWRL rules and Pellet-based reasoning). Consequently, we anticipate that internationalization of the system should be relatively straightforward, not requiring wholesale rewrites of crucial components. This should allow us to handle web pages in any language, given appropriate linguistic knowledge sources. Since OntoES does not need to parse out the grammatical structure of webpage text, only lower-level lexical (word-based) information is necessary for linguistic processing.

The system's lexical knowledge is highly modular, with specific resources encoded as user-selectable lexicons. The information used to build up existing content for the English lexicons includes a mix of implicit knowledge and existing resources. Some lexicon entries were created by students during class and project work; other entries were developed from existing lexical resources (e.g. the US Census Bureau for personal names, the World Factbook for country names, Ethnologue for language names, etc.). We are developing analogous lexicons for other languages, and adapting OntoES as necessary to accommodate them in its processing. As was the case for English, this involves some hand-crafting of relevant material, as well as finding and converting existing data sources in other languages for targeted types of lexical information. Often this is relatively straightforward: for example, WordNet is a sizable and important component for English OntoES, and similar and compatible resources exist for other languages. However, we also need to rely on linguistic knowledge and experience to find, convert, and implement appropriate cross-linguistic lexical resources.

In the realm of cross-linguistic extraction systems, OntoES has a clear advantage. We claim that ontologies, which lie at the crux of our extraction approach, can serve as viable interlinguas. We are currently substantiating this claim. Since an ontology represents a conceptualization of items and relationships of interest (e.g. interesting properties of a car, information needed to set up a doctor's appointment, etc.), a given ontology should be appropriate cross-linguistically with perhaps occasionally some slight cultural adaptation. For example, in our prior work on extraction from obituaries [5] we found that worldwide cultural and dialect differences were readily apparent even in English material. Certain terms for events like "tenth day kriya", "obsequies", and "cortege" were found only in English obituaries announcing events outside of America. Since our lexical resources serve as a "grounding" of the lowest-level concepts from ontologies with the lexical content of the web

pages, substituting one language's lexicon for another's provide OntoES a true cross-linguistic capability.

2.3 Ongoing Work

Our current work involves several separate but related tasks. We are locating annotated corpora in other languages amenable for evaluation purposes, and collecting and annotating interesting multilingual web material of our own. We are also developing prototype lexicons and recognizers for these target languages. Of course, our work requires us to develop and adapt prototype ontologies for target languages for sample concepts in data-rich domains.

In addition, we are enhancing extraction ontologies by enabling them to (1) explicitly discover and extract relationships among object instances of interest, and (2) discover patterns of interest from which they can more certainly identify and extract both object instances and relationship instances of interest. This involves devising, investigating, designing, coding, and evaluating algorithms for compound recognizers and for pattern discovery and patterned information extraction.

Finally, we are evaluating system performance using standard metrics and gold-standard annotated data.

3. CONCLUSION

Though an interesting effort in its own right, we expect our multilingual extraction work to also contribute to our larger effort to create a Web of Knowledge [7, 9]. Our research centers around resolving some of the tough technical issues involved in a community-wide effort to deploy the semantic web [16] and in concert with efforts at Yahoo!, Google, and elsewhere to extract information from the web and integrate it into community portals to enable community members to better discover, search, query, and track interesting community information [3, 10, 13]. Multilingual extraction ontologies have the far-reaching potential to play a significant role as semantic-web work finds its way into mainstream use in global communities.

4. REFERENCES

[1] M. Al-Muhammed and D. Embley. Ontology-based constraint recognition for free-form service requests. In *Proceedings of the 23rd International Conference on Data Engineering (ICDE'07)*, pages 366–375, Istanbul, Turkey, 2007.

[2] P. Buitelaar, P. Cimiano, P. Haase, and M. Sintek. Towards linguistically grounded ontologies. In *Proceedings of the 6th European Semantic Web Conference (ESWC'09)*, pages 111–125, Heraklion, Greece, 2009.

[3] P. DeRose, W. Shen, F. Chen, A. Doan, and R. Ramakrishnan. Building structured web community portals: A top-down, compositional, and incremental approach. In *Proceedings of the 33rd Very Large Database Conference (VLDB'07)*, pages 23–28, Vienna, Austria, 2007.

[4] B. J. Dorr. *Machine Translation: A view from the lexicon*. MIT Press, Cambridge, MA, 1993.

[5] D. Embley, D. Campbell, Y. Jiang, S. Liddle, D. Lonsdale, Y.-K. Ng, and R. Smith. Conceptual-model-based data extraction from multiple-record web pages. *Data & Knowledge Engineering*, 31(3):227–251, 1999.

[6] D. Embley, D. Campbell, S. Liddle, and R. Smith. Ontology-based extraction and structuring of information from data-rich unstructured documents. In *Proceedings of the 7th International Conference on Information and Knowledge Management (CIKM'98)*, pages 52–59, Washington D.C., 1998.

[7] D. Embley, S. Liddle, D. Lonsdale, G. Nagy, Y. Tijerino, R. Clawson, J. Crabtree, Y. Ding, P. Jha, Z. Lian, S. Lynn, R. Padmanabhan, J. Peters, C. Tao, R. Watts, C. Woodbury, and A. Zitzelberger. A conceptual-model-based computational alembic for a web of knowledge. In *Proceedings of the 27th International Conference on Conceptual Modeling (ER08)*, pages 532–533, 2008.

[8] D. Embley, C. Tao, and S. Liddle. Automating the extraction of data from HTML tables with unknown structure. *Data & Knowledge Engineering*, 54(1):3–28, 2005.

[9] D. Embley and A. Zitzelberger. Theoretical foundations for enabling a web of knowledge. In *Proceedings of the 6th International Symposium on Foundations of Information and Knowledge Systems (FoIKS10)*, Sophia, Bulgaria, 2010.

[10] A. Halevy, P. Norvig, and F. Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, March/April 2009.

[11] L. Hunter, Z. Lu, J. Firby, W. B. Jr., H. Johnson, P. Ogren, and K. Cohen. OpenDMAP: An open source, ontology-driven, concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression. *BMC Bioinformatics*, 9(8), 2008.

[12] K. Kishida. Technical issues of cross-language information retrieval: A review. *Information Processing and Management: an International Journal*, 41:433–455, 2005.

[13] R. Kumar, B. Pang, R. Ramakrishnan, A. Tomkins, P. Bohannon, S. Keerthi, and S. Merugu. A web of concepts. In *Proceedings of the 2009 Symposium on Principles of Database Systems*, pages 1–12, Providence, RI, 2009.

[14] D. Lonsdale, D. W. Embley, Y. Ding, L. Xu, and M. Hepp. Reusing ontologies and language components for ontology generation. *Data & Knowledge Engineering*, 69:318–330, 2010.

[15] C. Tao, D. Embley, and S. Liddle. FOCIH: Form-based ontology creation and information harvesting. In *Proceedings of the 28th International Conference on Conceptual Modeling (ER 2009)*, pages 346–359, Gramado, Brazil, 2009.

[16] W3C (World Wide Web Consortium) *Semantic Web Activity Page*. http://www.w3.org/2001/sw/.

CLOVA: An Architecture for Cross-Language Semantic Data Querying

John McCrae
Semantic Computing Group, CITEC
University of Bielefeld
Bielefeld, Germany
jmccrae@cit-ec.uni-bielefeld.de

Jesús R. Campaña
Department of Computer Science and Artificial Intelligence
University of Granada
Granada, Spain
jesuscg@decsai.ugr.es

Philipp Cimiano
Semantic Computing Group, CITEC
University of Bielefeld
Bielefeld, Germany
cimiano@cit-ec.uni-bielefeld.de

ABSTRACT

Semantic web data formalisms such as RDF and OWL allow us to represent data in a language independent manner. However, so far there is no principled approach allowing us to query such data in multiple languages. We present CLOVA, an architecture for cross-lingual querying that aims to address this gap. In CLOVA, we make a distinction between a language independent *data layer* and a language independent *lexical layer*. We show how this distinction allows us to create modular and extensible cross lingual applications that need to access semantic data. We specify the search interface at a conceptual level using what we call a *semantic form specification* abstracting from specific languages. We show how, on the basis of this conceptual specification, both the query interface and the query results can be localized to any supported language with almost no effort. More generally, we describe how the separation of the lexical layer can be used with a principled ontology lexicon model (LexInfo) in order to produce application-specific lexicalisations of properties, classes and individuals contained in the data.

Categories and Subject Descriptors

H.5.m [**Information Interfaces and Presentation**]: User Interfaces; I.2.1 [**Artificial Intelligence**]: Applications and Expert Systems; I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods; I.2.7 [**Artificial Intelligence**]: Natural Language Processing

General Terms

Design, Human Factors, Languages

Keywords

Multilingual Semantic Web, Ontology Localisation, Software Architecture

1. INTRODUCTION

Data models and knowledge representation formalisms in the Semantic Web allow us to represent data without reference to natural language[1]. In order to facilitate the interaction of human users with semantic data, supporting language-based interfaces in multiple languages is crucial. However, currently there is no principled approach supporting the access of semantic data across multiple languages. To fill this gap, we present in this paper an architecture we call CLOVA (Cross-Lingual Ontology Visualisation Architecture) designed for querying semantic data in multiple languages. A developer of a CLOVA application can define the search interface independently of any natural language by referring to ontological relations and classes within a *semantic form specification* (SFS), which represents a declarative and conceptual representation of the search interface with respect to a given ontology. We have designed an XML-based language which is inspired by the Fresnel language [2] for this purpose. The search interface can then be automatically localised by the use of a lexicon ontology model such as LexInfo [4], enabling the system to automatically generate the form in the appropriate language. The queries to the semantic repository are generated on the basis of the information provided in the SFS and the results of the query can be localised using the same method as used for the localisation of the search interface. The CLOVA framework is generic in the sense that it can be quickly customised to new scenarios, new ontologies and search forms and additional languages can be added without changing the actual application, even at run time if we desire.

The paper is organised as follows. Section 2 describes state of the art on information access across languages and points out basic requirements for cross lingual systems. Section 3 describes the CLOVA framework for rapid development of cross-lingual search applications accessing semantic data. We conclude in Section 4.

2. RELATED WORK

Providing access to information across languages is an important topic in a number of research fields. While our work is positioned in the area of the Semantic Web, we discuss work related to a number of other research areas, including databases, cross-language information retrieval as well as ontology presentation and visualisation.

[1]This holds mainly for RDF triples with resources as subjects and objects. String data-type elements are often language-specific.

WWW2010, April 26-30, 2010, Raleigh, North Carolina.

2.1 Database Systems

Supporting cross-language data access is an important topic in the area of database systems, albeit one which has not received very prominent attention (see [8]). An important issue is certainly the one of character encoding as we need to represent characters for different languages. However, most of the current database systems support Unicode so that this issue is not a problem anymore. A more complex issue is the representation of content in the database in such a way that information can be accessed across languages. There seems to be no consensus so far on what the optimal representation of information would be such that cross-language access can be realised effectively and efficiently. One of the basic requirements for multilingual organisation of data mentioned by Kumaran et al. [8] is the following:

"The basic multilingual requirement is that the database system must be capable of storing data in multiple languages."

This requirement seems definitely too strict to us as it assumes that the representation of data is language-dependent and that the database is supposed to store the data in multiple languages. This rules out language-independent approaches which do not represent language-specific information in the database at all.
The following requirement by Kumaran et al. is one we can directly adhere to:

Requirement 1 (Querying in multiple languages)
Data must be queriable using query strings in any (supported) language.

In fact, we will refer to the above as *Requirement 1a* and add the following closely related *Requirement 1b*: *'The results of a query should also be presented in any (supported) language.'* Figure 1 summarises all the requirements discussed in this section. However, it does not strictly follow from this that the data should be stored in multiple languages in the database. In fact, it suffices that the front end that users interact with supports different languages and is able to translate the user's input into a formal (language-independent) query and localises the results returned by the database management system (DBMS) into any of the supported languages.

A further important requirement by Kumaran et al. we subscribe to is related to interoperability:

Requirement 2 (Interoperability)
The multilingual data must be represented in such a way that it can be exchanged across systems.

This feature is certainly desirable. We will come back to this requirement in the context of our discussion of the Semantic Web (see below). The next two requirements mentioned by Kumaran et al. are in our view questionable as they assume that the DBMS itself has built-in support for multiple languages:

- **String equality across scripts**: A Multilingual database system should support *lexical joins* allowing to join information in different tables even if the relevant attributes of the join are in different scripts.
- **Linguistic equivalences**: Multilingual database systems should support *linguistic joins* which exploit predefined mappings between attributes and values across languages. For example, we might state explicitly that the attributes *"marital status"* (in English) and *"Familienstand"* are equivalent and that the values *"married"* and *"verheiratet"* are equivalent.

In fact, those two requirements follow from Kumaran et al.'s assumption that the database should store the data in multiple languages. If this is the case then we certainly have to push all the cross-language querying functionality into the DBMS itself. This is rather undesirable from our point of view as every time a new language is added to the system, the DBMS needs to be modified to extend the linguistic and lexical equivalences. Further, the data is stored redundantly (once for every language supported). Therefore, we actually advocate a system design where the data is stored in a language-independent fashion and the cross-lingual querying functionality as well as result localisation is external to the DBMS itself, implemented as pre- and post-processing steps, respectively.

In fact, we would add the following requirement to any system allowing to access data across languages:

Requirement 3 (Language Modularity)
The addition of further languages should be modular in the sense that it should not require the modification of the DBMS or influence the other languages supported by the system.

As a consequence, the capability of querying data across languages should not be specific to a certain implementation of a DBMS but work for any DBMS supporting the data model in question.

One of the important issues in representing information in multiple languages is avoiding redundancy (see [6]). Hoque et al. indeed propose a schema to give IDs to every piece of information and then include the language information in a dictionary table. This is perfectly in line with Semantic Web data models (RDF in particular) where URIs are used to uniquely identify resources. Dictionaries can then be constructed expressing how the elements represented by the URIs are referred to across languages. This thus allows to conceptually separate the data from the dictionary. This is a crucial distinction that CLOVA also adheres to (see below).

2.2 Cross-language Information Retrieval

In the field of information retrieval, information access across languages has also been an important topic, mainly in the context of the so called *Cross-Language Evaluation Forum*[2] (see [10] for the proceedings of CLEF 2008). Cross-language information retrieval (CLIR) represents an extreme case of the so called vocabulary mismatch problem well-known from information retrieval. The problem, in short, is the fact that a document can be highly relevant to a query in spite of not having any words in common with the query. CLIR represents an extreme case in the sense that if a query and a document are in different languages, then the word overlap and consequently every vector-based similarity measure will be zero.

[2]http://www.clef-campaign.org/

In CLIR, the retrieval unit is the document, while in database systems the retrieval unit corresponds to the information units stored in the data base. Therefore, the requirements with respect to multilinguality are rather different for CLIR and multilingual database systems.

2.3 Semantic Web

Multilinguality has been so far an underrepresented topic in the Semantic Web field. While on the Semantic Web we encounter similar problems as in the case of databases, there are some special considerations and requirements. We will consider further important requirements for multilinguality in the context of the Semantic Web. Before, we introduce the crucial distinction between the **data layer (proper)** and the **lexical layer**. We will see below that the conceptual separation between the data and the dictionary is even more important in the context of the Semantic Web. According to our distinction, the data layer contains the application-relevant data while the lexical layer merely contains information about how the data is realised/expressed in different languages and acts like a dictionary. We note that this distinction is a conceptual one as the data in both layers can be stored in the same DBMS. However, this might not always be possible in a decentralised system such as the Semantic Web:

Requirement 4 (Data and Lexicon Separation)
We require a clear separation between the data and lexicon layer in the Semantic Web. The addition of further languages should be possible without modifying the data layer. This means that the proper data layer and the lexical layer are cleanly separated and data is not stored redundantly.

In the Semantic Web, the parties interested in accessing a certain data source are not necessarily its owners (in contrast to standard centralised database systems as considered by Kumaran et al.). As a corollary it follows that if a user requires access to a data source in language x he might not have the permission to enrich the data source by data represented in the language x.

A further relevant requirement in the context of the Semantic Web is the following:

Requirement 5 (Sharing of Lexica)
Lexica should be represented declaratively and in a form which is independent of specific applications such that it can be shared.

It is very much in the spirit of the Semantic Web that information should be interoperable and thus reusable beyond specific applications. Following this spirit, it seems desireable that (given that data representation is language-independent) the language-specific information how certain resources are expressed in various languages can be shared across systems. This can be accomplished by declaratively described lexica which can be shared.

Multilinguality has been approached in RDF through the use of its `label` property, which can assign labels with language annotations to URIs. The SKOS framework [9] further expands on this by use of `prefLabel`, `altLabel`, `hiddenLabel`. These formalisms are sufficient for providing simple representation of language information. However, as more complex lexico-syntactic information is required, in turn more complex representations are necessary. A more formal distinction of the "data layer" and "lexical layer" is provided by *lexicon ontology models* of which the most prominent models are the Linguistic Information Repository (LIR) and LexInfo (see [4]).

2.4 Ontology Presentation and Visualisation

Fresnel [2] is a display vocabulary that describes methods of data presentation in terms of *lenses* and *formats*. In essence the *lens* in Fresnel selects which values are to be displayed and the *format* selects the formatting applied to each part of the lens. This provides many of the basic tools for presenting semantic web data. However it does not represent multilinguality within the vocabulary and it is not designed to present a queriable interface to the data. There exist many forms of ontology visualisation methods through the use of trees, and other structures to display the data contained within the ontology, a survey of which is provided in [7]. These are of course focussed mainly on displaying the structure of the ontology and do not attempt to convert the ontology to natural language. Furthermore, for very large data sources, it is impractical to visualise the whole ontology at one time and hence we wish only to select a certain section of it and hence require a query interface to perform this task.

3. MULTILINGUAL ACCESS AND QUERYING USING CLOVA

CLOVA addresses the problem of realising localised search interfaces on top of language-independent data sources, abstracting the work flow and design of a search engine and providing the developer with a set of tools to define and develop a new system with relatively little effort. CLOVA abstracts lexicalisation and data storage as services, providing a certain degree of independence from data sources and multilingual representation models.

The different modules of the system have been designed with the goal of providing very specific, non-overlapping and independent tasks to developers working on the system deployment concurrently. User interface definition tasks are completely separated from data access and lexicalisation, allowing developers of each module to use different resources as required.

CLOVA as an architecture does not fulfil any of the aforementioned requirements (as they should be fulfilled by lexicalisation services), but provides a framework to fully exploit cross-lingual services meeting these requirements. The application design allows to separate conceptual representations from language dependant lexical representations, making user interfaces completely language independent in order to later localise them to any supported language.

3.1 System Architecture

The CLOVA architecture is designed to enable the querying of semantic data in a language of choice, while still presenting queries to the data source in a language-independent form. CLOVA is modular, reusable and extensible and as such is easily configured to adapt to different data sources, user interfaces and localisation tools[3].

[3]A Java implementation of CLOVA is available at `http://www.sc.cit-ec.uni-bielefeld.de/clova/`

Req. No	Implication	Status
Req. 1a	Querying in multiple languages	REQUIRED
Req. 1b	Result localisation in multiple languages	REQUIRED
Req. 2	Data interoperability	REQUIRED
Req. 3	Language modularity	REQUIRED
Req. 4a	Separation between data and lexical layer	DESIRED TO SUPPORT Req. 3
Req. 4b	Language-independent data representation	DESIRED TO AVOID REDUNDANCY
Req. 5	Declarative representation of lexica	DESIRED FOR SHARING LEXICAL INFORMATION

Figure 1: Requirements for multilingual organisation of data

Figure 2 depicts the general architecture of CLOVA and its main modules. The *form displayer* is a module which translates the semantic form specification into a displayable format, for example HTML. Queries are performed by the *query manager* and then the results are displayed to the user using the *output displayer* module. All of the modules use the *lexicaliser* module to convert the conceptual descriptions (i.e., URIs) to and from natural language. Each of these modules are implemented independently and can be exchanged or modified without affecting the other parts of the system.

We assume that we have a data source consisting of a set of properties referenced by URIs and whose values are also URIs or language-independent data values. We shall also assume that there are known labels for each such URI and each language supported by the application. If this separation between the lexical layer and the data layer does not already exist, we introduce elements to create this separation. It is often necessary to apply such manual enrichment to a data source, as it is not trivial to identify which strings in the data source are language-dependent, however we find that is often a simple task to perform by identifying which properties have language-dependent ranges, or by using XML's language attribute.

We introduce an abstract description of a search interface by way of XML called a *semantic form specification*. It specifies the relevant properties that can be queried by using the URIs in the data source, thus abstracting from any natural language. We show how this can be used to display a form to the user and to generate appropriate queries once he/she has filled in the form. The *query manager* provides a backend that allows us to convert our queries using information in the form into standard query languages such as SPARQL and SQL. Finally, we introduce a lexicalisation component, which is used to translate between the language-independent forms specified by the developer and the localised forms presented to the user. We describe a lexicaliser which builds on a complex lexicon model and demonstrate that it can provide more flexibility with respect to the context and complexity of the results we wish to lexicalise.

3.2 Modules

3.2.1 Semantic Form Specification

One of the most important aspects of the architecture is the *Semantic Form Specification* (SFS), which contains all the necessary information to build a user interface to query the ontology. In the SFS the developer specifies the ontology properties to be queried by the application via their URIs. This consists of a form for which we specify a *domain*, i.e., the class of objects we are querying as defined in the database

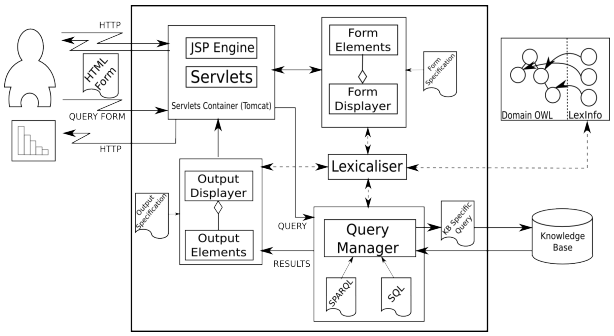

Figure 2: CLOVA general architecture

by an RDF `type` declaration or similar. If this is omitted we simply choose all individuals in the data source. The SFS essentially consists of a list of fields which are to be used to query the ontology. Each field contains the following information:

- **Name**: An internal identifier is used to name the input fields for HTML and HTTP requests.
- **Query output**: This defines whether this field will be included in these results. Valid values are *always*, *never*, *ask* (the user could decide wether to include the field in the results or not), *if_empty* (if the field has not been queried it is included in the output), *if_queried* (if the field is queried, it is included in the output) and *ask_default_selected* (the user decides, but as default the field will be shown).
- **Property**: represents the URI for the ontology property to be queried through the field. An indication of `reference=self` in place of a URI means that we are querying the domain of the search. Such queries are useful for querying the lexicalisation of the object being queried or limiting the query to a fixed set of objects.
- **Property Range**: We define a number of types (called *property ranges*) that describe the data that a field can handle. It differs from the data types of RDF or similar in that we also describe how the data should be queried as well. For example, while it is possible to describe both the revenue of a company and the age of an employee as integers in the database, it is not sensible to query revenue as a single value, whereas it is often useful to query age as a single value. These property ranges provide an abstraction of these properties in the data and thus support the generation of appropriate forms

and queries. The following property ranges are built-in into CLOVA:

- *String*, *Numeric*, *Integer*, *Date*: Simple data-type values. Note that *String* is intended for representing language-independent strings, e.g. IDs, not natural language strings. The numeric and date ranges are used to query precise values like "age" and "birth date".
- *Range*, *Segment*, *Set*: These are defined relative to another property range and specify how a user can query the property in question. *Range* specifies that the user should query the data by providing an upper and/or lower bound, e.g. "revenue", "number of employees". *Segment* is similar but requires that the developer divides the data up into pre-defined intervals. *Set* allows the developer to specify a fixed set of queriable values, e.g. "marital status".
- *Lexicalised Element*: Although we assume all data in the source is defined by URIs, it is obviously desirable that the user can query the data using natural language. This property range in fact allows to query for URIs through language-specific strings that need to be resolved by the system to the URI in question. The strings introduced into this field are processed by the lexicaliser to find the URI to which they belong which is then used in the corresponding queries. For example, locations can have different names in different languages, e.g. "*New York*" and "*Nueva York*", but the URI in the data source should be the same.
- *Complex*: A complex property is considered to be a property composed of other sub-properties. For example, searching for a "key person" within a company can be done by searching for properties of the person, e.g., "name", "birth place". This nested form allows us to express queries over the structure of an RDF repository or other data source.
- *Unqueriable*: For some data, methods for efficient querying cannot be provided, especially binary data such as images. Thus we defined this field to allow the result to still be extracted from the data source and included in the results.

The described property ranges are supported natively by CLOVA, but it is also possible to define new property ranges and include them in the SFS XML document. The appropriate implementation for a form display element that can handle the newly defined property range has to be provided of course (see Section 3.2.2).

- **Rendering Properties**: There is often information for a particular rendering that cannot be provided in the description of the property ranges alone. Thus, we allow for a set of context specific properties to be passed to the rendering engine. Examples of these include the use of auto-completion features or an indication of the type of form element to display, i.e. a *Set* can be displayed as a drop-down list, or as a radio button selection.

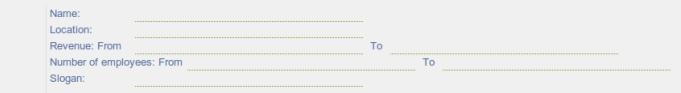

Figure 3: HTML form generated for a SFS document

The SFS document is in principle similar to the concept of a "lens" in the Fresnel display vocabulary [2] in that it describes the set of fields in the data that should be used for display and querying. However, by including more information about methods for querying the data, we provide a description that can be used for both presentation and querying of the data.

Example: Suppose that we want to build a small web application that queries an ontology with information about companies stored in an RDF repository. The application should ask for company names, companies' revenue, and company locations. The syntax of a SFS XML document for that application is shown below:

```
<!--xmlns:dbpedia="http://dbpedia.org/ontology/"-->
<form domain="dbpedia:Company">
    <fields>
        <field name="Name" output="ALWAYS">
            <property reference="self"/>
            <property-range>
                <lexicalised-property-range/>
            </property-range>
            <rendering context="html">
                <property name="autocompletion" value="yes"/>
            </rendering>
        </field>

        <field name="Location" output="ASK">
            <property uri="&dbpedia;Organisation/location"/>
            <property-range>
                <lexicalised-property-range/>
            </property-range>
        </field>

        <field name="Revenue"  output="ASK_DEFAULT_SELECTED">
            <property uri="&dbpedia;Organisation/revenue"/>
            <property-range>
                <ranged-property-range>
                   <continuous-property-range>
                       <min>0</min>
                   </continuous-property-range>
                </ranged-proprety-range>
            </property-range>
        </field>
    </fields>
</form>
```

3.2.2 Form Displayer

The form displayer consists of a set of *form display elements* defined for each property range. It processes the SFS by using these elements to render the fields in a given order. The implementation of these elements is dependent on the output method. The form display elements are rendered using Java code to convert the document to XHTML[4].

Figure 3 shows an example of rendering of an SFS which includes the fields in the example above. In this rendering the field "name" is displayed as a text field as it refers to the lexicalisation of this company. The location of a company for instance is represented as a text field. However, in spite of the fact that the data is represented in the data source as a language independent URI, the user can query by specifying

[4]The CLOVA project also provides XSLT files to perform the same task

the name of the resource in their own language (e.g., a German user querying "München" receives the same results as an English user querying "Munich"). Finally, the revenue is asserted as a continuous value which is queried by specifying a range and is thus rendered with two inputs allowing the user to specify the upper and/or lower bounds of their query. A minimum value on this range allows for client-side data consistency checks. In addition, check boxes are appended to fields in order to allow users to decide if the fields will be shown in the results, according to the output parameter in the SFS.

3.2.3 Query Manager

Once the form is presented to the user, he or she can fill the fields and select which properties he or she wishes to visualise in the results. When the query form is sent to the Query Manager, it is translated into a specific query for a particular knowledge base. We have provided modules to support the use of SQL queries using JDBC and SPARQL queries using Sesame [3]. We created an abstract query interface which can be used to specify the information required in a manner that is easy to convert to the appropriate query language allowing us to change the knowledge base, ontology and back end without major problems. The query also needs to be preprocessed using the *lexicaliser* due to the presence of language-specific terms introduced by the user which need to be converted to language independent URIs.

3.2.4 Output Displayer

Once the query is evaluated, the results are processed by the *output displayer* and an appropriate rendering shown to the user. The displayer consists of a number of display elements, each of which represents a different visualisation of the data, including not only simple tabular forms, but also graphs and other visual display methods. As with the form displayer, all of these elements are lexicalised in the same manner as the form displayer.

In general we might restrict the types of data that components will display as not every visualisation paradigm is suitable for any kind of data. For example, a bar chart showing foundation year and annual income would be both uninformative and difficult to display due to the scale of values. For this reason we provide an *Output Specification* to define the set of available display elements and sets of values they can display. These output specifications consist of a list of output elements described as follows:

- **ID**: Internal identifier of the output element displayed.
- **URI**: A reference to the output resource specified as a URI.[5]
- **Fields**: The set of fields used by this element. These should correspond by `name` to elements in the SFS.
- **Display properties**: Additional parameters passed to the display element to modify its behaviour. Some of these parameters include the possibility to ignore incomplete data, or to define the subtypes of a chart to display. These parameters are class dependant so that each output element has its own set of valid parameters.

[5]These can reference Java classes by linking to the appropriate class file or location in a JAR file

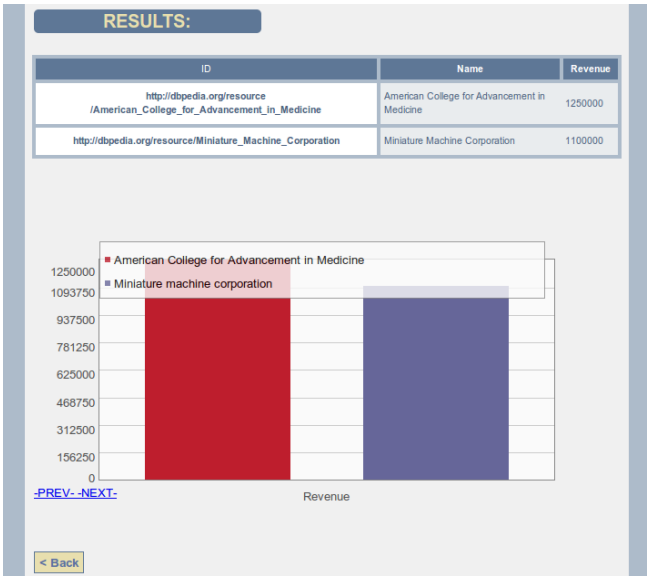

Figure 4: HTML result page for the example

The following output specification defines two output elements to show results.

```
<!-- xmlns:clova="jar:file:clova-html.jar!/clova/html/output/"
   xmlns:dbpedia="http://dbpedia.org/ontology/"-->
 <output>
    <elements>
        <element id="HTable" URI="&clova;HTableDisplayElement">
            <fields>
                <all/>
            </fields>
        </element>
        <element id="BarChart" URI="&clova;GraphDisplayElement">
            <fields>
                <field name="revenue"/>
            </fields>
            <display>
                <property name="Type" value="barChart"/>
            </display>
        </element>
    </elements>
</output>
```

The first element displays a table containing all the results returned by the query, while the second output element shows a bar chart for the property "Revenue". The HTML output generated for a given output specification containing the above mentioned descriptions is shown in Figure 4.

3.2.5 Lexicaliser

Simple lexicon models can be provided by language annotations, for example RDF's `label` and SKOS's `prefLabel`, and developing a lexicaliser is then as simple as looking up these labels for the given resource URI. This approach may be suitable for some tasks. However, we sometimes require lexicalisation using extra information about the context and would like to provide lexicalisation of more than just URIs, e.g. when lexicalising triples. While RDF labels can be attached to properties and individuals for instance, there is no mechanism that allows to compute a lexicalization for a triple by composing together the labels of the property and the individuals. This is a complex problem and we will leave a full investigation and evaluation of this for future work.

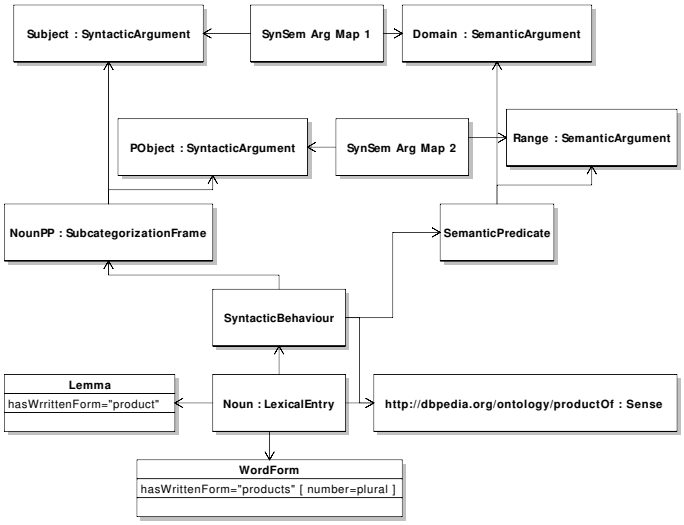

Figure 5: A simplified example of a LexInfo aggregate

Furthermore, it is often desirable to have fine control over the form of the lexicalisation, for example, the ontology label may be "company location in city". However, we may wish to have this property expressed by the simpler label "location". By using a lexicon ontology model we can specify the lexicalisation in a programmatic way, and hence adapt it to the needs of the particular query interface. For these reasons we primarily support lexicalisation through the use of the LexInfo [4] lexicon ontology model and its associated API[6], which is compatible with the LMF Vocabulary [5].

The LexInfo model:

A LexInfo model is essentially an OWL model describing the lexical layer of an ontology specifying how properties, classes and individuals are expressed in different languages. We refer to the task of producing language-specific representation of elements in the data source including triples as *lexicalisation* of the data. The corresponding LexInfo API organises the lexical layer mainly by defining so called *aggregates* which describe the lexicalisation of a particular URI, specifying in particular the lexico-syntactic behaviour of certain lexical entries as well as their interpretation in terms of properties, classes and individuals defined in the data. An aggregate essentially bundles all the relevant individuals of the LexInfo model needed to describe the lexicalization of a certain URI. This includes a description of syntactic, lexical and morphological characteristics of each lexicon entry in the lexicon. Indeed, each aggregate describes a lexical entry together with its lemma and several word forms (e.g. inflectional forms such as the plural etc.). The syntactic behaviour of a lexical entry is described through subcategorization frames making the required syntactic arguments explicit. The semantic interpretation of the lexical entry with respect to the ontology is captured through a mapping ("syn-sem argument map") from the syntactic arguments to the semantic arguments of a semantic predicate which stands proxy for an ontology element in the ontology. Finally the aggregate is linked through a `hasSense` link to the URI in the data layer it lexicalises. An example of an aggregate is given in figure 5. For details the interested reader is referred to [4].

[6]Available at `http://lexinfo.googlecode.com/`

LILAC:

In order to produce lexicalisations of ontology elements from a LexInfo model we use a simple rule language included with the LexInfo API called LILAC (LexInfo Label Analysis & Construction). A LILAC rule set describes the structure of labels and can be used for both generating the lexicon from labels and generating lexicalisations from the lexicon. In general we assume that lexicons are generated from some set of existing labels, which may be extracted from annotations in the data source, e.g., RDFS's `label`, from the URIs in the ontology or from automatic translations of these labels from another language. The process of generating aggregates from raw labels requires that first the part of speech tags are identified by a tagger such as TreeTagger. Then, the part-of-speech tagged labels are parsed using a LR(1)-based parser (see [1]). The API then handles these parse trees and converts them into LexInfo aggregates.

LILAC rules are implemented in a symmetric manner so that they can be used to both generate the aggregates in the lexicon ontology model (e.g. by analysing the labels of a given ontology) as well as lexicalise those aggregates.

A simple example rule for a label such as "revenue of" is:

```
Noun_NounPP -> <noun> <preposition>
```

This rule states that the lexicalisation of a `Noun_NounPP Aggregate` is given by first using the written form of lemma of the "noun" of the aggregate followed by the lemma of "preposition" of the aggregate. LILAC also supports the insertion of literal terms and choosing the appropriate word form in the following manner:

```
Verb_Transitive -> "is" <verb> [ participle,
    tense=past ] "by"
```

This rule can be used to convert a verb with transitive behaviour into a passive form (e.g., it transforms "eats" into "is eaten by").

LILAC can create lexicalisations recursively for phrase and similar, for example to lexicalise an aggregate for "yellow moon", the following rules are used. Note that in this cases the names provided by the aggregate class are not available so the name of the type is used instead:

```
NounPhrase -> <adjective> <NounPhrase>
NounPhrase -> <noun>
```

The process for lexicalisation proceeds as follows: for each ontology element (identified by a URI) that needs to be lexicalised, the LexInfo API is used to find the lexical entry that refers to the URI in question. Then the appropriate LILAC rules are invoked to provide a lexicalization of the URI in a given language.

As this process requires only the URI of the ontology element, by changing the LexInfo model and providing a reusable set of LILAC rules the language of the interface can be changed to any suitable form. It is important to emphasize that the LILAC rules are language-specific and thus need to be provided for each language supported.

Another issue is that we desire that our users are capable of searching for elements by their lexicalised form. LexInfo can support this as well. This involves querying the lexicon for

all lexical entries that have a word form matching the query and returning the URI that the lexical entry is associated to. Once we have mapped all language-specific strings to URIs, the query can be handled using the query manager as usual. For example if the user queries for "food" then the LexInfo model could be queried for all lexical entries that have either a lemma or word form matching this literal. The URIs referred to by this word can then be used to query the knowledge base. This means that a user can query in their own language and expect the same results, for example the same concept for "food processing" will be returned by an English user querying "food" and a Spanish user querying for "alimento" (part of the compound noun "Procesado de los alimentos").

3.3 CLOVA for company search

We developed a search interface for querying data about companies using CLOVA, which is available at `http://www.sc.cit-ec.uni-bielefeld.de/clova/demo`. For this application we used data drawn from the DBPedia ontology, which we entered into a Sesame store. We used the labels of the URIs to generate the lexicon model for English, and used the translations provided by DBPedia's wikipage links (themselves derived from WikiPedia's "other languages" links), to provide labels in German and Spanish. As properties were not translated in this way, the translations for these elements were manually provided. These translations were converted into a LexInfo model through the use of about 100 LILAC rules. About 20 of these rules were selected to provide lexicalisation for the company search application. In addition, we selected the form properties and output visualisations by producing a semantic form specification as well as an output specification. These were rendered by the default elements of the CLOVA HTML modules, and the appearance was further modified by specifying a CSS style-sheet. In general, the process of adapting CLOVA involves creating a lexicon, which could be a LexInfo model or a simpler representation such as with RDF's `label` property, and then producing the semantic form specification and output specification. Adapting CLOVA to a different output format or data back end, it requires implementing only a set of modest interfaces in Java.

4. CONCLUSION

We have presented an architecture for querying semantic data in multiple languages. We started by providing methods to specify the creation of forms, the querying of the results and presentation of the results in a language-independent manner through the use of URIs and XML specifications. By creating this modular framework we provide an interoperable language-independent description of the data, which could be used in combination with a lexicalisation module to enable multilingual search and querying. We then separated the data source into a language-independent data layer and a language-dependent lexical layer, which allows us to modularise each language and made the lexical information available separately on the semantic web. In this way we achieved all the requirements we set out in Figure 1. We described an implementation of this framework, which was designed to transform abstract specifications of the data into HTML pages available on the web and performed its lexicalisations by the use of LexInfo lexicon ontology models [4] providing fine control on the lexicalisations used in a particular context.

Acknowledgements

This work has been carried out in the context of the Monnet STREP Project funded by the European Commission under FP7, and partially funded by the "Consejería de Innovación Ciencia y Empresa de Andalucía" (Spain) under research project P06-TIC-01433.

5. REFERENCES

[1] A. Aho, R. Sethi, and J. Ullman. Compilers: principles, techniques, and tools. *Reading, MA,*, 1986.

[2] C. Bizer, R. Lee, and E. Pietriga. Fresnel: A browser-independent presentation vocabulary for rdf. In *Proceedings of the Second International Workshop on Interaction Design and the Semantic Web, Galway, Ireland*. Citeseer, 2005.

[3] J. Broekstra, A. Kampman, and F. Van Harmelen. Sesame: A generic architecture for storing and querying rdf and rdf schema. *Lecture Notes in Computer Science*, pages 54–68, 2002.

[4] P. Buitelaar, P. Cimiano, P. Haase, and M. Sintek. Towards linguistically grounded ontologies. In *Proceedings of the European Semantic Web Conference (ESWC)*, pages 111–125, 2009.

[5] G. Francopoulo, N. Bel, M. George, N. Calzolari, M. Monachini, M. Pet, and C. Soria. Lexical markup framework (LMF) for NLP multilingual resources. In *Proceedings of the workshop on multilingual language resources and interoperability*, pages 1–8. Association for Computational Linguistics, 2006.

[6] A. S. M. L. Hoque and M. Arefin. Multilingual data management in database environment. *Malaysian Journal of Computer Science*, 22(1):44–63, 2009.

[7] A. Katifori, C. Halatsis, G. Lepouras, C. Vassilakis, and E. Giannopoulou. Ontology visualization methods: a survey. *ACM Computing Surveys (CSUR)*, 39(4):10, 2007.

[8] A. Kumaran and J. R. Haritsa. On database support for multilingual environments. In *Proceedings of the IEEE RIDE Workshop on Multilingual Information Management,*, 2003.

[9] A. Miles, B. Matthews, M. Wilson, and D. Brickley. SKOS Core: Simple knowledge organisation for the web. In *Proceedings of the International Conference on Dublin Core and Metadata Applications*, pages 12–15, 2005.

[10] C. Peters, T. Deselaers, N. Ferro, J. Gonzalo, G. F. Jones, M. Kurimo, T. Mandl, A. Peñas, and V. Petras. *Evaluating Systems for Multilingual and Multimodal Information Access*, volume 5706. Springer, 2008.

Cross-Lingual Ontology Mapping and Its Use on the Multilingual Semantic Web

Bo Fu, Rob Brennan, Declan O'Sullivan
Knowledge and Data Engineering Group, School of Computer Science and Statistics,
Trinity College Dublin, College Green, Dublin 2, Ireland
{bofu, rob.brennan, declan.osullivan}@scss.tcd.ie

ABSTRACT
Ontology-based knowledge management systems enable the automatic discovery, sharing and reuse of structured data sources on the semantic web. With the emergence of multilingual ontologies, accessing knowledge across natural language barriers has become a pressing issue for the multilingual semantic web. In this paper, a semantic-oriented cross-lingual ontology mapping (SOCOM) framework is proposed to enhance interoperability of ontology-based systems that involve multilingual knowledge repositories. The contribution of cross-lingual ontology mapping is demonstrated in two use case scenarios. In addition, the notion of appropriate ontology label translation, as employed by the SOCOM framework, is examined in a cross-lingual ontology mapping experiment involving ontologies with a similar domain of interest but labelled in English and Chinese respectively. Preliminary evaluation results indicate the promise of the cross-lingual mapping approach used in the SOCOM framework, and suggest that the integrated appropriate ontology label translation mechanism is effective in the facilitation of monolingual matching techniques in cross-lingual ontology mapping scenarios.

Keywords
Cross-Lingual Ontology Mapping; Appropriate Ontology Label Translation; Matching Assessment Feedback; Querying of Multilingual Knowledge Repositories.

1. INTRODUCTION
The promise of the semantic web is that of a new way to organise, present and search information that is based on meaning and not just text. Ontologies are explicit and formal specifications of conceptualisations of domains of interests [11], thus are at the heart of semantic web technologies such as semantic search [8] and ontology-based information extraction [2]. As knowledge and knowledge representations are not restricted to the usage of a particular natural language, multilinguality is increasingly evident in ontologies as a result. Ontology-based applications therefore must be able to work with ontologies that are labelled in diverse natural languages. One way to realise this is by means of cross-lingual ontology mapping (CLOM).

In this paper, a summary of current CLOM approaches is presented in section 2. A semantic-oriented cross-lingual ontology mapping (SOCOM) framework that aims to facilitate mapping tasks carried out in multilingual environments is proposed and discussed in section 3. To illustrate possible applications of the SOCOM framework on the multilingual semantic web, two use case scenarios including cross-language document retrieval and personalised querying of multilingual knowledge repositories are presented in section 4. An overview of the initial implementation of the proposed framework is given in section 5. Section 6 presents an experiment that engages the integrated framework in a mapping scenario that involves ontologies labelled in English and Chinese, and discusses the evaluation results and findings from this experiment. Finally, work in progress is outlined in section 7.

WWW 2010, April 26-30, 2010, Raleigh, North Carolina, USA.

2. STATE OF THE ART
Current CLOM strategies can be grouped into five categories, namely manual processing, corpus-based approach, instance-based approach, linguistic enrichment of ontologies and the two-step generic approach. A costly *manual* CLOM process is documented in [13], where the English version of the AGROVOC [1] thesaurus is mapped to the Chinese Agriculture Thesaurus. Given large and complex ontologies, such an approach would be infeasible. Ngai et al. [16] propose a *corpus-based* approach to align the English thesaurus WordNet [2] and the Chinese thesaurus HowNet[3]. As bilingual corpora are not always available to domain-specific ontologies, it is difficult to apply their approach in practice. The *instance-based* approach proposed by Wang et al. [24] generates matching correspondences based on the analysis of instance similarities. It requires rich sets of instances embedded in ontologies, which is a condition that may not always be satisfied in the ontology development process. Pazienza & Stellato propose a *linguistically motivated* mapping method [17], advocating a linguistic-driven approach in the ontology development process that generates enriched ontologies with human-readable linguistic resources. To facilitate this linguistic enrichment process, a plug-in for the Protégé[4] editor – OntoLing [5] was also developed [18]. Linguistically enriched ontologies may offer strong evidence when generating matching correspondences. However, as such enrichment is not currently standardised, it is difficult to apply the proposed solution.

Trojahn et al. [23] present a multilingual ontology mapping framework, where ontology labels are first represented with collections of phrases in the target natural language. Matches are then generated using specialized monolingual matching agents that use various techniques (i.e. structured-based matching algorithms, lexicon-based matching algorithms and so on). However, as Shvaiko & Euzenat state in [20], "despite the many component matching solutions that have been developed so far, there is no integrated solution that is a clear success". Often various techniques are combined in order to generate high quality matching results [12], searching for globally accepted matches

[1] http://aims.fao.org/website/AGROVOC-Thesaurus/sub
[2] http://wordnet.princeton.edu
[3] http://www.keenage.com/html/e_index.html
[4] http://protege.stanford.edu
[5] http://art.uniroma2.it/software/OntoLing

can lead to a limited matching scope. In 2008, an OAEI[6] test case that involves the mapping of web directories written in English and Japanese was designed. Only one participant – the RiMOM tool – was able to submit results [26], by using a Japanese-English dictionary to translate labels from the Japanese web directory into English first, before applying monolingual matching procedures. This highlights the difficulty of exercising current monolingual matching techniques in CLOM scenarios.

Trojahn et al's framework and RiMOM's approach both employ a *generic two-step* method, where ontology labels are translated into the target natural language first and monolingual matching techniques are applied next. The translation process occurs in isolation of the mapping activity, and takes place independently of the semantics in the concerned ontologies. As a result, inadequate and/or synonymic translations can introduce "noise" into the subsequent matching step, where matches may be neglected by matching techniques that (solely) rely on the discovery of lexical similarities. This conception is further examined in [9], where strong evidence indicates that to enhance the performance of existing monolingual matching techniques in CLOM scenarios, appropriate ontology label translation is key to the generation of high quality matching results. This notion of selecting appropriate ontology label translations in the given mapping context is the focus of the SOCOM framework and the evaluation shown in this paper.

Notable work in the field of (semi-)automatic ontology label translation conducted by Espinoza et al. [7] introduces the LabelTranslator tool, which is designed to assist humans during the ontology localisation process. Upon selecting the labels of an ontology one at a time, ranked lists of suggested translations for each label are presented to the user. The user finally decides which suggested translation is the best one to localise the given ontology. In contrast to the LabelTranslator tool, the ontology rendition process of the SOCOM framework presented in this paper differs in its input, output and design purpose. Firstly, our rendering process takes formally defined ontologies (i.e. in RDF/OWL format) as input, but not the labels within an ontology. Secondly, it outputs formally defined ontologies labelled in the target natural language, but not lists of ranked translation suggestions. Lastly, our rendering process is designed to facilitate further machine processing (more precisely, existing monolingual ontology matching techniques), whereas the LabelTranslator tool aims to assist humans.

3. THE SOCOM FRAMEWORK

Given ontologies O_1 and O_2 (see Figure 1) that are labelled in different natural languages, O_1 is first transformed by the SOCOM framework into an equivalent of itself through the *ontology rendering* process as O_1'. O_1' contains all the original semantics of O_1 but is labelled in the natural language that is used by O_2. O_1' is then matched to O_2 using *monolingual matchers* to generate candidate matches, which are then reviewed by the *matching assessment* mechanism in order to establish the final mappings.

Ontology renditions are achieved by structuring the translated ontology labels in the same way as the original ontology O_1, and assigning these translation labels to new namespaces to create well-formed resource URIs in O_1' (for more details, please see [9]). Note that the structure of O_1 is not changed during this process, as Giunchiglia et al. [10] point out, the conceptualisation of a particular ontology node is captured by its label and its position in the ontology structure. Thus, the ontology rendering process should not modify the position of a node, because doing so would effectively alter the semantics of the original ontology.

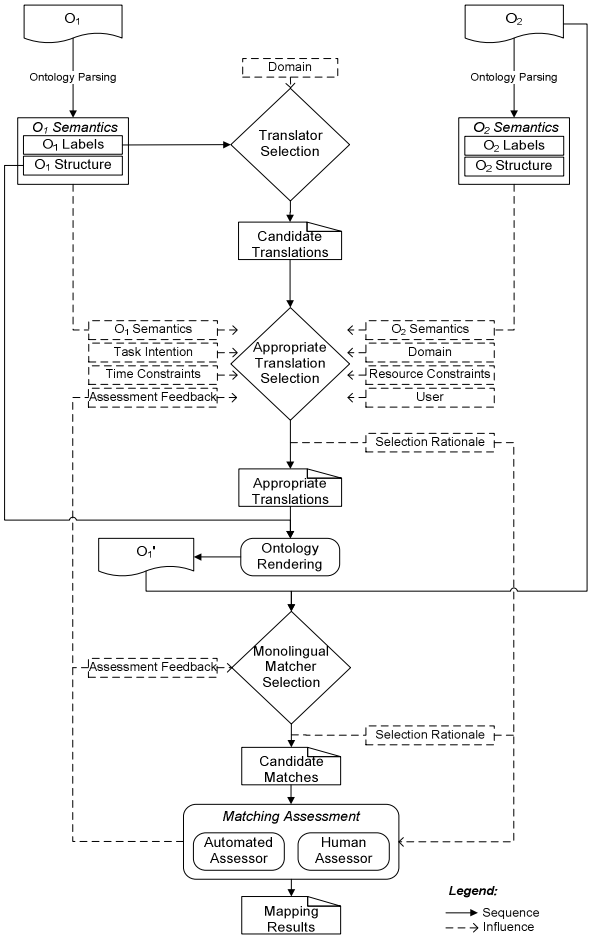

Figure 1. SOCOM Framework Workflow Overview

In contrast to the generic approach, where the translation of ontology labels takes place in isolation from the ontologies concerned, the SOCOM framework is semantic-oriented and aims to identify the most appropriate translation for a given label. To achieve this, firstly, suitable translation tools are selected at the *translator selection* point to generate candidate translations. This selection process is influenced by the knowledge domain of the concerned ontologies. For general knowledge representations, off-the-shelf machine translation (MT) tools or thesauri can be applied. For specific domains such as the medical field, specialised translation media are more appropriate. Secondly, to identify the most appropriate translation for a label among its candidate translations, the *appropriate translation selection* process is performed. This selection process is under the influence of several information sources including the source ontology semantics, the target ontology semantics, the mapping intent, the operating domain, the time constraints, the resource constraints, the user and finally the matching assessment result feedback. These influences are explained next.

The *semantics defined in O_1* can indicate the context that a to-be-translated label is used in. Given a certain position of the node with this label, the labels of its surrounding nodes (referred to as surrounding semantics in this paper) can be retrieved and studied. For example, for a class node, its surrounding semantics can be

[6] http://oaei.ontologymatching.org

represented by the labels of its super/sub/sibling-classes. For a property node, its surrounding semantics can be represented by the labels of the resources which this property restricts. For an individual, the surrounding semantics can be characterised by the label of the class it belongs to. Depending on the granularity of the given ontologies in a mapping scenario, an ontological resource's surrounding semantics should be modelled with flexibility. For example, if the ontologies are rich in structure, immediate surrounding resource labels (e.g. direct super/sub relations) alone can form the content of the surrounding semantics. If the ontologies are rich in instance, where the immediate surrounding label (e.g. the class an instance belongs to) alone is weak to provide the instance's context of use, indirect (e.g. all super/sub classes declared in the ontology) resource labels should be included in the surrounding semantics. The goal of obtaining surrounding semantics of a given resource is to provide the translation selection process with additional indications of the context a resource is used in[7].

As O_1 is transformed so that it can be best mapped to O_2, the *semantics defined in* O_2 therefore can act as broad translation selection rules. When several translation candidates are all linguistically correct for a label in O_1, the most appropriate translation is the one that is most semantically similar to what is used in O_2. An example of appropriate ontology label translation is shown in Figure 2, where the source ontology is labelled in Chinese and is mapped to an English target ontology. The class *摘要* from the source ontology has translation candidates *abstract* and *summary*. To determine the most appropriate translation, the defined semantics of the target ontology can influence the translation selection process. To understand how this is possible, consider three scenarios. Figure 2a demonstrates a situation where a class named *Summary* exists in the target ontology. In this case, *Summary* would be considered as more appropriate than *abstract* since it is the exact label used by the target ontology. Figure 2b illustrates another scenario where the target ontology contains a class named *Sum*. From a thesaurus or a dictionary, one can learn that *Sum* is a synonym of *summary*, therefore, instead of using either *abstract* or *summary*, *Sum* will be chosen as the appropriate translation in this case. Figure 2c shows a third scenario where both *Abstract* and *Summary* exist in the target ontology, the appropriate translation is then concluded by studying the surrounding semantics. The source class *摘要* has a super-class *出版物* (with translation candidates *publication* and *printing*), two sibling-classes *章节* (with translation candidates *chapter* and *section*) and *书籍* (with translation candidates *book* and *literature*). Its surrounding semantics therefore include: {*publication*, *printing*, *chapter*, *section*, *book*, *literature*}. Similarly, in the target ontology, the surrounding semantics of the class *Summary* contains: {*BookChapter*, *Reference*}, and the surrounding semantics of the class *Abstract* would include: {*Mathematics*, *Applied*}. Using string comparison techniques, one can determine that the strings in the surroundings of the target class *Summary* are more similar to those of the source class. *Summary* therefore would be the appropriate translation in such a case. Note that the SOCOM framework is concerned with searching for appropriate translations (from a mapping point of view) but not necessarily the most linguistically correct translations (from a natural language processing point of view), because our motivation for translating ontology labels is so that the ontologies can be best mapped[8]. This should not be confused with translating labels for the purpose of ontology localisation, where labels of an ontology are translated so that it is "adapted to a particular language and culture" [21].

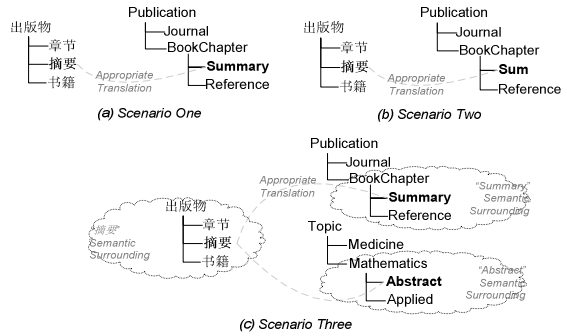

Figure 2. Examples of Appropriate Label Translation

In addition to using the embedded semantics of the given ontologies, *task intention* can also influence the outcome of the translation selection process as it captures some of the mapping motives. Consider a CLOM scenario where the user is not comfortable with all the natural languages involved, and would like to test just how meaningful/useful it is to map the given ontologies. In such a case, the selection of translation candidates need not be very sophisticated, thus results returned from off-the-shelf MT tools can be acceptable. The *domain* of the ontologies is another influence on the translation selection process. For example, if O_1 and O_2 are domain representations where each one is associated with collections of documents in different natural languages, lists of frequently used words in these documents can be collected. The translation candidate that is ranked highest on these lists would be deemed as the most appropriate translation. Moreover, *time constraints* can influence the translation selection process. If the mappings must be conducted dynamically such as the work presented in [5], the translation selection consequently must be fast, where it might not make use of all the resources that are available to it. On the other hand, not all of the aforementioned resources will be available in every CLOM scenario. *Resource constraints* therefore can have an impact on the outcome of the translation selection process. Furthermore, *users*, at times, can have the expertise that is not obtained by the system, and should influence the translation selection process when necessary. Lastly, *matching result feedback* can influence the future selection of appropriate translations (discussed next).

[7] The generation of surrounding semantics presented in this paper does not attempt to estimate the semantic relatedness between concepts, it is a procedure performed within readily defined ontologies in a cross-lingual ontology mapping scenario that aims to gather the context of use for a particular resource in the given ontologies. Though one might assume that the SOCOM framework would work best when ontologies with similar granularity are presented, this however, is not a requirement of the framework. As already mentioned, the surrounding semantics are modelled with flexibility, where indirectly related concepts in the ontology would be collected as long as the surrounding well illustrates the context of use for a particular ontological resource.

[8] Note that the appropriate ontology label translation mechanism presented in this paper does not attempt to disambiguate word senses, as the appropriateness of a translation is highly restricted to the specific mapping scenarios, thus it is not a form of natural language processing technique.

Once O_1' is generated, various monolingual matching techniques can be applied to create matches between O_1' and O_2. The selection of these monolingual matchers depends on the feedback generated from the mapping result assessment. *Assessment feedback* can be implicit (i.e. pseudo feedback) or explicit. Pseudo feedback is obtained automatically, where the system assumes matches that meet certain criteria are correct. For example, "correct" results may be assumed to be the ones that have confidence levels of at least 0.5. The precision of the matches generated can then be calculated for each matching algorithm used, which will allow the ranking of these algorithms. The ranking of the MT sources can also be determined upon establishment of the usage of each MT source (i.e. as percentages) among the "correct" matches. Based on these rankings, the top performing MT tools and matching algorithms can then be selected for the future executions of the SOCOM framework. Explicit feedback is generated from users and is more reliable than pseudo feedback, which can aid the mapping process in the same way as discussed above.

Matching assessment feedback allows insights into how the correct mappings are generated, in particular, which translation tool(s) and matching algorithm(s) are most suitable in the specified CLOM scenario. Such feedback in turn could influence the future selection of appropriate label translations and the monolingual matching techniques to use. Finally, the feedback should be influenced by the *selection rationale* employed during the translation selection process and the monolingual matching process. Such rationale can be captured as metadata as part of the mapping process and include information such as the influence sources used, translation tools used, monolingual matching techniques used, similarity measures of semantic surroundings and so on. The use of matching assessment feedback addresses one of the scalability issues that arise. Consider a mapping scenario where the concerned ontologies contain thousands of entities, one way to rapidly generate mapping results and improve mapping quality dynamically is to use the pseudo feedback. For the first, e.g. 100 mapping tasks, assume the ones that satisfy certain criteria are correct, detect how they are generated, and keep using the same techniques for the remaining mapping tasks. This assessment process can also be recursive where the system is adjusted for every few mapping tasks. Finally, explicit feedback involves users in the mapping process, which contributes towards addressing one of the challenges, namely user involvement in ontology matching as identified by Shvaiko & Euzenat in [20].

4. USE CASES

The notion of using conceptual frameworks such as thesauri and ontologies in search systems [6] [4] for improved information access [19] and enhanced user experiences [22] is well researched in the information retrieval (IR) and the cross-lingual IR (CLIR) community. However, the use of ontology mapping as a technique to aid the search functions in IR has been relatively limited. The most advanced work of using ontology alignment in CLIR, to the best of our knowledge, is Zhang et al.'s statistical approach presented in [25], which does not involve translations of ontology labels. To avail statistical analysis such as latent semantic indexing, singular value decomposition, directed acyclic graphs and maximal common subgraph on document collections, parallel corpora must be generated beforehand. However, this often is an expensive requirement and may not always be satisfied. Also, by applying statistical techniques only, such an approach ignores the existing semantic knowledge within the given ontologies in a mapping scenario. Hence alternative solutions are in need. The SOCOM framework presented in this paper can contribute towards this need. Its contribution can be demonstrated through two use cases as shown in Figures 3 & 4.

User generated content such as forums often contain discussions on how to solve particular technical problems, and a large amount of content of this type is written in English. Consider a scenario illustrated in Figure 3, where the user whose preferred natural language is Portuguese is searching for help on a forum site, but the query in Portuguese is returning no satisfactory results. Let us assume that the user also speaks English as a second language and would like to receive relevant documents that are written in English instead. To achieve this, domain ontologies in Portuguese and English can be extracted based on text presented in the documents using such as Alani et al.'s approach [1]. Mappings can then be generated pre-runtime using the SOCOM framework between the Portuguese ontology and the English ontology, and stored as RDF triples. At run time, once a query is issued in Portuguese, it is first transformed using such as Lopez et al.'s method [14] to associate itself with a concept in the Portuguese domain ontology. This Portuguese concept's corresponding English concept(s) can then be obtained by looking it up in the mapping triplestore. Once the system establishes which English concepts to explore further, their associated documents in English can be retrieved.

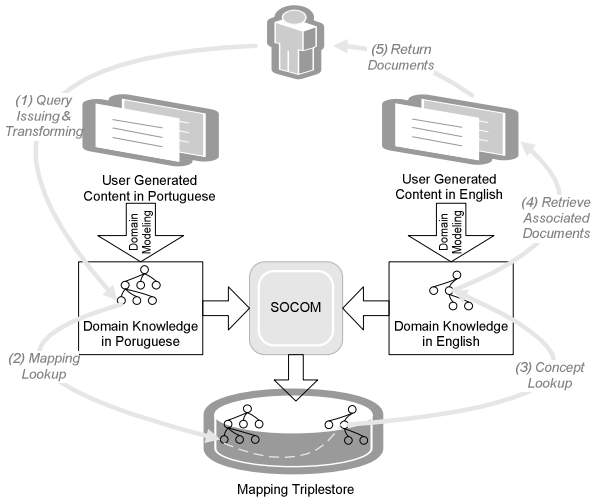

Figure 3. SOCOM Enabled Cross-Language Document Retrieval

Personalisation can also be enhanced with the integration of the SOCOM framework in scenarios such as the one shown in Figure 4, where a user is bi/multi-lingual and would like to receive documents in a restricted knowledge domain in various natural languages as long as they are relevant. To achieve this, ontology-based user models[9] containing knowledge such as user interests and language preferences can be generated pre-runtime using approaches such as [3]. Similar to the previous scenario, domain ontologies labelled in different natural languages can be obtained from sets of documents. In Figure 4, knowledge representations in English, French, German and Spanish are obtained in ontological form. Mappings of the user model and the various domain ontologies can then be generated using the

[9] User modelling is a well researched area particularly in adaptive hypermedia and personalised search systems, however, this is outside the scope of this paper.

SOCOM framework. At run time, a user query is transformed to be associated with a concept or concepts in the user model. By looking up in the mapping triplestore, the matched concepts in various knowledge repositories (the German and the Spanish knowledge repositories in the case of Figure 4) can be obtained, which will then lead to the retrieval of relevant documents in different natural languages.

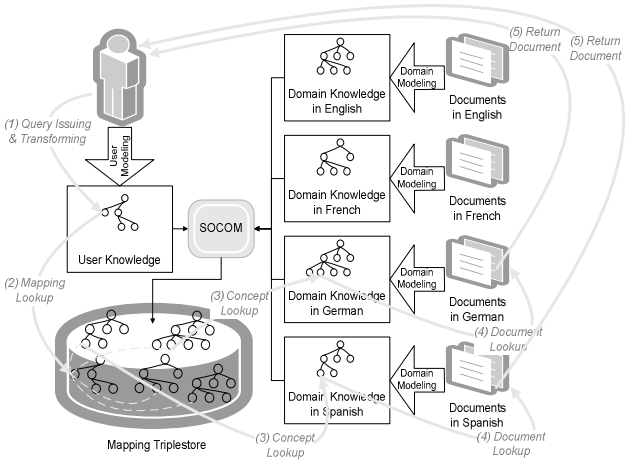

Figure 4. Personalised Querying of Multilingual Knowledge Repositories with SOCOM

5. IMPLEMENTATION

To examine the soundness of the appropriate ontology label translation selection process proposed in the SOCOM framework, an initial implementation of the proposal has been completed that uses just the semantics within the given ontologies in a CLOM scenario. This light-weight translation selection process (i.e. one that includes semantics in O_1 and semantics in O_2, but excludes the six other influence sources as shown in Figure 1) is the focus of the implementation and the evaluation presented in this paper.

This initial SOCOM implementation integrates the Jena 2.5.5 Framework[10] to parse the formally defined input ontologies. To collect candidate translations for ontology labels in O_1, the GoogleTranslate[11] 0.5 API and the WindowsLive[12] translator are used[13]. Synonyms of ontology labels in O_2 are generated by querying WordNet[14] 2.0 via the RiTa[15] API. Ontology labels are often concatenated to create well-formed URIs (as white spaces are not allowed), e.g. a concept *associate professor* can be labelled as *AssociateProfessor* in the ontology. As the integrated MT tools cannot process such concatenated labels, they are split into sequences of their constituent words before being passed to the MT tools. This is achieved by recognising concatenation patterns. In the previous example, white spaces are inserted before each capital letter found other than the first one. The candidate translations are stored in a translation repository, whereas the synonyms are stored in a lexicon repository. Both repositories are stored in the eXist[16] 1.0rc database.

The appropriate translation selection process invokes the repositories in the database via the XML:DB[17] 1.0 API, to compare each candidate translation of a given source label to what is stored in the lexicon repository. An overview of this appropriate translation selection process can be seen in Figure 5. If a one-to-one match (note that the match found in the lexicon repository can be either a target label used in O_2, or a synonym of a target label that is used in O_2) is found, the (matched target label or the matched synonym's corresponding) target label is selected as the appropriate translation. If one-to-many matches (i.e. when several target labels and/or synonyms in the lexicon repository are matched) are found, the surrounding semantics (see section 3) of the matched target labels are collected and compared to the surrounding semantics of the source label in question. Using a space/case-insensitive edit distance string comparison algorithm based on Nerbonne et al.'s method [15], the target label with surrounding semantics that are most similar to those of the source resource is chosen as the most appropriate translation. If no match is found in the lexicon repository, for each candidate translation, a set of interpretative keywords are generated to illustrate the meaning of this candidate. This is achieved by querying Wikipedia[18] via the Yahoo Term Extraction Tool[19]. Using the same customised string comparison algorithm, the candidate with keywords that are most similar to the source label's surrounding semantics is deemed as the most appropriate translation.

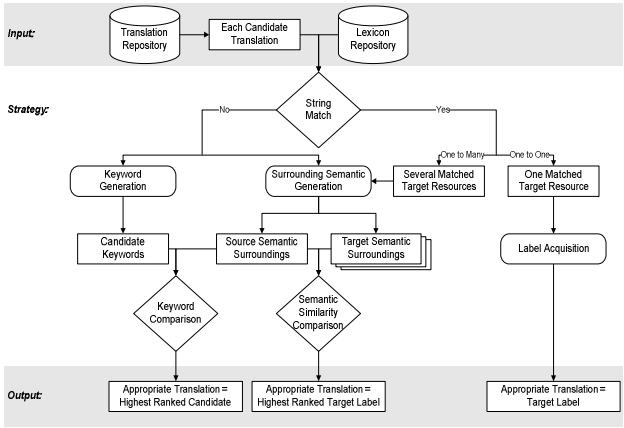

Figure 5. Overview of the Appropriate Ontology Label Translation Selection Process

Once appropriate translations are identified for each label in O_1, given the original source ontology structure, O_1' is generated using the Jena Framework. Finally, O_1' is matched to O_2 to generate candidate matches via the Alignment API[20] version 3.6.

6. EVALUATION

To evaluate the effectiveness of the integrated appropriate translation selection process, this initial implementation of the SOCOM framework is engaged in a CLOM experiment that

[10] http://jena.sourceforge.net
[11] http://code.google.com/p/google-api-translate-java
[12] http://www.windowslivetranslator.com/Default.aspx
[13] One could use a dictionary/thesaurus here, however, as the appropriate ontology label translation selection process in the SOCOM framework is not a word sense disambiguation mechanism (see section 3), off-the-self MT tools are efficient to collect candidate translations.
[14] http://wordnet.princeton.edu
[15] http://www.rednoise.org/rita
[16] http://exist.sourceforge.net
[17] http://xmldb-org.sourceforge.net/index.html
[18] http://www.wikipedia.org
[19] http://developer.yahoo.com/search/content/V1/termExtraction.html
[20] http://alignapi.gforge.inria.fr

involves ontologies labelled in Chinese and English describing the research community domain, against a baseline system – the generic approach, where labels are translated in isolation using just the GoogleTranslate 0.5 API and matches are generated using the Alignment API[21] version 3.6 (see [9] for more technical details of the implementation of the generic approach).

6.1 Experimental Setup

Figure 6 gives an overview of the experiment. A Chinese ontology CSWRC[22] is created manually by a group of domain experts (excluding the authors of this paper) based on the English SWRC[23] ontology. It contains 54 classes, 44 object properties and 30 data type properties. This Chinese ontology is matched to the English ISWC[24] ontology (containing 33 classes, 18 object properties, 17 data type properties and 50 instances) using the generic approach and the SOCOM approach, generating results M-G and M-S respectively.

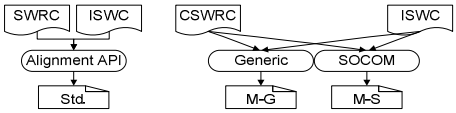

Figure 6. Cross-Lingual Ontology Mapping Experiments

As the CSWRC ontology is formally and semantically equivalent (with the same structured concepts but labelled in Chinese) to the SWRC ontology, a reliable set of gold standard (referred to as Std. in Figure 6) can be generated as matches found between the SWRC ontology and the ISWC ontology using the Alignment API[25]. By comparing results M-G and M-S to Std., this experimental design aims to find out which approach can generate higher quality matching results, when the concerned ontologies hold distinct natural languages and varied structures.

6.2 Experimental Results

Precision and recall[26] scores of M-G and M-S are calculated, see Figure 7, where a match is considered correct as long as the identified pair of corresponding resources is included in the gold standard Std., regardless of its confidence level.

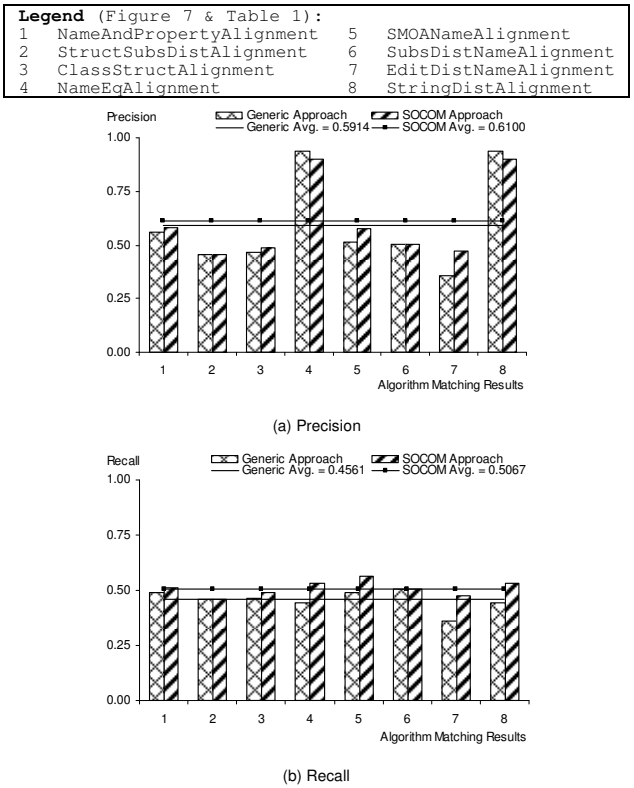

Figure 7. Overview of Precision and Recall when Disregarding Confidence Levels

Figure 7a shows that except the NameEqAlignment and the StringDistAlignment algorithm, all other matching methods indicate equal or higher precision when using the SOCOM approach. The aforementioned two algorithms employ strict string comparison techniques, where no dissimilarity between two labels is overlooked. Though this is a desirable characteristic at times, in this particular experiment setting, some matches are neglected in Std.. E.g. when using the StringDistAlignment algorithm, the gold standard was unable to establish a match between the class *AssociateProfessor* (in SWRC) and the class *Associate_ Professor* (in ISWC) because these labels are not identical, although this would have been a sound match if a human was involved or if preprocessing was undertaken. When the SOCOM approach is used to match CSWRC to ISWC, the most appropriate translation for the class *副教授* (associate professor) in the source ontology was determined as *Associate_Professor* since this exact English label was used in the target ontology. Consequently, a match with 1.00 confidence level between the two was generated in M-S. However, as this correspondence was not included in Std., such a result is deemed as incorrect. Similar circumstances led to the lower precision scores of the SOCOM approaches in cases that involve the NameEqAlignment and the StringDistAlignment algorithms. Nevertheless, on average, with a precision score at 0.61, the SOCOM approach generated more correct matching results than the generic approach overall. Furthermore, at an average recall score of 0.5067 (see Figure 7b), the SOCOM approach demonstrates that its correct results are always more complete than those generated by the generic approach.

As precision and recall each measures one aspect of the match quality, f-measure scores are calculated to indicate the overall

[21] The Alignment API 3.6 contains eight matching algorithms, namely NameAndPropertyAlignment, StructSubsDistAlignment, ClassStructAlignment, NameEqAlignment, SMOANameAlignment, SubsDistNameAlignment, EditDistNameAlignment and StringDistAlignment. For each correspondence found, a matching relationship is given and is accompanied by a confidence measure that range between 0 (not confident) and 1 (confident).

[22] http://www.scss.tcd.ie/~bofu/SOCOMExperimentJuly2009/Ontologies/CSWRC.owl

[23] http://ontoware.org/frs/download.php/298/swrc_v0.3.owl

[24] http://annotation.semanticweb.org/iswc/iswc.owl

[25] Based on the assumption that the CSWRC ontology is equivalent to the SWRC ontology, this experimental design aims to validate whether matches generated using the exact same matching algorithms would result the same or highly similar corresponding concepts.

[26] Given a gold standard with R number of matching results, and an evaluation set containing X number of results, if N number of them are correct based on the gold standard, then for this evaluation set precision = N/X, recall = N/R and f-meaure = 2/(1/precision + 1/recall).

quality[27]. Table 1 shows that the SOCOM approach generated results with at least equal quality compared to the generic approach. In fact, the majority of algorithms were able to generate higher quality matches when using the SOCOM approach, leading to an average of 0.5460 in its f-measure score. The differences in the two approaches' f-measure scores (when they exist) range from a smallest 1.9% (when using the NameAndPropertyAlignment algorithm) to a highest of 11.4% (when using the EditDistNameAlignment algorithm). Additionally, when using the SOCOM approach, bigger differences in f-measure can be seen in lexicon-based algorithms. Such a finding indicates that appropriate ontology label translation in the SOCOM framework contributes positively to the enhanced performances of matching algorithms, particularly those that are lexicon-based.

Table 1. **F-measure Scores when Disregarding Confidence Levels**

	Generic	SOCOM
1	.5233	.5421
2	.4574	.4574
3	.4651	.4884
4	.6000	.6667
5	.5020	.5714
6	.5039	.5039
7	.3571	.4714
8	.6000	.6667
Avg.	.5011	.5460

So far, the confidence levels of matching results have not been taken into account. To include this aspect in the evaluation, confidence means of the correct matches and their standard deviations are calculated. The mean is the average confidence of the correct matches found in a set of matching results, where the higher it is, the better the results. The standard deviation is a measure of dispersion, where the greater it is, the greater the spread in the confidence levels. Higher quality matching results therefore are those with higher means and lower standard deviations. On average, when using the SOCOM framework, the confidence mean is 0.7105. Whereas, a lower mean of 0.6970 is found in the generic approach. The standard deviation when using the SOCOM framework is 0.2134, which is lower than 0.2161 as found in the generic approach. These findings denote that matches generated using the SOCOM approach are of higher quality, because they are not only more confident but also less dispersed.

Moreover, average precision, recall and f-measure scores are collected at various thresholds. These scores are calculated when the conditions a correct result must satisfy adjust, i.e. a matching result is only considered correct when it is included in the gold standard, and it has confidence level of at least 0.25, 0.50, 0.75 or 1.00. An overview of the trends is shown in Figure 8. As the requirement for a correct matching result become stricter, the precision (Figure 8a) and recall (Figure 8b) scores both decline as a result, leading to a similar decreasing trend in the f-measure (Figure 8c) scores. The differences in the recall scores of the two approaches are greater than the differences of their precision scores. This finding suggests that the matches generated using the two approaches may appear similar in their correctness, but the ones generated by the SOCOM approach are more complete. Overall, the SOCOM approach always has higher precision, recall and f-measure scores than the generic approach no matter what the threshold is[28]. This finding further confirms that the matches generated using the SOCOM approach are of higher quality.

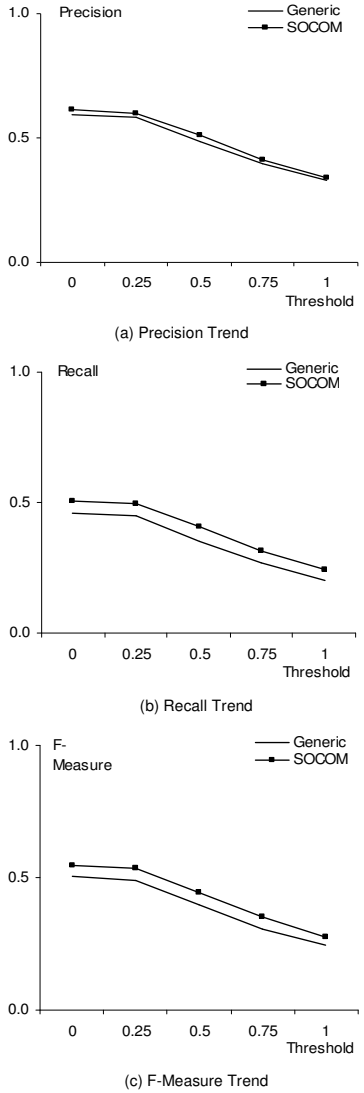

Figure 8. Trend Overview in Average Precision, Recall and F-Measure

Lastly, one can argue that the differences in the f-measure scores found between the generic and the SOCOM approach are rather small and therefore can be ignored. To validate the difference (if it exists) of the two approaches, paired t-tests are carried out on the f-measure scores collected across various thresholds, and a p-value of 0.001 is found. At a significance level of α=0.05, it can be concluded that the f-measure scores are statistically significant, meaning that the SOCOM approach generated higher quality matches than the generic approach.

[27] Note that neither precision nor recall alone is a measurement of the overall quality of a set of matching results, as the former is a measure for correctness and the latter is a measure for completeness. One can be sacrificed for the optimisation of the other, for example, when operating in the medical domain, recall may be sacrificed in order to achieve high precision; when merging ontologies, the opposite may be desired.

[28] Dotted lines of the generic and the SOCOM approach shown in Figure 8 are almost parallel to one another, this may be in part a result of the engineering approach deployed in the experiment (i.e. using the same tools in the implementation for both approaches). Further research, however, is needed to confirm the validity of this speculation.

7. CONCLUSIONS & FUTURE WORK

A semantic-oriented framework to cross-lingual ontology mapping is presented and evaluated in this paper. Preliminary evaluation results of an early prototype implementation illustrate the effectiveness of the integrated appropriate ontology label translation mechanism, and denote a promising outlook for applying CLOM techniques in multilingual ontology-based applications. The findings also suggest that a fully implemented SOCOM framework – i.e. one that integrates all the influence factors (discussed in section 2) – would be even more effective in the generation of high quality matches in CLOM scenarios.

The implementation of such a comprehensive SOCOM framework is currently on-going. It is planned to be evaluated using the benchmark datasets from the OAEI 2009 campaign, engaging the proposed framework in the mapping of ontologies that are written in very similar natural languages, namely English and French. In addition, the SOCOM framework is to be embedded in a demonstrator cross-language document retrieval system as part of the Centre for Next Generation Localisation, which involves several Irish academic institutions and a consortium of multi-national industrial partners aiming to develop novel localisation techniques for commercial applications.

8. ACKNOWLEDGMENT

This research is partially supported by Science Foundation Ireland (Grant 07/CE/11142) as part of the Centre for Next Generation Localisation (http://www.cngl.ie) at Trinity College Dublin.

9. REFERENCES

[1] Alani H., Kim S., Millard D. E., Weal M. J., Hall W., Lewis P. H., Shadbolt N. R.. Automatic ontology-based knowledge extraction from Web documents. IEEE Intelligent Systems 18, 1, 14-21, Jan. 2003

[2] Buitelaar P., Cimiano P., Frank A., Hartung M., Racioppa S.. Ontology-based information extraction and integration from heterogeneous data sources. International Journal of Human Computer Studies, 66, 11, 759-788, Nov. 2008

[3] Cantador I., Fernández M., Vallet D., Castells P., Picault J., Ribière M.. A multi-purpose ontology-based approach for personalised content filtering and retrieval. Advances in Semantic Media Adaptation and Personalization. Studies in Computational Intelligence, vol. 93, 25-51, 2008

[4] Castells P., Fernández M., Vallet D.. An adaptation of the vector-Space model for ontology-based information retrieval. IEEE Transactions on Knowledge and Data Engineering 19(2), Special Issue on Knowledge and Data Engineering in the Semantic Web Era, 261-272, Feb. 2007

[5] Conroy C., Brennan R., O'Sullivan D., Lewis D.. User evaluation study of a tagging approach to semantic mapping. In Proceedings of ESWC, 623-637, 2009

[6] De Luca E. W., Eul M., Nürnberger A.. Multilingual query-reformulation using an RDF-OWL EuroWordNet representation. In Proceedings of the Workshop on Improving Web Retrieval for Non-English Queries (iNEWS07), at SIGIR 2007, ISBN 978-84-690-6978-3, 55-61, 2007

[7] Espinoza M., Gómez-Pérez A., Mena E.. LabelTranslator – a tool to automatically localize an ontology. In Proceedings of ESWC, 792-796, 2008

[8] Fernandez M., Lopez V., Sabou M., Uren V., Vallet D., Motta E., Castells P.. Semantic search meets the Web. In Proceedings of IEEE ICSC, 253-260, 2008

[9] Fu B., Brennan R., O'Sullivan D.. Cross-lingual ontology mapping – an investigation of the impact of machine translation. In Proceedings of ASWC, LNCS 5926, 1-15, 2009

[10] Giunchiglia F., Yatskevich M., Shvaiko P.. Semantic matching: algorithms and implementation. Journal on Data Semantics, vol. IX, 1-38, 2007

[11] Gruber T.. A translation approach to portable ontologies. Knowledge Acquisition 5(2):199-220, 1993

[12] Li J., Tang J., Li Y., Luo Q.. RiMOM: A dynamic multistrategy ontology alignment framework. IEEE Transactions on Knowledge and Data Engineering, Vol. 21, No. 8, 1218-1232, 2009

[13] Liang A. C., Sini M.. Mapping AGROVOC and the Chinese agricultural thesaurus: definitions, tools, procedures. New Review of Hypermedia and Multimedia, 12:1, 51-62, 2006

[14] Lopez V., Uren V., Motta E., Pasin M.. AquaLog: an ontology-driven question answering system for organizational semantic intranets. Web Semantics. 5, 2, 72-105, Jun. 2007

[15] Nerbonne J., Heeringa W., Kleiweg P.. Edit distance and dialect proximity. Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison, 2nd ed. CSLI, Stanford, v-xv, 1999

[16] Ngai G., Carpuat M., Fung P.. Identifying concepts across languages: a first step towards a corpus-based approach to automatic ontology alignment. In Proceedings of the 19th International Conference on Computational Linguistics, vol.1, 1-7, 2002

[17] Pazienta M., Stellato A.. Linguistically motivated ontology mapping for the Semantic Web. In Proceedings of the 2nd Italian Semantic Web Workshop, 14-16, 2005

[18] Pazienza M. T., Stellato A.. Exploiting linguistic resources for building linguistically motivated ontologies in the Semantic Web. In Proceedings of OntoLex Workshop, 2006

[19] Shuang L., Fang L., Clement Y., Weiyi M.. An effective approach to document retrieval via utilizing WordNet and recognizing phrases. 27th Annual international ACM SIGIR Conference on Research and Development in information Retrieval, 266-272, ACM Press, 2004

[20] Shvaiko P., Euzenat J.. Ten challenges for ontology matching. In Proceedings of ODBASE, 1164-1182, 2008

[21] Suárez-Figueroa M. C., Gómez-Pérez A.. First attempt towards a standard glossary of ontology engineering terminology. In Proceedings of the 8th International Conference on Terminology and Knowledge Engineering (TKE'08), 2008

[22] Stamou, S., Ntoulas, A.. Search personalization through query and page topical analysis. User Modeling and User-Adapted Interaction 19, 1-2, 5-33., Feb. 2009

[23] Trojahn C., Quaresma P., Vieira R.. A framework for multi-lingual ontology mapping. In Proceedings of LREC, 1034-1037, 2008

[24] Wang S., Englebienne G., Schlobach S.. Learning concept mappings from instance similarity. In Proceedings of ISWC, 339-355, 2008

[25] Zhang L., Wu G., Xu Y., Li W., Zhong Y.. Multilingual collection retrieving via ontology alignment. In Proceeding of ICADL 2004, LNCS 3334, 510-514, Springer-Verlag, 2004

[26] Zhang X., Zhong Q., Li J., Tang J., Xie G., Li H.. RiMOM results for OAEI 2008. In Proceedings of the OM Workshop, 182-189, 2008

Rivière or *Fleuve*? Modelling Multilinguality in the Hydrographical Domain

Guadalupe Aguado-de-Cea, Asunción Gómez-Pérez,
Elena Montiel-Ponsoda, and Luis M. Vilches-Blázquez
Ontology Engineering Group
Dpto. de Inteligencia Artificial
Facultad de Informática, Universidad Politécnica de Madrid
28660, Boadilla del Monte, Madrid, Spain
{lupe, asun, emontiel, lmvilches}@fi.upm.es

ABSTRACT
The need for interoperability among geospatial resources in different natural languages evidences the difficulties to cope with domain representations highly dependent of the culture in which they have been conceived. In this paper we characterize the problem of representing cultural discrepancies in ontologies. We argue that such differences can be accounted for at the ontology terminological layer by means of external elaborated models of linguistic information associated to ontologies. With the aim of showing how external models can cater for cultural discrepancies, we compare two versions of an ontology of the hydrographical domain: *hydrOntology*. The first version makes use of the labeling system supported by RDF(S) and OWL to include multilingual linguistic information in the ontology. The second version relies on the *Linguistic Information Repository* model (LIR) to associate structured multilingual information to ontology concepts. In this paper we propose an extension to the LIR to better capture linguistic and cultural specificities within and across languages.

Categories and Subject Descriptors
I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods [semantic networks]

General Terms
Documentation, Design

Keywords
multilingual ontologies, hydrographical domain, LIR, ontology localization

1. INTRODUCTION
The symbiosis between ontologies and natural language has proven more and more relevant on the light of the growing interest and use of Semantic Web technologies. Ontologies that are well-documented in a natural language not only provide humans with a better understanding of the world model they represent, but also a better exploitation by the systems that may use them. This "grounding in natural language" is believed to provide improvements in tasks such as ontology-based information extraction, ontology learning and population from text, or ontology verbalization, as pointed out in [4].

Nowadays, there is a growing demand for ontology-based applications that need to interact with information in different natural languages, i.e., with multilingual information. This is the case of numerous international organizations currently introducing semantic technologies in their information systems, such as the Food and Agriculture Organization or the World Health Organization, to mention just a few. Such organizations have to manage information and resources available in more than a dozen of different natural languages, and have to customize the information they produce to a similar number of linguistic communities.

In the present research, we are concerned with a further use case: the geospatial information. The importance of multilingualism in this domain lies in the need for interoperability among multiple geospatial resources in different languages, and a flexible human interaction with multilingual information. For many years, geospatial information producers have focused on the collection of data in one or different languages without considering the interoperability level among them. If we take as example the case of Spain, data have been collected from multiple producers at different levels (national, regional and local), and in the different languages that are official in the country (Spanish, Catalan, Basque and Galician), but there have been no efforts to make these data interoperable.

Today, the widespread use of geospatial information and the globalization phenomenon have brought about a radical shift in the conception of this information. In this context, multilingualism has reached a pre-eminent position in the international scene. The rapid emergence of international projects such as EuroGeoNames[1] confirms this trend. The main goal of EuroGeoNames is to implement an interoperable internet service that will provide access to the official, multilingual geographical name data held at national level and make them available at European label.

Ideally, the "meaning" expressed by ontologies would provide the "glue" between geospatial communities [22] by capturing their knowledge and facilitating the alignment of heterogeneous and multilingual elements. However, this still remains an open issue because of the cultural and subjective discrepancies in the representation of geospatial information. This domain is a good

[1] http://www.eurogeographics.org/eurogeonames

Other international projects in a similar line of research are: eSDI-NET+ (http://www.esdinetplus.eu) or GIS4EU (http://www.gis4eu.org/)

exponent of what has been called *culturally-dependant domains* [8] that is, domains in which their categorizations tend to reflect the particularities of a certain culture. The geospatial domain has to do with the most direct experiences of humans with their environment, and it has, therefore, a very strong relation with how a certain community perceives and interacts with a natural phenomenon. A good example of these experiences can be found in [13]. This is inevitably reproduced in the different viewpoints and granularity levels represented by conceptualizations in this domain, which are, in its turn, reflected in the language.

However true that may be, we believe that interoperability is still possible by assuming a trade-off between what is represented in the ontology and what is captured in the ontology terminological (or lexical) layer [2]. Up to now, the representation of multilingualism in ontologies has not been a priority [1], and very few efforts have been devoted to the representation of linguistic information in ontologies, let alone multilingual information. We believe that a sound lexical (and terminological) model independent from the ontology that could capture cultural discrepancies, would pave the way for solving this problem.

In this paper, our purpose is to show how such an external and portable model created to associate lexical and terminological information to ontologies may account for categorization mismatches among cultures. This is the purpose of the *Linguistic Information Repository* (LIR) [15][19], a model created to capture specific variants of terms within and across languages. With the aim of showing this, we will compare the functionalities offered by two representation modalities to link linguistic and multilingual information with ontologies: the labelling system of RDF(S) and OWL vs. the LIR model. This comparison will be done on the basis of an ontology of the hydrographical domain: *hydrOntology*. Additionally, an extension of the LIR model to better account for categorization mismatches among cultures will be proposed.

The rest of the paper is structured as follows. In section 2 we present the state of the art on formalisms and models to represent linguistic information in ontologies. Then, in section 3 we try to characterize the problem of conceptual mismatches or discrepancies among conceptualizations in multilingual knowledge resources. Section 3 is devoted to a brief description of *hydrOntology*. The inclusion of linguistic information in the ontology by means of the RDF(S) labels is described in section 4. Then, the LIR model is presented in section 5, and its instantiation with the linguistic information related to *hydrOntology* is detailed in section 6. By describing the two versions of *hydrOntology*, we aim at showing the main benefits and drawbacks of each modelling modality. Finally, we conclude the paper in section 7.

2. The linguistic-ontology interface

Most of the ontologies available nowadays in the Web are documented in English, i.e., the human-readable information associated to ontology classes and properties consists of terms and glosses in English. Most of these ontologies, not to say all of them, make use of the *rdfs:label* and *rdfs:comment* properties of the RDF Schema vocabulary, a recommendation of the W3C Consortium to provide "a human-readable version of a resource's name"[3].

It is also specified that labels can be annotated using the "language tagging" facility of RDF literals[4], which permits to indicate the natural language used in a certain information object. The RDF(S) properties can be complemented by Dublin Core metadata [5] that have been created to describe resources of information systems. Examples of the Dublin Core Metadata elements are: title, creator, subject or description. Since it is possible to attach as many metadata as wished, this has been used to associate the same metadata in different natural languages to obtain an ontology documented in different natural languages, in other words, to obtain a multilingual ontology. This is precisely one of the main advantages of this representation modality, namely, associating as much information in different languages as wished.

However, we identify several drawbacks for an appropriate exploitation of the resulting multilingual ontologies:

(1) All annotations are referred to the ontology element they are attached to, but it is not possible to define any relation among the linguistic annotations themselves. This results in a bunch of unrelated data whose motivation is difficult to understand even for a human user.

(2) When different labels in the same language are attached to the same ontology element, absolute synonym or exact equivalence is assumed among the labels. As reported in [6] "identical meaning" among linguistic synonyms is rarely the case. It could be argued that in technical or specialized domains, absolute synonymy exists, but even in those domains, labels usually differ in "denotation, connotation, implicature, emphasis or register" [5], what sometimes is reflected in the subcategorization frames they select (syntactic arguments they co-occur with). We will try to illustrate this in section 6.

(3) A similar situation arises when labels in different languages are attached to the same ontology element. In some cases, they will share the common meaning represented by the ontology element (Figure 1). However, the problem appears when a language understands a certain concept with a different granularity level to the one represented by the ontology concept, as illustrated in Figure 2 and Figure 3. In this case, if more fine-grained equivalents exist in one of the languages represented by several labels, it will be interesting to make those differences explicit for a suitable treatment of multilinguality.

(4) Finally, scalability issues should also be mentioned. If only a couple of languages are involved and not much linguistic information is needed, the RDF(S) properties can suffice. But if a higher number of languages are required, as seems to be the trend in the current demand, the linguistic information will become unmanageable.

On the light of the drawbacks outlined, additional approaches have been proposed to connect linguistic and ontological information. In this sense, we will first refer to the Linguistic Watermark initiative [17]. The Linguistic Watermark is a framework or metamodel for describing linguistic resources and

[2] In [3] the *terminological layer* in an ontology is defined as the terms or labels selected to name ontology elements.

[3] http://www.w3.org/TR/rdf-schema/

[4] http://www.isi.edu/in-notes/rfc3066.txt

[5] http://dublincore.org

using their content to "enrich and document ontological objects". The authors already propose a description for WordNet and a set of dictionaries called DICT. Their idea would be to directly import the linguistic information contained in those resources and integrate it in the ontology. However, it seems as if the reused information is included in the ontology by making use of the RDF(S) properties, and this shows the same disadvantages presented above. This approach is technologically supported by the OntoLing Protégé plugin[6].

A further effort to associate linguistic information to ontologies is represented by the LexInfo [4] model. This model is more in line with what we propose in this paper to enrich ontologies with a linguistic model that is kept separated from the ontology. LexInfo is a joint model that brings together two previous models LingInfo and LexOnto, and builds on the Lexical Markup Framework or LMF, an ISO standard created to represent the linguistic information in computational lexicons. As already mentioned, LexInfo offers an independent portable model that is to be published with arbitrary domain ontologies. LexInfo combines the representation of deep morphological and syntactic structures (segments, head, modifiers), as contained in the LingInfo model, with linguistic predicate-argument structures (subcategorization frames) for predicative elements such as verbs, as captured by LexOnto. Since its main objective is to provide an elaborate model to increase the expressivity of ontological objects in a certain language, it cares less for multilingual aspects and categorization discrepancies among languages.

Finally, we will briefly mention the Simple Knowledge Organization System (SKOS) [12], a model to represent the concept schema of thesauri in RDF(S) and OWL. This model also accounts for the representation of multilingual terms, but does not offer a complex machinery to deal with cultural discrepancies. As it has not been created with the purpose of associating linguistic information to ontologies, the semantic relations captured in the model are limited to hierarchical and associative relations among concepts.

3. Characterization of the multilingual representation problem

The reconciliation of different representations (within the same natural language) can be solved by establishing mappings among those representations. When facing representations in different languages, the mapping process results in a multilingual system. A collection of mapping approaches with monolingual and multilingual resources can be found in [9]. Our approach to tackle multilinguality, however, takes as a starting point one conceptualization to which information in different languages is attached. From the development viewpoint, reusing an existing conceptualization in the domain to transform it into a multilingual resource that can be shared among different speaking communities demands less time and efforts than having to conceptualize the same domain from scratch in each natural language, and then find the mappings or correspondences among concepts. Both approaches have to deal with the differences in conceptualizations that each culture makes. In the mapping approach, it is the mapping itself the one that establishes the equivalence links among ontologies, whereas in the second option, this can be solved at the terminological layer or by modifying the conceptualization (for a detailed analysis of modeling modalities to represent multilinguality in knowledge-based systems, see [1]).

To the best of our knowledge, the most recurrent conceptual discrepancies could be systematically classified as follows:

(a) 1:1 or exact equivalence (as illustrated in Figure 1)

(b) n:1 subsumption relation (isSubsumedBy) (illustrated in Figure 2)

(c) 1:n subsumption relation (subsumes) (represented by Figure 3)

In case (a) both conceptualizations or world views share the same structure and the same granularity level. This is normally reflected in the language by means of a word or term that designates that concept. In the situation represented by (b), the original conceptualization (the one belonging to the English language) makes a more fine grained distinction of a certain reality that does not correlate with the granularity level in the target representation of the same reality. In that case, the target concept is slightly more general, and it could be understood as encompassing the *n* concepts in the original conceptualization. This results in two terms in the English language, for instance, to designate those two concepts, whereas in the target culture, only one term is available. The last case (c) depicts the same situation as in (b) but exactly the other way round.

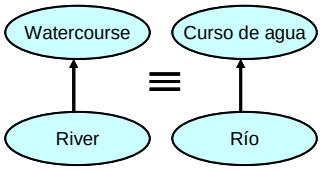

Figure 1. 1:1 or exact equivalence between conceptualizations

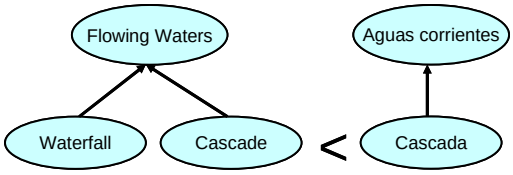

Figure 2. n:1 subsumption relation

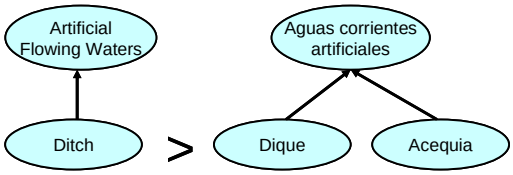

Figure 3. 1:n subsumption relation

However, if our objective is to rely on one ontology to "glue" the different conceptualizations of reality that cultures make, we will need to assume a trade-off between what is represented in the ontology and what is left out, so that every culture can feel that conceptualization as its own, and can meet its representation needs.

Coming back to case (c), if we agree on representing the view of the English culture in the ontology, we will be missing the granularity level of the Spanish world view. We think that those cultural discrepancies could still be reported at the terminological layer of the ontology. A further option would be to integrate the granularity level of the target culture in the common ontology,

[6] http://art.uniroma2.it/software/OntoLing/

but, here again, a certain compromise would be necessary. However, if there are more than two or three cultures and languages involved in the multilingual ontology the suggested option will not be an optimal one. In that case, one possible solution could be to include specific language modules in the ontology, and support different linearizations or visualizations of the same ontology according to the language selected. To the best of our knowledge, currently there is no system to support this latter option. Therefore, our proposal is to account for those categorization mismatches in an elaborated model of lexical and terminological information, separated from the ontology.

4. *hydrOntology*: an ontology of the hydrographical domain

hydrOntology [24] is an ontology in OWL that follows a top-down development approach. Its main goal is to harmonize heterogeneous information sources coming from several cartographic agencies and other international resources. Initially, this ontology was created as a local ontology that established mappings between different data sources (feature catalogues, gazetteers, etc.) of the Spanish National Geographic Institute (IGN-E). Its purpose was to serve as a harmonization framework among Spanish cartographic producers. Later, the ontology has evolved into a global domain ontology and it attempts to cover most of the concepts of the hydrographical domain.

hydrOntology has been developed according to the ontology design principles proposed by [10] and [2]. Some of its most important characteristics are that the concept names (classes) are sufficiently explanatory and are correctly written. Thus each class tries to group only one concept and, therefore, classes in brackets and/or with links ("and", "or") are avoided. According to certain naming conventions, each class is written with a capital letter at the beginning of each word, while object and data properties are written with lower case letters.

In order to develop this ontology following a top-down approach, different knowledge models (feature catalogues of the IGN-E, the Water Framework European Directive, the Alexandria Digital Library, the UNESCO Thesaurus, Getty Thesaurus, GeoNames, FACC codes, EuroGlobalMap, EuroRegionalMap, EuroGeonames, several Spanish Gazetteers and many others) have been consulted; additionally, some integration issues related to geographic information and several structuring criteria [25] have been considered. The aim was to cover most of the existing GI sources and build an exhaustive global domain ontology. For this reason, the ontology contains one hundred and fifty (150) relevant concepts related to hydrography (e.g. river, reservoir, lake, channel, and others), 34 object properties, 66 data properties and 256 axioms.

Currently, the *hydrOntology* ontology is available in two versions. The first one in Prótége makes use of the RDF(S) labeling model to document the ontology in natural language. In a subsequent stage, the ontology was associated to the Linguistic Information Repository (LIR) model, currently supported in the NeOn Toolkit. The first version of the ontology is available in Spanish and English, whereas in the second version two more languages were added: French and Catalan, as will be reported in section 6.

Regarding the first version of the ontology, *hydrOntology* was originally developed in Spanish, and therefore, the *labels* given to the concepts in the original ontology were in Spanish. Later on, English *labels* were also related to ontology concepts, and the language of those labels was specified by means of language tags. Definitions or glosses describing the concepts were also included in Spanish and English, if available, by making use of the *comment* property. Finally, one metadata element of Dublin Core (*source*) and one additional annotation (*provenance*) were used to report about the resources from which the different definitions (*comments*) and *labels* had been obtained, respectively. It must be noted that the process of documentation was not systematically carried out for different reasons, and not all types of annotations are available for every concept.

A snapshot of the class hierarchy of *hydrOntology* in the Protégé ontology editor can be seen in Figure 4. The concept *Río* (River) has been chosen for illustration. It has nine annotations related to it: three *provenance* annotations, two *comment* annotations, three *label* annotations, and one *source* annotation. As already reported, the *provenance* annotation gives information about the linguistic resources (glossaries, thesauri, dictionaries, etc.) labels have been obtained from. Since there are no mechanisms for relating the *label* (e.g. River) with its source of *provenance* (e.g. Water Framework Directive), the authors have decided to include the *label* in the provenance text for the sake of clarity (e.g.; "River – Water Framework Directive. European Union"@en).

Two *comments* are included, one in Spanish, and one in English, though no relation to any of the *labels* is given. Finally, three *label* annotations are given: two in Spanish (in addition to the one given in the URI, i.e., *Río*) and one in English. The two additional labels are *Curso de agua principal* (Main Watercourse), and *Curso fluvial* (Watercourse). According to the authors, the main difference among the three synonyms is the discourse register. The label *Río* is the general word, and would appear in general documents, whereas the other two additional labels would only come up in technical documentation managed by experts in the domain. It is worth noting that such fine-grained aspects could be relevant for certain indexing or information extraction tasks, but cannot be made explicit in the RDF(S) labeling model.

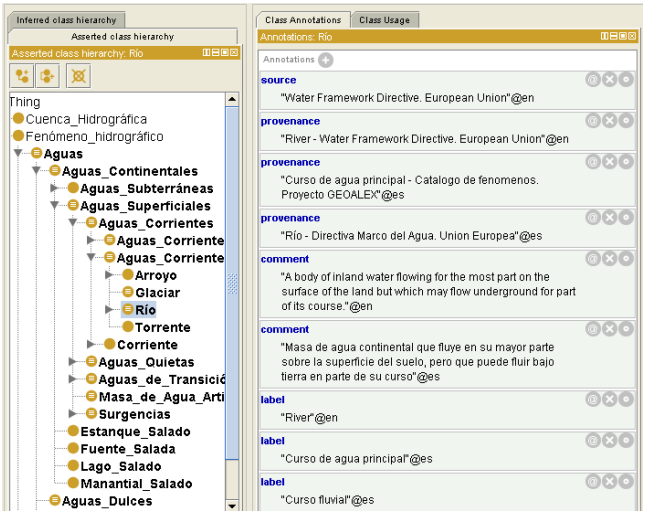

Figure 4. Snapshot of *hydrOntology* and the linguistic information associated to the *Río* ontology concept

Regarding the English translation, *River*, it is not possible to know to which of the Spanish labels is related to or is translation of. *River* is considered to be in a complete equivalence relation with *Río*, which would be appropriate in this case, but it is rarely the case, as explained in section 2. However, the RDF(S) labeling model does not offer any means to report about those cultural

differences that, more often than not, occur between two languages.

Because of these deficiencies in the representation of multilinguality in ontologies in OWL, and with the aim of giving response to the increasing demand for multilingual ontologies, the *Linguistic Information Repository* (LIR) model was developed. In the next section, we present the LIR model and how it aims at solving some of the representation problems identified so far.

5. LIR, a model for structuring the linguistic information associated to ontologies

The *Linguistic Information Repository* or LIR is a proprietary model expected to be published and used with domain ontologies. In itself, it has also been implemented as an ontology in OWL. Its main purpose is not to provide a model for a lexicon of a language, but to cover a subset of linguistic description elements that account for the linguistic realization of a domain ontology in different natural languages. A complete description of the current version of the LIR can be found in [14].

The lexical and terminological information captured in the LIR is organized around the Lexical Entry class. Lexical Entry is considered a union of word form (Lexicalization) and meaning (Sense). This ground structure has been inspired by the Lexical Markup Framework (LMF). The compliance with this standard is important for two main reasons: (a) links to lexicons modeled according to this standard can be established, and (b) the LIR can be flexibly extended with modular extensions of the LMF (or standard-compliant) modelling specific linguistic aspects, such as deep morphology or syntax, not dealt by LIR in its present stage. For more details on the interoperability of the LIR with further standards see [18].

The rest of the classes that make up the LIR are Language, Definition, Source, Note and Usage Context (see Figure 5). These can be linked to the Lexicalization and Sense classes. Each Lexicalization is associated to one Sense. The Sense class represents the meaning of the ontology concept in a given language. It has been modelled as an empty class because its purpose is to guarantee interoperability with other standards. The meaning of the concept in a certain language (which may not completely overlap with the formal description of the concept in the ontology) is "materialized" in the Definition class, i.e., is expressed in natural language. The Usage Context gives us information about how a word behaves syntactically in a certain language by means of examples. Source information can be attached to any class in the model (Lexicalization, Definition, etc.), and, finally, the Note class has been meant to include any information about language specificities, connotations, style, register, etc., and can be related to any class. By determining the Language of a Lexical Entry, we can ask the system to display only the linguistic information associated to the ontology belonging to a given language.

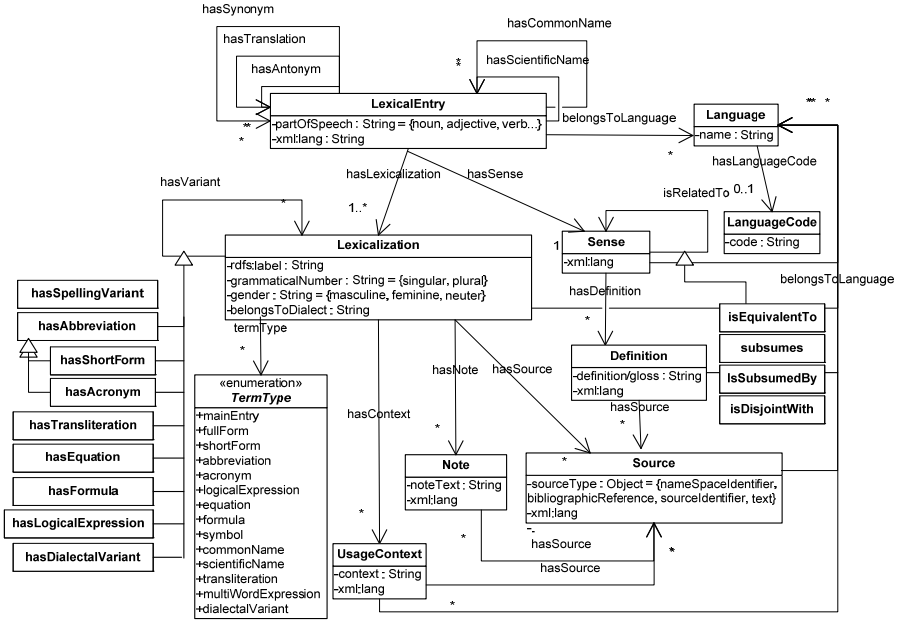

Figure 5. Overview of the LIR model with extensions to the isRelatedTo relation

Thanks to this set of linguistic descriptions, the LIR is capable of managing lexicalizations within one language, and their translations to other languages. Relations of synonymy can be expressed among lexicalizations in the same language, and the preferred lexicalization can be determined (main Entry), as well as other term variant relations (such as Acronym, Multi Word Expression or Scientific Name). Finally, relations of translation equivalence can be established among lexicalizations in different languages.

However, as we stated previously, more often than not lexicalizations in different languages are not exact equivalents, because the senses they represent do not completely overlap in their intensional and/or extensional descriptions. In order to account for cultural and linguistic specificities of languages, we propose an extension of the LIR to allow declaring semantic relations among the senses (Sense) of lexicalizations within and across languages. The semantic relations identified with this purpose are: equivalence (isEquivalentTo), subsumption (subsumes or isSubsumedBy), or disjointness (isDisjointWith).

So, the relation *isRelatedTo* that currently links senses (Sense) in the model is further specified.

6. Modeling multilinguality in *hydrOntology* with LIR

The current version of the LIR is supported by the LabelTranslator system[7], a plug-in of the NeOn Toolkit[8].As soon as an ontology is imported in the NeOn Toolkit, the whole set of classes captured in the LIR is automatically associated to each Ontology Element, specifically, to ontology classes and properties, by means of the relation "has Lexical Entry". In this way, the rest of linguistic classes organized around the Lexical Entry class are linked to an ontology element.

LabelTranslator [8] has been created for automating the process of ontology localization. Ontology Localization consists in adapting an ontology to the needs of a concrete linguistic and cultural community, as defined in [20]. Currently, the languages supported by the plug-in are Spanish, English and German. Once translations are obtained for the labels of the original ontology, they are stored in the LIR model. However, if the system does not support the language combination we are interested in, we can still use it to take advantage of the LIR API implemented in the NeOn Toolkit. In this sense, we can manually introduce the linguistic information necessary for our purposes.

As already mentioned, in the second version of *hydrOntology*, our purposes were to enrich the ontology in French and Catalan. With this aim, we imported the ontology originally documented in Spanish in the NeOn Toolkit, and automatically, all the linguistic classes of the LIR were associated to the concepts and properties in the ontology. The linguistic information associated to the URI of ontology concepts and properties in the original ontology automatically instantiated the LIR classes, i.e., a Lexical Entry was created for each ontology element, with its corresponding identifier (e.g., LexicalEntry-1), the Language of the label was identified and instantiated (e.g., Spanish), and a Lexicalization related to the Lexical Entry was also instantiated with the label in Spanish (e.g., Río). The rest of the linguistic information contained in the original ontology was not imported by the tool, and this fact was reported to the developers.

The next step was to manually introduce the labels in English, already available in the Protégé version of *hydrOntology*,. Since not all the concepts had been originally translated into English, we decided to make use of the LabelTranslator system to semi-automatically obtain translations for the original labels. The process was carried out in a semi-automatic way, and the translation candidates returned by LabelTranslator were evaluated by a domain expert. Since the purpose of this paper is not to evaluate the LabelTranslation plug-in, we will only refer to some of the results by way of example. To obtain more information about the experimental evaluation conducted with this tool, we refer to [8]. A table summarizing the results has been included in the Annex section[9] (see Table 1).

Then, the following step was the enrichment of the ontology with information in French and Catalan. Since these languages are not supported by LabelTranslator, we resorted to authoritative terminological resources in the domain [10], and manually introduced the information in the LIR by means of the LIR API (see Figure 6). For the sake of comparison, we will illustrate the results by taking the concept *River* as example, as in the case of the Protégé version of *hydrOntology*.

As shown in Figure 6, seven Lexical Entries with Part of Speech *noun* were associated to the concept *Río*: three in Spanish, one in English, one in Catalan and two in French. By clicking on each Lexical Entry we are able to visualize the rest of linguistic information associated to it: Lexicalizations, Senses, Usage Contexts, Sources and Notes.

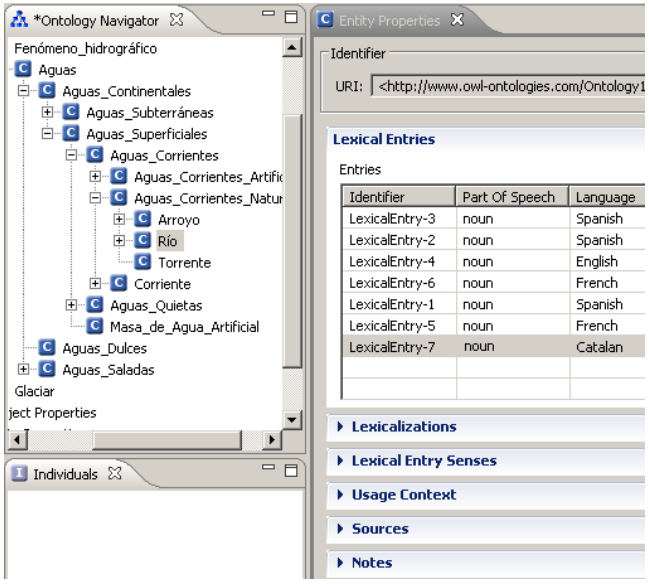

Figure 6. Linguistic Information associated to the concept *Río* in the LIR API (supported by LabelTranslator)

The three Lexical Entries in Spanish (*Río*, *Curso de agua principal*, and *Curso fluvial*) are related by means of the hasSynonym relation (see Figure 7 for Lexical Entry Relationships). The differences in use depending on register (formal vs. informal) are explained in the Note class. With the new extension to the LIR that we propose in this paper, the Senses of these Lexical Entries could additionally be related by an equivalence relation (isEquivalentTo).

Then, the three Lexical Entries in Spanish are related to the Lexical Entry in English (*River*), the one in Catalan (*Riu*), and the last two in French (*Rivière* and *Fleuve*) by means of the hasTranslation relation (see Figure 7). The Lexical Entry in English and the Lexical Entries in Spanish are considered equivalents in meaning, and the same happens with the Catalan equivalent. Therefore, their senses could also be related by the equivalence relation (isEquivalentTo).

[7] http://neon-toolkit.org/wiki/LabelTranslator

[8] http://neon-toolkit.org/

[9] The reason for not obtaining correct translations for some ontology terms may be due to the fact that the resources currently accessed by the system are quite general.

[10] For instance, the *Diccionari de l'Enciclopèdia Catalana* for the Catalan language (http://www.enciclopedia.cat), and the *Dictionnaire français d'hydrologie* for the French language (http://www.cig.ensmp.fr/~hubert/glu/indexdic.htm)

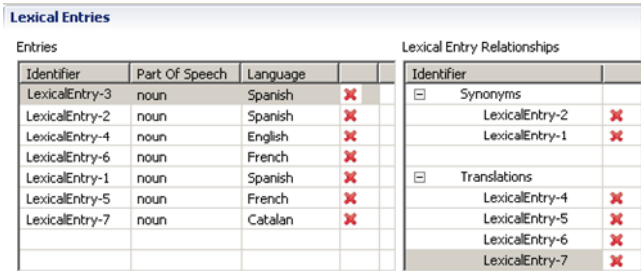

Figure 7. *Synonymy* and *Translation* Relationships among Lexical Entries

However, the two French Lexical Entries represent two more specific concepts that would stay in a relation of subsumption with the Spanish *Río*, the Catalan *Riu*, and the English *River*. This is an example of conceptual mismatch. The French understanding of river has a higher granularity level and identifies two concepts which are intensionally more specific, and extensionally do not share instances. These concepts are *Rivière* and *Fleuve*. According to the specialized resources accessed, *Rivière* is defined as a stream of water of considerable volume that flows into the sea or into another stream, and *Fleuve* is defined as a stream of water of considerable volume and length that flows into the sea. Therefore, in order to make explicit those differences in meaning, we relate them to two different Senses, and provide a definition in natural language for each of them (see Figure 6 for the Definition of *Rivière* in French). Then, with the new functionality of the LIR, we would establish a relation of subsumption between these two senses and the Spanish, English, and Catalan senses for *Río*, *River*, and *Riu* (isSubsumedBy).

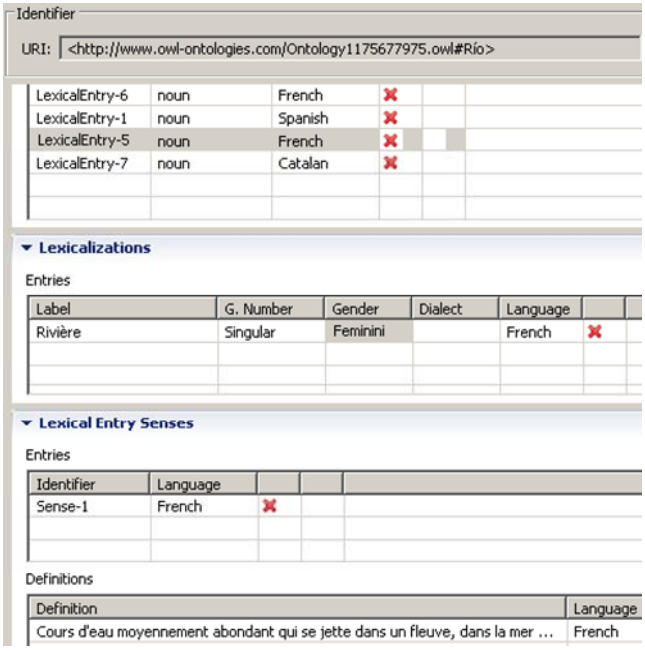

Figure 8. Lexicalization *Rivière* and its related Sense-1 and Definition in French

This further specification of the *isRelatedTo* relation among Senses allows accounting for categorization discrepancies among languages, which are not simply motivated by the fact that there are more lexicalizations in one language than in another, but by the different granularity levels that cultures make of the same world phenomenon. One could argue that these language specificities are only captured in the terminological layer of the ontology, but not in the conceptual model. However, this may suffice for certain ontology-based tasks such as information extraction or verbalization, whereas it may be insufficient for others. In that sense, a modification of the conceptualization to adapt the specificities of a certain language could be directly carried out by considering the lexical and terminological information contained in the LIR.

7. Conclusions

The aim of this paper has been twofold. On the one hand, we have discussed the difficulties involved in the interoperability of resources in the same domain created in different cultural settings, specifically because of the different granularity levels in which world phenomena are dealt with. We have described and illustrated the problematic issues of so called cultural dependent domains taking as example concepts of the hydrographical domain. On the other hand, our objective has been to compare two modalities for the representation of multilingual information in ontologies, with the aim of emphasizing the benefits of associating complex and sound lexical models to ontological knowledge. To achieve this we have presented in detail two versions of a multilingual ontology of the hydrographical domain, *hydrOntology*. The first version shows the representation possibilities offered by the OWL formalism to account for the multilingual information associated to ontology concepts in two languages: English and Spanish. The second version describes the representation possibilities of the Linguistic Information Repository (LIR), a proprietary model designed to associate lexical and terminological information in different languages to domain ontologies. Thanks to such a portable model, the lexical information can be structured for better exploitation purposes of ontology-based applications, and can account for linguistic and cultural discrepancies among languages.

8. ACKNOWLEDGMENTS

This work is supported by the Spanish R&D Project *Geobuddies* (TSI2007-65677C02) and the European Project *Monnet* (FP7-248458). We would also like to thank Óscar Corcho for valuable comments on a draft version of the paper.

9. REFERENCES

[1] Aguado de Cea, G., Montiel-Ponsoda, E., Ramos, J. A. (2007) Multilingualidad en una aplicación basada en el conocimiento TIMM, Monográfico para la revista SEPLN

[2] Arpírez JC, Gómez-Pérez A, Lozano A, Pinto HS (ONTO)2Agent: An ontology-based WWW broker to select ontologies. In: Gómez-Pérez A, Benjamins RV (eds) ECAI'98 Workshop on Applications of Ontologies and Problem-Solving Methods. Brighton, (UK), 1998, pp 16–24.

[3] Barrasa, J. Modelo para la definición automática de correspondencias semánticas entre ontologías y modelos relacionales. PhD Thesis, UPM, Madrid, Spain. 2007.

[4] Buitelaar, P., Cimiano, P., Haase, P. and Sintek, M. Towards Linguistically Grounded Ontologies. In Proceedings of the 6th Annual European Semantic Web Conference (ESWC2009), 111-125, 2009.

[5] DiMarco, Ch. Hirst, G. and Stede, M. The semantic and stylistic differentiation of synonyms and near-synonyms. In AAAI Spring Symposium on Building Lexicons for Machine Translation, 114–121, Stanford, CA, 1993.

[6] Edmonds, P. and Hirst, G. Near-synonymy and lexical choice. In Computational Linguistics, 28, 2, MIT Press, 105-144, 2002.

[7] Espinoza, M. Montiel-Ponsoda, E. and Gómez-Pérez, A. Ontology Localization. In Proceedings of the 5th Fifth International Conference on Knowledge Capture (KCAP), 33-40, 2009.

[8] Espinoza, M. Gómez-Pérez, A. and Mena, E. "Enriching an Ontology with Multilingual Information", Proc. ESWC'08, Tenerife (Spain), Springer LNCS, pp. 333-347, 2008.

[9] Euzenat, J. et al., Results of the Ontology Alignment Evaluation Initiative' 09. ISWC workshop on Ontology Matching (OM-2009). 2009.

[10] Gruber T.R. Toward principles for the design of ontologies used for knowledge sharing, International Journal of Human-Computer Studies, 1995, v.43 n.5-6.

[11] Guarino, N.: Formal Ontology and Information Systems, in Guarino, N. (ed.), Proceedings of FOIS98, 1998.

[12] Isaac, A., Summers, E. (Eds.) SKOS Simple Knowledge Organization System Primer. W3C, 2009. http://www.w3.org/TR/skos-primer/

[13] Mark, D. M., and Turk, A. G. Landscape Categories in Yindjibarndi: Ontology, Environment, and Language. In Kuhn, W., Worboys, M., and Timpf, S., Editors, Spatial Information Theory: Foundations of Geographic Information Science, LNCS No. 2825, Springer. 2003, pp. 31-49.

[14] Montiel-Ponsoda, E., Peters, W., Aguado de Cea, G., Espinoza, M. Gómez-Perez, A. and Sini, M. Multilingual and Localization support for ontologies. Technical report, D2.4.2 NeOn Project Deliverable, 2008.

[15] Montiel-Ponsoda, E., Aguado de Cea, G., Gómez-Pérez, A., and Peters, W. Modelling multilinguality in ontologies. In Coling 2008: Companion volume – Posters and Demonstrations, Manchester, UK, 67-70, 2008.

[16] Nowak, J., Nogueras-Iso, J., Peedell, S. Issues of multilinguality in creating a European SDI – The perspective for spatial data interoperability, in: Proceedings of the 11th EC GI & GIS Workshop, ESDI Setting the Framework. Alghero, Italy. 2005.

[17] Oltramari, A and Stellato, A. Enriching ontologies with linguistic content: An evaluation framework. In Proceedings of OntoLex 2008 Workshop at 6th LREC Conference in Marrakech, Morocco, 2008

[18] Peters, W. Gangemi, A. and Villazón-Terrazas, B. Modelling and re-engineering linguistic/terminological resources. Technical report, D2.4.4 NeOn Project Deliverable, 2010.

[19] Peters, W., Montiel-Ponsoda, E., Aguado de Cea, G., and Gómez-Pérez, A. Localizing ontologies in OWL. In Proceedings of the OntoLex07 Workshop at the ISWC in Busan, South Corea, 2007.

[20] Súarez-Figueroa, M.C. and Gómez-Pérez, A. A First Attempt towards a Standard Glossary of Ontology Engineering Terminology. In Proceedings of the 8th International Conference on Terminology and Knowledge Engineering (TKE2008), Copenhagen, 2008.

[21] Suarez-Figueroa, M.C. (coordinator). NeOn Development Process and Ontology Life Cycle. NeOn Project Deliverable 5.3.1 (2007).

[22] Tanasescu, V. Spatial Semantics in Difference Spaces, COSIT 2007, Melbourne, Australia. 2007.

[23] Thomson, M.K. and Béra, R. Relating Land Use to the Landscape Character: Toward an Ontological Inference Tool. In Winstanley, A. C. (Ed): GISRUK 2007, Proceeding of the Geographical Information Science Research UK Conference, Maynooth, Ireland, 2007, pp.83-87

[24] Vilches-Blázquez, L. M., Ramos, J. A., López-Pellicer, F. J., Corcho, O., Nogueras-Iso, J. An approach to comparing different ontologies in the context of hydrographical information. Popovich et al., (eds.): IF&GIS'09. LNG&C Springer. Pages: 193-207, 2009 St. Petersburg, Russia.

[25] Vilches-Blázquez L.M., Bernabé-Poveda M.A., Suárez-Figueroa M.C., Gómez-Pérez A., Rodríguez-Pascual A.F. "Towntology & hydrOntology: Relationship between Urban and Hydrographic Features in the Geographic Information Domain". In Ontologies for Urban Development. Studies in Computational Intelligence, Springer. 2007, vol.61, pp73–84

ANNEX

Table 1. Results of the semi-automatic translation with LabelTranslator

LabelTranslator	1st Candidate Translation	Candidate Translation	No Candidate Translation	Correct Translation
Aguas Continentales		✓		Inland waters
Aguas Subterráneas	✓			Groundwater
Acuífero	✓			Aquifer
Aguas superficiales		✓		Surface waters
Aguas corrientes			×	Flowing waters
Aguas Corrientes artificiales			×	Artificial flowing waters
Acequia		✓		Irrigation ditch
Canal	✓			Channel
Conducto			×	Pipe
Aguas Corrientes naturales			×	Natural flowing water
Arroyo	✓			Brook
Glaciar	✓			Glacier
Charca		✓		Pond
Torrente	✓			Torrent
Embalse	✓			Reservoir
Afluente		✓		Tributary
Guadi			×	Wadi
Aguas de transición			×	Transitional waters
Terma			×	Hot spring
Aguas costeras		✓		Coastal water
Mar	✓			Sea
Ribera	✓			Riverbank

Word order based analysis of given and new information in controlled synthetic languages

Normunds Grūzītis
Institute of Mathematics and Computer Science, University of Latvia
Raina bulv. 29, Riga, LV-1459, Latvia
normundsg@ailab.lv

ABSTRACT
When an OWL ontology, together with SWRL rules, is defined or verbalized in controlled natural language (CNL) it is important to ensure that the meaning of CNL statements will be unambiguously (predictably) interpreted by both human and machine. CNLs that are based on analytical languages (namely, English) impose a number of syntactic restrictions that enable the deterministic interpretation. Similar restrictions can be adapted to a large extent also for synthetic languages, however, a fundamental issue reveals in analysis of given (topic) and new (focus) information. In highly analytical CNLs, detection of which information is new and which has been already introduced is enabled by systematic use of definite and indefinite articles. In highly synthetic languages, articles are not typically used. In this paper we show that topic-focus articulation in synthetic CNLs can be reflected by systematic changes in word order that are both intuitive for a native speaker and formal for the automatic parsing.

Categories and Subject Descriptors
I.2.1 [**Artificial Intelligence**]: Applications and Expert Systems – *Natural language interfaces*; I.2.7 [**Natural Language Processing**]

General Terms
Design, Experimentation, Languages

Keywords
Ontology Verbalization, Controlled Natural Language, Synthetic Language, Word Order, Information Structure, Topic-Focus Articulation, Anaphoric References

1. INTRODUCTION
One of the fundamental requirements in definition and verbalization of ontology structure, restrictions, and implication rules is the unambiguous interpretation (in terms of the underlying formalism) of controlled natural language (CNL) statements, so that the CNL user could easily predict the precise meaning of the specification he/she is writing or reading; that also includes the resolving of anaphoric references. To enable deterministic construction of discourse representation structures (DRS), several widely accepted restrictions are used in CNLs (e.g., in Attempto Controlled English [2]): a set of interpretation rules for potentially ambiguous syntactic constructions (an issue that is still present even in a highly restricted syntactic subset of natural language), a monosemous lexicon (i.e., domain-specific terminology), an assumption that the antecedent of a definite noun phrase (NP) is the most recent and most specific accessible NP that agrees in gender and number, and some other limitations.

There are several sophisticated CNLs that provide seemingly informal means for bidirectional mapping between controlled English and OWL [10]. Experiments show that the underlying principles of English-based CNLs can be successfully adapted also for other rather analytical languages, for example, for Afrikaans [7]. Moreover, Ranta and Angelov [8] have shown that the Grammatical Framework (GF), a formalism for implementation of multilingual CNLs, provides convenient means for writing parallel grammars that simultaneously cover similar syntactic fragments of several natural languages. Thus, if the abstract and concrete grammars are carefully designed, GF provides a syntactically and semantically precise translation from one CNL to another (again, assuming that the domain-specific translation equivalents are monosemous). This potentially allows exploitation of powerful tools that are already developed for one or the other "dialect" of controlled English also for non-English CNLs. For instance, the ACE parser [2] could be used for DRS construction, paraphrasing and mapping to OWL, and ACE verbalizer [5] could be used in the reverse direction, facilitating cross-lingual ontology development, verbalization, and querying.

While it seems promising and straightforward for rather analytical CNLs that share common fundamental characteristics, allowing (apart from other) for explicit detection of anaphoric references, it raises issues in the case of highly synthetic languages, where explicit linguistic markers, indicating which information is already given (anaphors) and which is new (antecedents), in general, are not available (here we are talking about individuals that are referenced by NPs, not by anaphoric pronouns). In analytical CNLs, analysis of the information structure of a sentence is based on the strict word order (basically, the subject-verb-object or SVO pattern) and systematic use of definite and indefinite articles. In highly synthetic languages, articles are rarely used and are "compensated" by more implicit linguistic markers; typically, by changes in the neutral word order, which is enabled by rich inflectional paradigms and syntactic agreement.

As a case study, we have chosen Latvian [6], a member of the Baltic language group (together with Lithuanian). Baltic languages are among the oldest of the remaining Indo-European languages. Syntactically they both are very closely related and are highly synthetic with rich morphology; however, the definiteness feature is not encoded even in noun endings as it is in case of Bulgarian [1], for instance. Thus, we will describe the correspondence between the given/new information and word order patterns in terms of topic-focus articulation [3]. Although the topic (theme) and focus (rheme) parts of a sentence in general

WWW 2010, April 26-30, 2010, Raleigh, North Carolina.

Latvian are not always reflected by systematic changes in the word order [9], in this paper we demonstrate that changes in word order are reliable markers in the case of controlled Latvian, allowing for systematic reconstruction of "missing" articles.

Few other markers may be used in Latvian to indicate that an NP is an anaphoric reference, namely, definite and indefinite endings of adjectives and participles, if they are used as attributes. However, such markers are optional and non-reliable even in controlled language — attributes in domain-specific terms (multi-word units) often have definite endings by default. We might also impose the usage of artificial determiners, using indefinite and demonstrative pronouns, but then it would be Latvian-like controlled language, not a subset of actual Latvian. The problem is even more apparent in case of Lithuanian that has not been historically influenced by German. Therefore the only formal and general feature that indicates the status of an NP is its position in a sentence — whether it belongs to the topic or focus part. Our hypothesis is that the requirement for compliance with predefined word order patterns in a controlled synthetic language is not only reasonable, but also makes the CNL more natural and is intuitively satisfiable by a native speaker. As our experiments show, the proposed approach is directly applicable for Lithuanian and can be adapted also for majority of Slavic languages (e.g., Russian and Czech) — the closest siblings to Baltic languages.

2. TERMINOLOGICAL STATEMENTS

In this paper we focus on terminological (TBox) statements of OWL ontologies [15] that are supplemented with limited data integrity constraints in form of SWRL rules [18] and SPARQL queries [16, 17] (see Section 3). In this section we will consider different types of statements defining atomic and complex classes, properties, and property restrictions of a simplified university ontology. Statements are given in parallel in Manchester OWL Syntax [4], ACE [2], and ACE compliant controlled Latvian.

In Figure 1 the example ontology is visualized according to the UML profile for OWL [14] — a user-friendly notation that unveils the structure of the ontology in a highly comprehensible form, but it is not well suited to capture complex restrictions and integrity constraints. In the following two sections we will do both verbalize the UML-defined structure and use CNL to define additional restrictions and integrity constraints.

2.1 Classes

Statements defining class hierarchies consist of subject and subject complement. Subject (topic) is always universally quantified, predicate noun — always existentially quantified:

(1) `Class: Professor SubClassOf: Teacher`
Every professor is a teacher.
Katrs profesors ***ir*** ~~***kāds***~~ ***pasniedzējs***.

For the universal quantifier there is a corresponding determiner/pronoun both in English and Latvian (as well as in other analytical and synthetic languages). As to the existential quantifier there is no counterpart for the indefinite determiner in Latvian. We could artificially use an indefinite pronoun instead, but such construction would be more than odd in this case. Besides the fact that the subject complement is always indefinite it also always appears in the focus part of a sentence (here and further — formatted in bold), i.e., it always is new information and, thus, the explicit linguistic marker (indefinite pronoun) can be omitted without introducing any ambiguities. Similarly, class equivalence can be defined by stating subclass axioms in both directions (in two separate statements), and class disjointness — by substituting the determiner "every" with its antonym "no":

(2) `DisjointClasses: Assistant, Professor`
No assistant is a professor.
Neviens asistents ***nav profesors***.

In Latvian (and in many other synthetic languages), a negated pronoun is used for the negative universal quantifier, and the statement is negated twice by the copula, but these are minor syntactic differences; the information structure remains the same. This assumption can be directly extended to complex classes that are combined from atomic ones by applying logical constructors:

(3) `Class: Course SubClassOf: owl:Thing and (MandatoryCourse or OptionalCourse)`
Every course is something that is a mandatory course or that is an optional course.
Katrs kurss ***ir kaut kas****, kas* ***ir obligātais kurss*** *vai kas* ***ir izvēles kurss***.

So far about cases when the verb phrase (VP) is a predicate nominal. Another type of constructors for complex classes are property restrictions — VPs consisting of a transitive verb complemented by a direct object. In the following statement such VP is used to implicitly specify an anonymous superclass.

(4) `Class: Teacher SubClassOf: teaches some Course`
Every teacher teaches a course.
Katrs pasniedzējs ***pasniedz*** [*kādu*] ***kursu***.

Here the role of word order comes in. In Latvian, if the object comes after the verb (the neutral word order) it belongs to the focus part of the sentence (new information), but if it precedes the verb — to the topic part (given information). In the case of the inverse use of a property, the word order is changed for the whole statement (in both languages), moving the agent to the focus:

(5) `Class: Course SubClassOf: inverse (teaches) some Teacher`
Every course is taught by a teacher.
*Katr****u*** *kurs****u*** ***pasniedz*** [*kād****s***] ***pasniedzēj****s*.

Combinations of the introduced syntactic phrases can be further used to explicitly specify complex superclasses (Statement 6) and general class axioms, where anonymous is either the subclass (Statement 7), or both the super- and the sub-class (Statement 8).

(6) `Class: Student SubClassOf: Person and (inverse (enrolls) some AcademicProgram)`
Every student is a person that is enrolled by an academic program.
Katrs students ***ir persona****, ko* ***uzņem*** [*kāda*] ***akadēmiskā programma***.

(7) `Class: owl:Thing and (teaches some MandatoryCourse) SubClassOf: Professor`
Everything that teaches a mandatory course is a professor.
Katrs, kas ***pasniedz*** [*kādu*] ***obligāto kursu****,* ***ir profesors***.

(8) `Class: owl:Thing and (inverse (includes) some AcademicProgram) SubClassOf: inverse (teaches) some Teacher`
Everything that is included by an academic program is taught by a teacher.
Katru, ko ***ietver*** [*kāda*] ***akadēmiskā programma****,* ***pasniedz*** [*kāds*] ***pasniedzējs***.

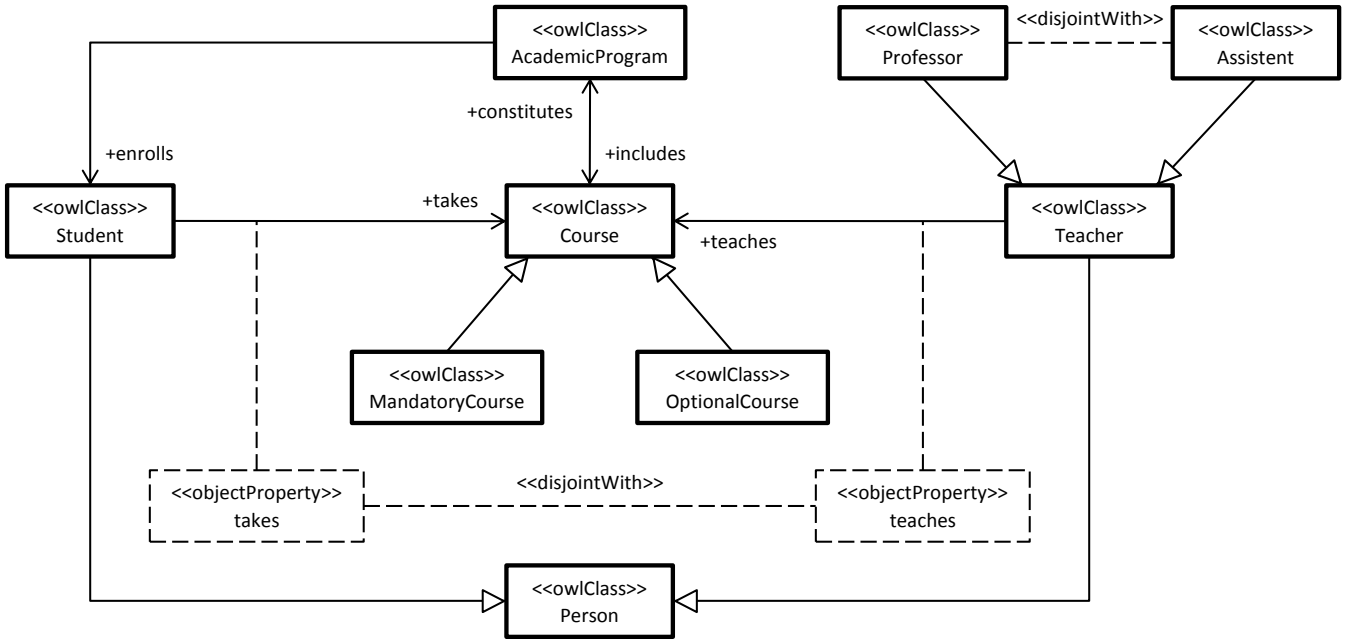

Figure 1. The structure of a simplified university ontology, visualized by the OWL2UML plug-in [19]. No complex class expressions or data integrity constraints are included. The automatically generated diagram is also slightly simplified for printing purposes.

In English, both active and passive voice sentences are still SVO sentences; the inverse direction of the property is indicated by the passive voice. In Latvian, the voice remains active, but the syntactic functions of both NPs are interchanged in such a case, making it an OVS sentence (semantic roles remain the same in both languages). Thus, it turns out that the object stands to the left from the verb — in the topic part, indicating that it should be given information. However, recall that in TBox statements there is no doubt which determiner has to be assigned with the topic — it is always universally quantified (unless it is an anaphoric pronoun that links a relative clause to its anchor).

Also, it should be mentioned that in the above provided SVO statements (4–8) the indefinite pronoun "kāds" is given in square brackets, which means that in these cases it might be optionally used as a counterpart for the indefinite article. This conforms to the language intuition (see Section 4) and emphasizes the indefiniteness of the NP — the explicit marker improves readability (interpretation) as there is no relative clause associated with the NP, which would then serve as an indicator of indefiniteness.

Another aspect that should be mentioned is that there has not been made any differentiation between animate and inanimate things — quantifiers "everything", "something", "nothing", and the relative pronoun "that" are used in all cases. This makes Statement 7 in English and Statement 8 in Latvian odd, which has been noticed by our respondents (see Section 4). However, we ignore this issue in this paper for the sake of simplicity and for compliance with the ACE verbalizer [5] (although ACE parser supports such differentiation, it is discarded at the ontological level).

2.2 Properties

The already introduced syntactic constructions can also be used to define properties and their restrictions. Thus, to specify the domain and range of a property, in the topic part of the statement one has to refer the universal class, which is then specified in the relative clause, referring the property of interest. The subject is complemented by a predicate nominal in the focus part (see the statements 9–10). Note that usage of definite and indefinite NPs (i.e., references to concrete classes) is not possible in property definitions (except when stating the domain and range), as this would go beyond the expressivity of OWL (see the next section).

(9) `ObjectProperty: teaches Domain: Teacher`
Everything that teaches something is a teacher.
Katrs, kas ***kaut ko pasniedz, ir pasniedzējs****.*

(10) `ObjectProperty: teaches Range: Course`
Everything that is taught by something is a course.
Katrs, ko ***kaut kas pasniedz, ir kurss****.*

Although property hierarchies, characteristics, and chains can be defined in the same manner, by additionally exploiting the anaphoric pronoun "it", in many cases usage of if-then constructions together with variables is at least more concise, if not more comprehensible as well (especially in property chaining). Exceptions are functional and inverse functional properties that are defined by cardinality restrictions and can be stated more naturally without anaphoric references, and reflexive and irreflexive properties — their verbalization in CNL is more natural if the reflexive pronoun "itself" is used.

For instance, the definition of a property chain given in Statement 11 could be paraphrased by using pronouns instead of variables — "*Everything that includes something that is taken by something enrolls it.*" — but such paraphrase would more likely confuse a human interpreter, especially when resolving the anaphoric reference.

(11) `ObjectProperty: enrolls SubPropertyChain: includes o inverse (takes)`
If [*something*] *X includes something that is taken by* [*something*] *Y then X enrolls Y.*
Ja [*kaut kas*] *X ietver kaut ko, ko ņem* [*kaut kas*] *Y, tad X uzņem Y.*

Note that the pronoun "something" may be omitted in the apposition phrases — this not only makes the statements more comprehensible, but also allows to reduce or even to hide the issue of discrimination between animate and inanimate things (accordingly, X and Y in Statement 11). For instance, when declaring that two properties are disjoint, we can avoid the use of the indefinite pronoun at all:

(12) `DisjointProperties: teaches, takes`
If X teaches Y then X does not take Y.
Ja X pasniedz Y, tad X neņem Y.

The concise form, however, has a drawback in the case of a highly synthetic CNL. In English, the strict word order (and the change of the voice) enables the unambiguous detection of which variable represents the agent in any of the SVO chunks. In Latvian, provided that all the properties involved are represented by transitive verbs (instead of comparative phrases, for instance, as in "*X is smaller than Y*"), the agent/subject can be recognized only due to the different ending if compared with the object; the verb itself does not change. Plain variables, of course, are not inflected. Although for a human interpreter it usually causes no ambiguities (due to the rich background knowledge, and knowledge of lexical semantics), suffixes have to be added to the variables to enable the automatic parsing. Nevertheless, this is still a more user-friendly solution (see Statement 13) than the use of the artificial apposition phrases. Moreover, even if indefinite apposition phrases are used, they are applicable only in the if-clauses; for the then-clauses definite apposition phrases should be introduced, making such statements even more unnatural.

(13) `ObjectProperty: includes InverseOf: constitutes`
If X includes Y then X is constituted by Y.
Ja X-s ietver Y-u, tad X-u veido Y-s.

Although property axioms can be seen as a special case, variables may be used in statements defining classes as well (e.g., Statement 7 in Section 2.1 can be paraphrased in ACE as "*If X teaches a mandatory course then X is a professor*."). The formal nature of CNL then becomes explicit more widely, losing the seeming naturalness that, of course, is not a self-purpose; variable constructions should be allowed as an alternative to improve readability in certain cases (e.g., for tracking coreferences in complex rules). Allowing for such alternatives, however, introduces an issue in the verbalization direction — how to decide (encode in the grammar) in which cases variables are preferable over indefinite and definite NPs.

Nevertheless, variable constructions are partially out of the scope of this paper, as there is no need for information structure analysis to cope with utterances of anaphoric pronouns and variables. Note that we have already violated the word order guidelines in some of the previous examples — in Latvian, the indefinite pronoun "kaut kas" typically goes before the verb if it is not specified by a relative clause (see the statements 9–10). Thus, formally it belongs to the topic, although it is always new information. But, again, this causes no ambiguities.

In overall, if we are restricting our synthetic CNL to cover terminological statements (class and property definitions) only, information structure analysis is not necessary at all: since OWL axioms are variable-free, any noun phrase that is not explicitly universally quantified is existentially quantified. However, we have now laid the foundations to extend controlled Latvian for support of data integrity constraints.

3. DATA INTEGRITY CONSTRAINTS

In this section we will add some implication rules and data integrity queries to our example ontology, making an actual exploitation of the topic-focus articulation (TFA) — when specifying integrity constraints, one cannot avoid the usage of variables or definite/indefinite NPs.

3.1 SWRL Rules

In SWRL rules [18], variables are used, which cause at least one anaphoric reference, when a rule is verbalized in CNL. In terminological statements, changes in the word order are caused only due to an inverse use of a property, but in the case of rules (verbalized in controlled Latvian), the word order has to be changed also to indicate whether an NP (the subject or the object) introduces a new individual or is an anaphoric reference:

(14) `Rule: Student(?x1), MandatoryCourse(?x2), AcademicProgram(?x3), enrolls(?x3, ?x1), includes(?x3, ?x2) -> takes(?x1, ?x2)`
*Every mandatory course that **is included by an academic program is taken by every student** that **is enrolled by the academic program**.*
*Katru obligāto kursu, ko **ietver** [kāda] **akadēmiskā programma, ņem katrs students**, ko [šī] akadēmiskā programma **uzņem**.*

In the above statement, inverse properties are used in both relative clauses, causing the swapping of the subject and the object. In the first case, the subject ("akadēmiskā programma") stands to the right from the verb, indicating that it belongs to the focus part — new information. In the second case, the subject (again, "akadēmiskā programma") stands to the left from the verb — in the topic part of the clause, indicating that this is already given information (a reference to the individual introduced in the first relative clause). Thus, in the latter relative clause, the property (verb) is alone in the focus — the new information is the relationship between the two already given individuals (the student and the academic program).

The English and Latvian verbalizations of the above rule are more clearly aligned in Figure 2 (the optional "articles" are not used).

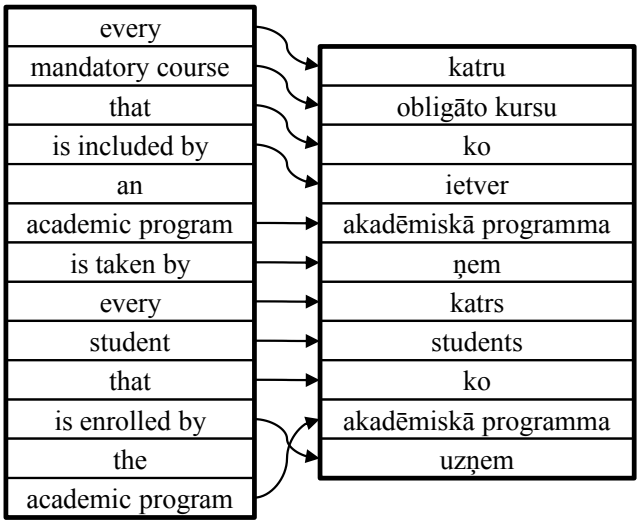

Figure 2. A word alignment graph (generated by the Grammatical Framework [8]), showing that given information in Latvian is reflected by changes in the neutral word order.

To emphasize that an NP is an anaphoric reference, we could optionally use the demonstrative pronoun "šis" ("this"), which usually sounds natural in Latvian. Theoretically, this would allow us to place the NP also to the right from the verb, but both stylistically and intuitively the preferable position is still to the left, which causes the correct intonation.

Let us consider one more rule (Statement 15) where the indefinite pronoun "kāds" is not anymore offered as an optional attribute. It is omitted due to the cascade of relative clauses that modify all the indefinite NPs.

(15)
```
Rule: Person(?x1), MandatoryCourse(?x2),
AcademicProgram(?x3), enrolls(?x3, ?x1),
includes(?x3, ?x2), takes(?x1, ?x2) ->
Student(?x1)
```

*Every person that **takes a mandatory course** that **is included by an academic program** that **enrolls the person is a student**.*

*Katra persona, kas **ņem obligāto kursu**, ko **ietver akadēmiskā programma**, kas* [*šo*] *personu **uzņem**, **ir students**.*

Due to the cases when the usage of the indefinite and demonstrative pronouns might improve the readability of a rule, an implementation of an additional concrete grammar of surrogate Latvian, where the usage of indefinite and demonstrative pronouns in the role of articles is always mandatory, would be a simple but naive trade-off. Such a coarse-grained grammar could be used to paraphrase (on demand) the purely TFA-based sentences, both user-provided and auto-generated (verbalized). Although reading of such surrogate statements perhaps is easier than writing, this would be confusing to the end-users anyway (see the next section). To provide more fine-grained paraphrases and to protect the users from a confusingly verbose look-ahead editor, the TFA-based grammar can be improved at least by distinguishing two types of NPs: those that are modified by relative clauses, and those that are not. In the latter case, usage of the indefinite or demonstrative pronoun is preferable.

However, if a statement is written in synthetic language, the best paraphrase could be its translation into analytical language (e.g., English), or alternatively (for an advanced user) — in a human-readable formal syntax like the Manchester OWL Syntax.

3.2 SPARQL Queries

When executing SWRL rules, potentially new facts are inferred and added to the ontology (ABox) whenever the body of a rule is satisfied. If this is the intention and if the rule (its verbalization) does not include negated atoms or disjunctions then it is the right choice; otherwise we would end up with unwanted entailments or would not be able to translate the statement into SWRL. For example, if we would try to redefine the property chaining defined in Statement 11 (in Section 2.2) as a more specific rule (by referring to the concrete classes) — "*Every student that takes a course that is included by an academic program is enrolled by the academic program.*" — the effect would be that a student, taking a course that is included by another academic program, is automatically enrolled by that program.

By asking SPARQL queries we can verify integrity constraints relying on the closed world assumption (negation as failure) [13], without introducing unintended entailments. Thus, Statement 11 can be alternatively (but not equally) specified in CNL as the following query:

(16)
```
ASK WHERE {
  ?x1 rdf:type Student.
  ?x1 takes ?x2.
  ?x2 rdf:type Course.
  ?x3 rdf:type AcademicProgram.
  ?x3 enrolls ?x1.
  NOT EXISTS {?x3 includes ?x2}}
```

Is there a student that takes a course that is not included by an academic program that enrolls the student?

*Vai **ir kāds students**, kas **ņem kursu**, ko **neietver akadēmiskā programma**, kas studentu **uzņem**?*

In the case of consistency checking, the ASK form of a query (yes/no question) is entirely appropriate and its verbalization syntax is not much different from that of rules. Note that in such queries the first NP is always indefinite (although appearing in the topic) and the corresponding indefinite pronoun is always explicitly attached to it.

The current SPARQL specification [16] does not directly provide an operator for negation as failure; it is possible by combining the OPTIONAL, FILTER and !BOUND operators. However, if the !BOUND operator is applied to a variable (in our example, x3) that is used also outside the OPTIONAL block then the result, of course, will not be what expected. Therefore, for the sake of simplicity, we have used the NOT EXISTS pattern that will be provided by the SPARQL 1.1 specification [17].

4. EVALUATION

To verify whether the proposed assumptions on which the proof-of-concept implementation [20] is based are linguistically motivated and universally applicable in highly synthetic CNLs, an initial evaluation was performed. About ten linguists (both Latvians and Lithuanians), specialized in Baltic languages, received nearly twenty examples, covering different types of statements and different levels of complexity. For each example several alternative translations were given in Latvian and Lithuanian in parallel (see Table 1). Among the alternative choices were the "literal" (surrogate) translation, the pure TFA-based translation, and a combination of the previous two, in order to seemingly improve the readability. The respondents were asked to sort the choices (in their native language) by priority from 1 to 3 (1 goes for the best translation of the original statement in ACE), or rejected at all (0). The respondents were introduced with the basic limitations of CNL, but they were also asked to follow their language intuition.

Table 1. Example statements, evaluated by a Lithuanian respondent. The English statement is the benchmark, the Latvian translations (in this case) — for comparison.

*Every student is **a** person that is enrolled by **an** academic program.*		
A_{LV}	*Katrs students ir **kāda** persona, ko uzņem **kāda** akadēmiskā programma.*	
A_{LT}	*Kiekvienas studentas yra **koks nors** asmuo, kurį priima **kokia nors** akademinė programa.*	**0**
B_{LV}	*Katrs students ir persona, ko uzņem akadēmiskā programma.*	
B_{LT}	*Kiekvienas studentas yra asmuo, kurį priima akademinė programa.*	**1**
C_{LV}	*Katrs students ir persona, ko uzņem **kāda** akadēmiskā programma.*	
C_{LT}	*Kiekvienas studentas yra asmuo, kurį priima **kokia nors** akademinė programa.*	**2**

The respondents were also invited to give an alternative translation for each example, if none of the proposed ones was enough satisfactory. This option was used rather frequently, resulting in some interesting suggestions. Those that can be systematized will be taken into account.

In overall, most of the literal translations, using the artificial "articles", were rejected or assigned with the lowest priority. There was no consensus, however, whether the indefinite and demonstrative pronoun in certain cases should be used to improve readability or not; even the same respondent usually did not act consistently among different examples. However, in most cases, the usage of the pronoun is preferred, if the NP is not modified by a relative clause.

It should be mentioned that almost all the respondents were disappointed with the uniform approach to animate and inanimate things. Although this is not directly related to the topic of this paper, this issue has to be taken into account, which means that one more feature has to be incorporated in the domain-specific lexicons (in the noun and pronoun entries) and exploited in the grammars. However, the issue will still remain, if other tools are used in the workflow (e.g., the ACE verbalizer).

Based on these results, an improved grammar is being developed, which will be evaluated by a wider audience.

5. CONCLUSION

We have shown that in controlled Latvian, which is a highly synthetic CNL, where definite and indefinite articles are not used, the topic-focus articulation can be reflected by systematic changes in the neutral word order. This provides a simple and reliable mechanism (guidelines) for deterministic (predictable) analysis of the information structure of a sentence, enabling automatic detection of anaphoric NPs. As the very initial evaluation confirms, native speakers tend to follow such guidelines rather intuitively. Moreover, in languages where the semantic and pragmatic aspects of the sentence are more studied [11], the general correlations between the word order and given/new information are being taught even in language learning courses for beginners [12].

At the time of writing, the proof-of-concept implementation of Latvian-English CNL covers most of the syntactic constructions that were introduced in the previously given examples. The aim for the near future is to extend the TFA-based grammar to cover the full expressivity of terminological statements and rules, while remaining compliant with ACE. Future work is (a) to introduce support for assertional statements — the problem is how to determine, whether the subject noun (in the case of the neutral word order) represents given or new information, (b) to make a more detailed investigation on data integrity queries that are important in practical applications and will make a rather extensive use of anaphoric references, and (c) to consider pros and cons for using the GF Resource Library [8].

6. ACKNOWLEDGMENTS

This research has been supported by the European Social Fund. The author would like to thank Gunta Nešpore, Baiba Saulīte, Guntis Bārzdiņš and Kārlis Čerāns for the valuable discussions on linguistic and semantic aspects, as well as the reviewers for the detailed comments and suggestions, and the respondents for their help in the initial evaluation of the proposed approach.

7. REFERENCES

[1] Angelov, K. Type-Theoretical Bulgarian Grammar. In *6th International Conference on Natural Language Processing* (Gothenburg, Sweden, 2008), LNCS/LNAI 5221, Springer.

[2] Fuchs, N.E., Kaljurand, K., and Schneider, G. Attempto Controlled English Meets the Challenges of Knowledge Representation, Reasoning, Interoperability and User Interfaces. In *19th International FLAIRS Conference* (Melbourne Beach, Florida, 2006), AAAI Press, 664--669.

[3] Hajičová, E. *Issues of Sentence Structure and Discourse Patterns*. Charles University, Prague, 1993.

[4] Horridge, M., Drummond, N., Goodwin, J., Rector, A., Stevens, R., and Wang, H. The Manchester OWL syntax. In *2nd International OWLED Workshop* (Athens, Georgia, 2006).

[5] Kaljurand, K., and Fuchs, N.E. Verbalizing OWL in Attempto Controlled English. In *3rd International OWLED Workshop* (Innsbruck, Austria, 2007).

[6] Nau, N. *Latvian*. (Languages of the World / Materials 217). Lincom, Munich, 1998.

[7] Pretorius, L., and Schwitter, R. Towards Processable Afrikaans. In *Workshop on Controlled Natural Language* (Marettimo Island, Italy, 2009), CEUR, vol. 448.

[8] Ranta, A., and Angelov, K. Implementing Controlled Languages in GF. In *CNL 2009 Workshop* (Marettimo Island, Italy, 2009), LNCS/LNAI 5972, Springer (to appear).

[9] Saulīte, B. Linguistic Markers of Information Structure in Latvian. In *18th International Congress of Linguists* (Seoul, Korea, 2008), The Linguistic Society of Korea, 3067--3076.

[10] Schwitter, R., Kaljurand, K., Cregan, A., Dolbear, C., and Hart, G. A Comparison of three Controlled Natural Languages for OWL 1.1. In *4th International OWLED Workshop* (Washington, DC, 2008).

[11] Sgall, P., Hajičová, E., and Panevová, J. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Reidel, Dordrecht, 1986.

[12] Short, D. *Teach Yourself Czech*. Hodder Education, London, 2003.

[13] Sirin, E., and Tao, J. Towards Integrity Constraints in OWL. In *6th International OWLED Workshop* (Chantilly, Virginia, 2009), CEUR Workshop Proceedings, vol. 529.

[14] Ontology Definition Metamodel. OMG Adopted Specification. http://www.omg.org/spec/ODM/1.0/ (2009)

[15] OWL 2 Web Ontology Language. W3C Recommendation. http://www.w3.org/TR/owl2-primer/ (2009).

[16] SPARQL Query Language for RDF. W3C Recommendation. http://www.w3.org/TR/rdf-sparql-query/ (2008)

[17] SPARQL Query Language 1.1. W3C Working Draft. http://www.w3.org/TR/sparql11-query/ (2010)

[18] SWRL: A Semantic Web Rule Language Combining OWL and RuleML. W3C Member Submission. http://www.w3.org/Submission/SWRL/ (2004).

[19] http://protegewiki.stanford.edu/index.php/OWL2UML

[20] http://eksperimenti.ailab.lv/cnl/

Metadata Synchronization between Bilingual Resources: Case Study in Wikipedia

Eun-kyung Kim
Korea Advanced Institute of Science and Technology
Yuseong-gu, Guseong-dong
Daejeon, Republic of Korea
kekeeo@world.kaist.ac.kr

Matthias Weidl
Universität Leipzig
Department of Computer Science
Johannisgasse 26,
D-04103 Leipzig, Germany
mam07jct@studserv.uni-leipzig.de

Key-Sun Choi
Korea Advanced Institute of Science and Technology
Yuseong-gu, Guseong-dong
Daejeon, Republic of Korea
kschoi@world.kaist.ac.kr

ABSTRACT

In this paper, we present a conceptual study aimed at understanding the impact of international resource synchronization in Wikipedia and DBpedia. In the absence of any information synchronization, each country would construct its own datasets and manage it from its users. Moreover the cooperation across the various countries is adversely affected. The solution is based on the analysis of Wikipedia infobox templates and on experimentation such as term translation.

Categories and Subject Descriptors

H.3.m [**Information Systems**]: Miscellaneous

General Terms

Experimentation

Keywords

Semantic Web, Multilingual, Wikipedia, DBpedia

1. INTRODUCTION

Wikipedia is an international project which is a web-based, free-content encyclopedia and is written collaboratively. Wikipedia has made tremendous effect in the web. It has grown rapidly into one of the largest reference web sites. The main advantage of using Wikipedia is its wide coverage of concepts and languages. Wikipedia currently comprises more than 260 languages. However, Wikipedia still lacks sufficient support for non-English languages. The 22% articles in all Wikipedias belong to the English language version. Although English has been accepted as a global standard to exchange information between different countries, companies and people, the majority of users are attracted by projects and web sites if the information is available in their native language as well. One can also assume that the proportion of users on the Internet who do not speak English will continue to rise.

Due to the differences in the amount of information between English and non-English languages in Wikipedia, there needs to not only avoid information loss, but also enrich informations.

WWW2010, April 26-30, 2010, Raleigh, North Carolina.

In this paper we presents our efforts to create a multilingual system for translation-based information synchronization. Our work is based on idea that Wikipedia as a multilingual corpus and especially translation results between different editions provide valuable multilingual resources to the web. To explore this problem, we developed *Metadata Synchronization*, a platform for translation-based data synchronization between English Wikipedia and Korean Wikipedia. It aims to translate infoboxes from the English Wikipedia into Korean and insert it into the Korean Wikipedia. Because Wikipedia offers a number of structural elements, in particular, the infobox template is used to express structured information about a condensed set of important facts relating to the article[10]. The infobox is manually created by authors that create or edit an article. As a result, many articles have no infoboxes and other articles contain infoboxes which are not complete. Moreover, even the interlanguage linked articles do not use the same infobox template or contain different amount of information. Interlanguage links are links from any page describing an entity in one Wikipedia language to a page describing the same subject in another language. This problem raises an important issue about multi-lingual access on the Web.

The rest of this paper is organized as follows: In section 2, we describe related work. The framework and details of the proposed approach are given in section 3. Section 4 and 5 discuss the experimentation and results. At the end, we summarize the obtained results and point our future work.

2. RELATED WORK

Wikipedia represents a valuable source of knowledge to extract semantic information between concepts. [9] focuses on research that extracts and makes use of the concepts, relations, facts and descriptions found in Wikipedia, and organizes the work into four broad categories: applying Wikipedia to natural language processing; using it to facilitate information retrieval and information extraction; and as a resource for ontology building.

In [5], the authors suggests a methodology that semantic information can be extracted from Wikipedia by analyzing the links between categories. They try to provide a semantic schema for Wikipedia which could improve its search capabilities and provide contributors with meaningful suggestions for editing the Wikipedia pages.

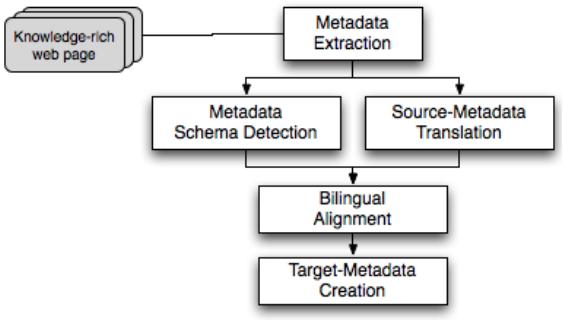

Figure 1: The workflow of the proposed metadata synchronization approach.

DBpedia[1][2] is a large, on-going, project which concentrates on the task of converting Wikipedia content into structured knowledge, and make it usable for the Semantic Web. An important component of DBpedia is harvesting of the information present in infoboxes. The infobox extraction algorithm detects such templates and recognizes their structure and saves it in RDF triples. DBpedia reached a high-quality of the extracted information and offers datasets of the extracted Wikipedia in 91 different languages. However, DBpedia still lacks sufficient support for non-English languages. First, DBpedia only extracts data from non-English articles that have an inter-language link to an English article. Therefore all other articles cannot be queried by DBpedia. Another reason is that the Live Extraction Server [7] only supports the timely extraction of the English Wikipedia. Due to the differences in the number of resources between English and non-English languages in DBpedia, there needs to be a synchronization among them.

[4] presents a method for cross-lingual alignment of template and infobox attributes in Wikipedia. The alignment is used to add and complete templates and infoboxes in one language with information derived from Wikipedia in another language.

In the area of cross-language data fusion, another project has been launched[6]. The goal is to extract infobox data from multiple Wikipedia editions and fusing the extracted data among editions. To increase the quality of articles, missing data in one edition will be complemented by data from other editions. If a value exists more than once, the property which is most likely correct will be selected.

3. FRAMEWORK OF METADATA SYNCHRONIZATION APPROACH

Our metadata synchronization algorithm consists of two consecutive steps: a metadata extraction step and a metadata translation step. Figure 1 illustrates the workflow of our approach. In order to illustrate our approach, we examine a case study in Wikipedia by using the Infobox as metadata. Wikipedia is considered to be one of the most successful collaborative editing communities on the web, we consider it as an interesting example to discuss multilingual synchronization.

In particular, Web pages pose main problems to multilingual resource:

[1]http://dbpedia.org/

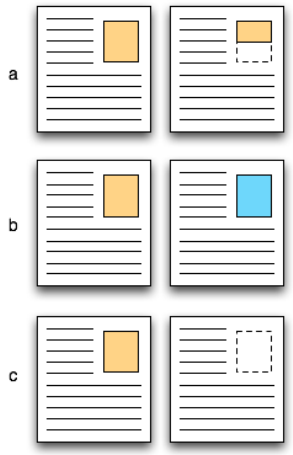

Figure 2: The form of the three template-pairs between different languages' infoboxes. (a) describes S-group, (b) describes D-group and (c) describes M-group.

- Because the number of pages in the non-English ares smaller than the English community.
- Because languages differ in grammar as well, it is obvious that every language uses its own kind of formats and templates.
- The majority of users are attracted by web sites if the information is available in their native language as well.

To explore this problem, we try to synchronize between the knowledge-poor web pages (The Korean Wikipedia: it consists of about 130,000 articles at March 2010.) and the knowledge-rich web pages (The English Wikipedia: 3.24M articles at March 2010) using translation.

At first, we extract infoboxes both from Wikipedia dump[2] (for Korean infoboxes) and DBpedia (for English infoboxes), and then use interlanguage links to determine which infoboxes are equivalent in both languages. Interlanguage links are links from an article in one Wikipedia language to the same subject in other Wikipedia languages [1]. For example, the article *Banana* in the English language Wikipedia may have interlanguage links to the article 바나나 in the Korean language Wikipedia, *Banane* in the German language Wikipedia, and so on.

After extracting infobox template in both languages, we retrieve the template-pairs between two languages. At the same time, we also extract schema of infoboxes. This schema reflects which properties is the most relevant properties for each infobox. There are some articles have more than one infobox template, we did not deal with this case in this paper.

There are several types of imbalances of infobox information between two different languages. According to the presence of infobox, we classified interlanguages linked pairs of articles into three groups: Short Infobox (S), Distant Infobox (D), and Missing Infobox (M):

[2]http://download.wikimedia.org/

- The ***S-group*** contains pairs of articles which use the same infobox template but have a different amount of information. For example, an English-written article and a non-English-written article, which have an interlanguage link and use the same infobox template, but have a different amount of template attributes.

- The ***D-group*** contains pairs of articles which use different infobox templates. D-group emerges due to the different degrees of each Wikipedia communities' activities. In communities where many editors lively participate, template categories and formats are more well-defined and more fine-grained. For example, *philosopher, politician, officeholder* and *military person* templates in English are matched just *person* template in Korean. It appears not only a Korean Wikipedia but also non-English Wikipedias.

- The ***M-group*** contains pairs of articles where an infobox exists on only one side.

The forms of each types are shown in Figure 2.

We also refined template schema that is necessary for several reasons. The first reason is that many templates use the abbreviation for its name, for example, `SW books`, `NRHP` and `MLB player`. In the cases of `SW books` and `NRHP`, these stand for `Star Wars books` and `National Register of Historic Places` respectively. However, a `SW` is widely used for an abbreviation of a `software` and `NRHP` is used only in Wikipedia. Thus it is difficult to understand.

The second reason is that the template name does not have same meaning as the common sense in a dictionary. For example, according to the Wiktionary[11] `television` means 'An electronic communication medium that allows the transmission of real-time visual images, and often sound'. However, for example, the `Template:Infobox television` represents television program. Its attributes are shown in below:

- **Infobox_television**={show_name, image, caption, genre, format, creator, developer, writer, director, producer, country, opentheme, ...}

Moreover different format of properties with same meaning should be refined. In order to overcome the problems of multiple property names being used for the same type, for example, 'birth place', 'birthplace and age' and 'place birth' are mapped to 'birthPlace'. In addition, the 'img' property is replaced into the 'image', because the later form is better to translate. If this is solved, it will be a great help for organizing templates and the efficiency of our synchronization framework will be higher.

We extracted English triples in infoboxes are simply translated into Korean triples. For the dictionary-based term translation [8], we use bilingual dictionary which is originally created for Englsih-to-Korean translations from Wikipedia interlanguage links [1], with our pre-constructed bilingual resources from SWRC [3]. The dictionary based translation is easy to set up. It just requires access to a bilingual dictionary. However, using multiple translation entries from a dictionary can generate large amounts of ambiguity which is a classic problem of translation. Another problem comes from the uncollected names. Such as proper names, uncollected names are still not correctly translated using the dictionary-based translation. After we constructed the translation resource, we added several translation patterns for multi-terms. Multi-terms are set of single terms such as *Computer Science*. We extracted 11 translation patterns(9 patterns for syntactic solving and 2 patterns for Date) using bilingual alignment of collocations (See Table 1). It uses the simple method of word co-occurrence in two or more aligned sentences across languages. We used the English-Korean multilingual corpus from SWRC. It consists of 60,000 pairs of sentences. We can choose from any of those aligned parallel sentences and determine the relative word order in the collocation [3].

[3]Semantic Web Research Center http://swrc.kaist.ac.kr

Table 1: Syntactic Translation Patterns between English and Korean multi-terms

English	Korean
A B	A B
A and B	A B
A or B	A B
A, B and C	A B C
the A	A
A of B	B A
A in B	B 에서 A
A from B	B 에서 A
A on B	B 의 A

After the translation step, we try to align the bilingual metadata. For the alignment, we manually constructed the mapping table between Korean and English template schemas based on translation results.

The details of this method are described in the following.

4. EXPERIMENTATION

We introduce the details of experimental dataset and the processing of the data.

4.1 Metadata Extraction

We need two types of data sets for experimentation. One type is the Wikipedia dump and the other type of data is the DBpedia dataset. In this paper, we only use Korean and English data, but our approach can be applied on any language of data in Wikipedia. We extracted 37,576 Korean articles contain infobox using Wikipedia dumps at March, 2010. Once the data is ready, all articles are parsed to extract attribute-value pairs of infobox. We got 1,042 infobox templates in Korean. We got 2,792 infobox templates in English using DBpedia dataset, and 1,642 template-pairs. It is noted lots of templates are duplicate uses. Thus, the number of template-pairs is much bigger than the number of templates in Korean. However, there are many articles having infobox without the name of template, it is a hurdle in the progress and development of the infobox extraction.

4.2 Metadata Translation

We executed the translation from English triples in infoboxes to Korean triples. In our experiments, we used DBpedia 3.4 as resource for the translation, a comparison of datasets is as follows:

- English Triples in DBpedia: 43,974,018

- Korean Dataset (Existing Triples/Translated Triples): 354,867/12,915,169

We can get translated Korean triples over 30 times larger than existing Korean triples. However, its quality is still quite poor. Because lots of triples include sentences or phrases such as below:

TRIPLE 1. *"%21%21%21", "currentMembers", "**Nic Offer Allan Wilson Mario Andreoni Tyler Pope Dan Gorman Sean McGahan Shannon Funchess Paul Quattrone**"*

TRIPLE 2. *"14226_Hamura", "discoverer", "**Lincoln Laboratory Near-Earth Asteroid Research Team**"*

Also, a large amount of translated triples consists of only numbers. Our translation approach is compared to a statistical machine translation system by Google Translate API. The Google Translate API is an initial prototype used a statistical MT system based on Moses and trained on Europarl. Overall, the Google Translate API performed better and executed faster than our dictionary based translation. However, in case of proper nouns and abbreviations, our approach yields slightly higher accuracy than Google Translate API.

5. DISCUSSIONS

Today, we created the metadata using translated results. However, we did not check the consistency between existing things and new things. To solve this problem, we try to utilize this consistency management to construct a large and fine-grained ontology by using infobox template. Thus we have built a template ontology, OntoCloud[4], from DBpedia and Wikipedia to efficiently build the template structure.

The construction of OntoCloud consists of the following steps: (1) extracting templates of DBpedia as concepts in an ontology, for example, the Template:Infobox *Person* (2) extracting attributes of these templates, for example, *name* of Person. These attributes are mapped to properties in ontology. (3) constructing the concept hierarchy by set inclusion of attributes, for example, *Book* is a subclass of *Book series*. For the ontology building, similar types of templates are mapped to one concept. For example, the *Template:infobox baseball player* and *Template:infobox asian baseball player* describe *baseball player*. Using the OntoCloud, we could be check the consistency of templates.

6. CONCLUSIONS

As the web grows in number of pages and amount of information, there is an increasing interest towards supporting tasks such as organizing, and enriching. We have proposed a novel idea on using term translation not only to synchronize but also to enrich information of multilingual web resources, and presented an effective approach to implement this idea in Wikipedia. Our work is ongoing for technical improvements, such as better alignment between bilingual metadata, and more precise translating. After the verification of template consistency, the Korean Wikipedia and DBpedia can be updated automatically. This will be helpful to guarantee that the same information can be recognized in different languages. Moreover, will be helpful to edit articles and to create infoboxes when a new article is created. It can support the authors by suggesting the right template. As future work, it is planned to support more standardization for the Korean language and improve the quality of the translated datasets.

[4]http://swrc.kaist.ac.kr/ontocloud/

7. REFERENCES

[1] E. Adar, M. Skinner, and D. S. Weld. Information arbitrage across multi-lingual wikipedia. In *WSDM '09: Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 94–103, New York, NY, USA, 2009. ACM.

[2] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. In *Proceedings of 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference (ISWC+ASWC 2007)*, pages 722–735. November 2008.

[3] R. Basili, D. De Cao, D. Croce, B. Coppola, and A. Moschitti. Cross-language frame semantics transfer in bilingual corpora. In *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, chapter 27, pages 332–345. 2009.

[4] G. Bouma, S. Duarte, and Z. Islam. Cross-lingual alignment and completion of wikipedia templates. In *CLIAWS3 '09: Proceedings of the Third International Workshop on Cross Lingual Information Access*, pages 21–29, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[5] S. Chernov, T. Iofciu, W. Nejdl, and X. Zhou. Extracting semantic relationships between wikipedia categories. In *In 1st International Workshop: "SemWiki2006 - From Wiki to Semantics" (SemWiki 2006), co-located with the ESWC2006 in Budva*, 2006.

[6] C. B. Eugenio Tacchini, Andreas Schultz. Experiments with wikipedia cross-language data fusion. In *5th Workshop on Scripting and Development for the 5th Workshop on Scripting and Development for the Semantic Web (SFSW2009)*, 2009.

[7] S. Hellmann, C. Stadler, J. Lehmann, and S. Auer. Dbpedia live extraction. In *Proc. of 8th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE)*, volume 5871 of *Lecture Notes in Computer Science*, pages 1209–1223, 2009.

[8] O. Levow. Dictionary-based techniques for cross-language information retrieval. *Information Processing and Management*, 41:523–547, 2005.

[9] O. Medelyan, D. Milne, C. Legg, and I. H. Witten. Mining meaning from wikipedia. May 2009.

[10] F. Wu and D. S. Weld. Automatically refining the wikipedia infobox ontology. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 635–644, New York, NY, USA, 2008. ACM.

[11] T. Zesch, C. Müller, and I. Gurevych. Extracting lexical semantic knowledge from wikipedia and wiktionary. In *Proc. of the 6th Conference on Language Resources and Evaluation (LREC)*, 2008.

Modeling Wordlists via Semantic Web Technologies

Shakthi Poornima
Department of Linguistics
State University of New York at Buffalo
Buffalo, NY USA
poornima@buffalo.edu

Jeff Good
Department of Linguistics
State University of New York at Buffalo
Buffalo, NY USA
jcgood@buffalo.edu

ABSTRACT

We describe an abstract model for the traditional linguistic wordlist and provide an instantiation of the model in RDF/XML intended to be usable both for linguistic research and machine applications.

Categories and Subject Descriptors

E.2 [**Data**]: Data Storage Representations

Keywords

wordlists, interoperation, RDF

1. INTRODUCTION

Lexical resources are of potential value to both traditional descriptive linguistics as well as computational linguistics.[1] However, the kinds of lexicons produced in the course of linguistic description are not typically easily exploitable in natural language processing applications, despite the fact that they cover a much larger portion of the world's languages than lexicons specifically designed for NLP applications. In fact, one particular descriptive linguistic product, a wordlist, can be found for around a third to a half of the world's seven thousand or so languages, though wordlists have not played a prominent role in NLP to the best of our knowledge.

Wordlists are widely employed by descriptive linguists as a first step towards the creation of a dictionary or as a means to quickly gather information about a language for the purposes of language comparison (especially in parts of the world where languages are poorly documented). Because of this, they exist for many more languages than do full lexicons. While the lexical information they contain is quite sparse, they are relatively consistent in their structure across resources. As we will see, this makes them good candidates for exploitation in the creation of a multilingual database consisting of rough translational equivalents which lacks precision, but has coverage well-beyond what would otherwise be available.

This paper describes an effort to convert around 2700 wordlists covering more than 1500 languages (some wordlists represent dialects) and close to 500,000 forms into an RDF format to make them more readily accessible in a Semantic Web context.[2] This may well represent the largest single collection of wordlists anywhere and certainly represents the largest collection in a standardized format. While the work described here was originally conceived to support descriptive and comparative linguistics, we will argue that the use of Semantic Web technologies has the additional beneficial effect of making these resources more readily usable in other domains, in particular certain NLP applications.

We approach this work as traditional, not computational linguists, and our current goal is to encode the available materials not with new information but rather to transfer the information they contain in a more exploitable format. Semantic Web technologies allow us to represent traditional linguistic data in a way we believe remains faithful to the original creator's conception and, at the same time, to produce a resource that can serve purposes for which it was not originally intended (e.g., simplistic kinds of translation). Our work, therefore, indicates that Semantic Web offers a promising approach for representing the work of descriptive linguists in ways of use to computational linguists.

2. MODELING A WORDLIST

We illustrate the basic structure of a wordlist in (1), which gives a typical presentation format. Here, the language being described is French, with English labels used to index general meanings.

(1)

MAN	*homme*
WOMAN	*femme*

The information encoded in a wordlists is quite sparse. In general, they give no indication of morphosyntactic features (e.g., part of speech), nor of fine-grained semantics. Meanings are most usually indexed simply by the use of labels drawn from languages of wider communication (e.g., English or Spanish), though the intent is not to translate between languages but, rather, to find the closest semantic

[1]Funding provided for the work described here has been provided by NSF grant BCS-0753321 in the context of a larger-scale project, Lexicon-Enhancement via the Gold Ontology, headed by researchers at the Institute for Language Information and Technology at Eastern Michigan University. More information can be found at `http://linguistlist.org/projects/lego.cfm`.

[2]These wordlists were collected by Timothy Usher and Paul Whitehouse in the context of traditional comparative linguistic research, and represent an enormous effort without which the work described here would not have been possible.

WWW2010, April 26-30, 2010, Raleigh, North Carolina.

match in the target language for what is presumed to be a general concept. The notional relationship between a meaning and a form in a wordlist is not one of defining (as is the case in a monolingual dictionary) or translating (as is the case of a bilingual dictionary), but rather something we term *counterpart* following [1]. This is not a particularly precise relation, but it is not intended to be. Specifying too much precision in the meaning-form relationship would make it difficult to collect wordlists rapidly, which is otherwise one of their most desirable features.

The concepts that one sees in traditional linguistic wordlists have often been informally standardized across languages and projects through the use of what we call here *concepticons*. Concepticons are curated sets of concepts, minimally indexed via words from one language of wider communication but, perhaps, also described more elaborately using multiple languages (e.g., English and Spanish) as well as illustrative example sentences. They may include concepts of such general provenance that counterparts would be expected to occur in almost all languages, such as TO EAT, or concepts only relevant to a certain geographical region or language family. For instance, Amazonian languages do not have words for MOSQUE, and Siberian languages do not have a term for TOUCAN [1, p.5-6].

To the extent that the same concepticon can be employed across wordlists, it can be understood as a kind of interlingua, though it is not usually conceptualized as such by descriptive linguists. The concepticon we are employing is based on three available concept lists. The most precise and recently published list is that of the Loanword Typology (LWT) project [1], which consists of around 1400 entries.

3. WORDLISTS AND SEMANTIC WEB

Each wordlist in our RDF datanet consists of two components: metadata and a set of entries. The metadata gives relevant identifying information for the wordlist e.g., a unique identifier, the ISO 639-3 code, the related Ethnologue language name, alternate language names, reference(s), the compilers of the wordlist, etc. The entries set consists of all entries in the wordlist. The structure of our entries is quite simple, consisting of a reference to an external concepticon entry in the concepticon employed by our project paired with a form in the target language using the counterpart relationship discussed above. Obviously, this structure could be elaborated. However, it is sufficient for this first stage of a work and, we believe, serves as an appropriate baseline for further specification.

In cases where there is more than one form attached to a concept, we create two concept-form mappings. For instance, the entry in (2) from a wordlist of North Asmat, a language spoken in Indonesia, associates the concept GRANDFATHER with two counterparts, whose relationship to each other has not been specified in our source.

(2) GRANDFATHER: *-ak*, *afak*

An RDF/XML fragment describing one of the two forms in (2) is given in Figure 1 for illustrative purposes. In addition to drawing on standard RDF constructs, we also draw on descriptive linguistic concepts from GOLD[3] (General Ontology for Linguistic Description), which is intended to be a sharable ontology for language documentation and description. The key data encoded by our RDF representation of wordlists is the counterpart mapping between a particular wordlist concepts (`lego:concept`) drawn from our concepticon and a form (`gold:formUnit`) found in a given wordlist.

```
<rdf:RDF xmlns:rdf="...>
    <lego:concept rdf:about="...">
    <lego:hasCounterpart>
       <gold:LinguisticSign rdf:about="...">
          <gold:inLanguage>
            <gold:Language rdf:about="..."/>
          </gold:inLanguage>
          <gold:hasForm>
            <gold:formUnit>
              <gold:stringRep>-ak</gold:stringRep>
            </gold:formUnit>
          </gold:hasForm>
       </gold:LinguisticSign>
    </lego:hasCounterpart>
    </lego:concept>
</rdf:RDF>
```

Figure 1: Wordlist Entry RDF Fragment

An important feature of our RDF model, illustrated in Figure 1 is that the counterpart relation does not relate a meaning directly to a form but rather to a linguistic sign (`gold:LinguisticSign`) whose form feature then contains the relevant specification. This structure would allow additional information (e.g., part of speech, definition, example) about the lexical element specified by the given form to be added to the representation at the level of the linguistic sign, if it were to become available.

4. PROSPECTS

The data model described here was originally designed to promote lexical data interoperability for descriptive linguistic purposes. At the same time, it makes visible the similarities between a concepticon and an interlingua, thus opening up the possibility of straightforward exploitation of a data type produced in a descriptive linguistic context in NLP contexts. Furthermore, by expressing the model in the form of an RDF graph rather than a more parochial XML format, it can be more easily processed. Potential NLP applications for this datanet involve tasks where simple word-to-word mapping across languages may be useful. One such example is the PanImages[4] search of the PanLex project which facilitates cross-lingual image searching. More work could be done to promote interoperability, of course. For example, we could devise an LMF [2] expression of our model, though we leave this for the future.

5. REFERENCES

[1] In M. Haspelmath and U. Tadmor (eds.), *Loanwords in the world's languages: A comparative handbook*. 2009.

[2] G. Francopoulo, et al. Multilingual resources for NLP in the lexical markup framework (LMF). *Language Resources and Evaluation*, 43:57–70, 2009.

[3] `http://linguistics-ontology.org/`. Similar ontologies such as SKOS could also be used in lieu of GOLD.

[4] `http://www.panimages.org/`

Multilingual Ontology-based User Profile Enrichment

Ernesto William De Luca, Till Plumbaum, Jérôme Kunegis, Sahin Albayrak
DAI Lab, Technische Universität Berlin
Ernst-Reuter-Platz 7, 10587 Berlin, Germany
{ernesto.deluca, till.plumbaum, jerome.kunegis, sahin.albayrak}@dai-labor.de

ABSTRACT

In this paper, we discuss the possibility of enriching user profiles with multilingual information. Nowadays, the English language is the de facto standard language of commerce and science, however users can speak and interact also in other languages. This brings up the need of enriching the user profiles with multilingual information. Therefore, we propose to combine ontology-based user modeling with the information included in the RDF/OWL EuroWordNet hierarchy. In this way, we can personalize retrieval results according to user preferences, filtering relevant information taking into account the multilingual background of the user.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

RDF/OWL, Web 2.0, Multilingualism, EuroWordNet

Keywords

Multilingual Semantic Web, User Modeling

1. INTRODUCTION

At present most of the demand for text retrieval is well satisfied by monolingual systems, because the English language is the de facto standard language of commerce and science. However, there is a wide variety of circumstances in which a reader might find multilingual retrieval techniques useful. Being able to read a document in a foreign language does not always imply that a person can formulate appropriate queries in that language as well. Furthermore, dealing with polysemic words seems to be more difficult in multilingual than in monolingual retrieval tasks.

Every text retrieval approach has two basic components: the first for representing texts (queries and documents) and the other for their comparison. This automated process is successful when its results are similar to those produced by human comparison between queries and documents. Queries and documents often differ from its length however. While the query is often quite short, documents might be up to hundreds of pages long. Moreover, users frequently adopt a vocabulary that is not contained in the documents, known as the paraphrase problem.

Multilingual Retrieval. When working in a multilingual environment, words have to be disambiguated both in the native and in the other languages. In this case the combination of multilingual text retrieval and word sense disambiguation (WSD) approaches is crucial [2]. In order to retrieve the same concept in different languages, some relations between the searched concept and its translations have to be built. WSD is used to convert relations between words into relations between concepts; sense disambiguation can be acquired for words, but it is more difficult for documents. To have accurate WSD, we need a larger coverage of semantic and linguistic knowledge than is available in current lexical resources.

Because we focus on multilingual concepts, we decided to use EuroWordNet [6], a variant of the most well-known available lexical database WordNet. In previous work, we extended the RDF/OWL WordNet representation [5] for multilingualism, leading to our own RDF/OWL EuroWordNet representation [3].

Ontology-based User Modeling. With the advent of the Web 2.0 and the growing impact of the Internet on our every day life, people start to use more and more different web applications. They manage their bookmarks in social bookmarking systems, communicate with friends on Facebook[1] and use services like Twitter[2] to express personal opinions and interests. Thereby, they generate and distribute personal and social information like interests, preferences and goals [4]. This distributed and heterogeneous corpus of user information, stored in the user model (UM) of each application, is a valuable source of knowledge for adaptive systems like information filtering services. These systems can utilize such knowledge for personalizing search results, recommend products or adapting the user interface to user preferences. Adaptive systems are highly needed, because the amount of information available on the Web is increasing constantly, requiring more and more effort to be adequately managed by the users. Therefore, these systems need more and more information about users interests, preferences, needs and goals and as precise as possible. However, this personal and social information stored in the distributed UMs usually exists in different languages due to the fact that we communicate with friends all over the world. Also, today's adaptive systems are usually part of web applications and typically only have access to the information stored in that specific ap-

WWW2010, April 26-30, 2010, Raleigh, North Carolina.

[1] http://www.facebook.com/
[2] http://twitter.com/

plication. Therefore, we enhance the user model aggregation process by adding valuable and important meta-data which leads to better user models and thus to better adaptive systems. For this reason, we propose a combination of RDF/OWL EuroWordNet within ontology-based aggregation techniques.

2. PROPOSED SEMANTIC USER MODELING AGGREGATION

RDF/OWL EuroWordNet opens new possibilities for overcoming the problem of language heterogeneity in different user models and thus allows a better user modeling aggregation. Therefore, we propose an ontology-based user modelling approach that combines mediator techniques to aggregate user models from different applications and utilize the EuroWordNet information to handle the multilingual information in the models. Based on this idea, we define some requirements that we have to fulfill.

Requirement 1: Ontology-based profile aggregation. We need an approach to aggregate information that is both application independent and application overarching. This requires a solution that allows us to semantically define relations and coherences between different attributes of different UMs. The linked attributes must be easily accessible by applications such as recommender and information filtering systems. In addition, similarity must be expressed in these defined relations.

Requirement 2: Integrating semantic knowledge. A solution to handle the multilingual information for enriching user profiles is needed. Hence, we introduce a method to incorporate information from semantic data sources such as EuroWordNet and to aggregate complete profile information. We decided to use an ontology as the conceptual basis of our approach to meet the first requirement explained above. Therefore a meta-ontology is used to link attributes of different UMs that contain equal or similar content.

The definition of a meta-model based on the meta-ontology can be divided into two steps. First, we define a concrete meta-model for a specific domain we want to work with, such as music, movies or personal information. The meta-model can be an already existing model, like FOAF[3] or a proprietary model that only certain applications understand. Next, we decribe how to connect multilingual attribute information stored in different user models.

3. MULTILINGUAL ONTOLOGY-BASED AGGREGATION

To enrich the user model with multilingual information, as described above, we decided to utilize the knowledge available in RDF/OWL EuroWordNet [3]. We want to leverage this information and use it for a more precise and qualitatively better user modeling. We treat the semantic external resources as a huge semantic profile that can be used to enrich the user model and add valuable extra information (see Figure 1). The aggregation of information into semantic profiles and user models is performed similarly to the approach described in [1], by using components that mediate between the different models. We extend this approach by using a combined user model, aggregated with the proposed ontology.

[3]http://www.foaf-project.org/

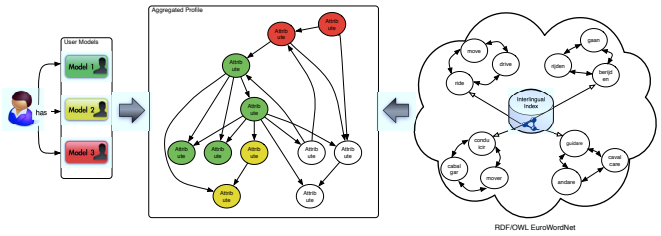

Figure 1: Integrating semantic knowledge about multilingual dependencies with the information stored in the user models.

To use the information contained in RDF/OWL EuroWordNet, we developed a framework that allows us to define several mediators that take the information from user models and trigger different sources in the Semantic Web for more information. These mediators are specialized components that read a user model and collect additional data from an external source.

4. CONCLUSION

In this paper, we presented the possibility of enriching user profiles with information included in the RDF/OWL EuroWordNet hierarchy to better filter results during the search process. This aggregated information can be used in our multilingual semantic information retrieval system that has been described in more details in [2]. In this work, we have shown that we can handle the high heterogeneity of distributed data, especially concerning multilingual heterogeneity, using aggregated user profiles that have been enriched with information contained in the RDF/OWL EuroWordNet representation. This gives us the possibility to personalize retrieval results according to user preferences, filtering relevant information taking into account the multilingual background of the user.

5. REFERENCES

[1] Shlomo Berkovsky, Tsvi Kuflik, and Francesco Ricci. Mediation of user models for enhanced personalization in recommender systems. *User Modeling and User-Adapted Interaction*, 18(3):245–286, 2008.

[2] Ernesto William De Luca. *Semantic Support in Multilingual Text Retrieval*. Shaker Verlag, Aachen, Germany, 2008.

[3] Ernesto William De Luca, Martin Eul, and Andreas Nürnberger. Converting EuroWordNet in OWL and extending it with domain ontologies. In *Proc. Workshop on Lexical-semantic and Ontological Resources*, 2007.

[4] Till Plumbaum, Tino Stelter, and Alexander Korth. Semantic web usage mining: Using semantics to understand user intentions. In *Proc. Conf. on User Modeling, Adaptation and Personalization*, pages 391–396, 2009.

[5] Mark van Assem, Aldo Gangemi, and Guus Schreiber. WordNet in RDFS and OWL. Technical report, W3C, 2004.

[6] Piek Vossen. Eurowordnet general document, version 3, final, 1999.