MSW 2011

Proceedings of the 2nd International Workshop on the

Multilingual Semantic Web

Collocated with the 10th International Semantic Web Conference (ISWC 2011)

Bonn, Germany, October 23, 2011.

Sponsored by:

Endorsed by:

ACL SIGSEM

Preface

Multilingualism has become an issue of major interest for the Semantic Web community, in light of the substantial growth of internet users that create and update knowledge all over the world in languages other than English. This process has been accelerated due to initiatives such as the Linked Data initiative, which encourages not only governments and public institutes to make their data available to the public, but also private organizations in domains as far apart as medicine, cartography or music. These actors publish their data sources in the languages they are available in, and, as such, in order to make this information available to an international community, multilingual knowledge representation, access and translation are an impending need.

This second edition of the MSW workshop focused on the representation of multilingual information in the Semantic Web and Linked Data, specifically addressing issues in the cross-lingual discovery of mappings between multilingual Linked Data vocabularies and data sets, and the cross-lingual querying of knowledge repositories. The 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 13 submissions to the workshop, from which the program committee accepted 5 as full papers and 5 as short papers. Taking into account only the full papers the selection rate amounts to 40%. The accepted papers cover a variety of topics regarding the representation of lexical objects in the Semantic Web, the creation and management of multilingual knowledge bases, as well as the cross-lingual linking of multilingual ontologies and data sets. The MSW Workshop program also included a keynote talk by Sebastian Hellmann.

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 for its support through the research grant for Monnet (FP7-248458), the Spanish Ministry of Science and Innovation for its support through the BabeLData project (TIN2010-17550), and the Science Foundation Ireland through Lion2 (SFI/08/CE/I1380). Special thanks also go to the European Project FlareNet (ECP-2007-LANG-617001) and the Special Interest Group on Computational Semantics (SIGSEM) of the Association for Computational Linguistics (ACL) for their endorsement.

Elena Montiel-Ponsoda
John McCrae
Paul Buitelaar
Philipp Cimiano

October, 2011

Table of Contents

MSW 2011 Organization

Organizing Committee

Elena Montiel-Ponsoda
Ontology Engineering Group (OEG), Departamento de Inteligencia Artificial
Universidad Politécnica de Madrid, España
http://www.oeg-upm.net/index.php/en/phd/52-emontiel

John McCrae
Semantic Computing Group, CITEC – University of Bielefeld
http://www.sc.cit-ec.uni-bielefeld.de/people/jmccrae

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

Program Committee

Ernesto William De Luca, Universitaet Magdeburg - Data and Knowledge Engineering Group, Germany

Vanesa López, KMI, Open University, UK

Gerard de Melo, Microsoft Research Cambridge, UK

Sergei Nirenburg, University of Maryland - Institute for Language and Information Technologies, USA

Alessandro Oltramari, Carnegie Mellon University, Department of Psychology, Pittsburgh, USA

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

Felix Sasaki, University of Applied Sciences Potsdam, Germany

Marta Sabou, Department of New Media Technology – MODUL University, Vienna

Philipp Sorg, AIFB – Karlsruhe Institute of Technology, Germany

Martin Volk, Universitaet Zürich - Institute of Computational Linguistics, Switzerland

Piek Vossen, Vrije Universiteit - Dept. of Language, Cognition and Communication, the Netherlands

Yong Yu, Computer Science and Engineering, Shanghai Jiao Tong University

Cross-Lingual Web API Classification and Annotation

Maria Maleshkova, Lukas Zilka, Petr Knoth, Carlos Pedrinaci

Knowledge Media Institute (KMi)
The Open University, Milton Keynes, United Kingdom
{m.maleshkova, l.zilka, p.knoth, c.pedrinaci}@open.ac.uk

Abstract. Recent developments on the Web are marked by the growing support for the Linked Data initiative, which encourages government and public organisations, as well as private institutions, to expose their data on the Web. This results in a plentitude of multi-lingual document collections where the original resources are published in the language, in which they are available. The challenges of multilingualism present on the Semantic Web are also reflected in the context of services on the Web, characterised by the rapid increase in popularity and use of Web APIs, as indicated by the growing number of available APIs and the applications built on top of them. Web APIs are commonly described in plain-text as part of Web pages, following no particular guidelines and conforming to no standards, despite some initial approaches in the area [1, 2]. Therefore, API providers publish descriptions in any language they see fit, making the service discovery and the subsequent processing of the documentation challenging tasks. In this paper, we present a cross-lingual approach that calculates semantic similarity of text to help classify and annotate Web APIs, based on their textual descriptions. Furthermore, we show how our solution can be implemented as part of SWEET [3], which is a tool that enables the semi-automated creation of semantic Web API descriptions. In addition, we demonstrate how the cross-lingual approach can be adopted to support the language-independent discovery of Web APIs.

1 Introduction

In the research context, English has established itself as a de-facto standard language for conducting and publishing work. It is therefore easy to forget that multilingualism is actually one of the main characteristics of the Semantic Web. The importance of language diversity is made evident by the growing support for the Linked Data initiative, which encourages government and public organisations, as well as private institutions, to expose their data on the Web. Since the document collections are published in the language, in which the original sources are available, the result is an abundance of multi-lingual resources. In comparison, the situation is quite similar in the context of services on the Web, where the past few years have been marked by the increasing popularity and use of Web APIs. The growing importance of Web APIs, also referred to as RESTful

services [4] (especially when conforming to the REST [5] architectural principles) was initially triggered by popular Web 2.0 applications like Facebook, Google, Flickr and Twitter that offer easy-to-use, publicly available APIs as means for accessing their resources. Currently, Web APIs not only enable retrieval and manipulation of different resources but also facilitate building of versatile applications based on combining heterogeneous data coming from diverse services.

Despite their proliferation, Web APIs are facing a number of limitations. The majority of the Web APIs have only textual descriptions that are given directly as part of HTML Web pages, disregarding efforts towards a common formal language for describing Web APIs [1, 2]. Providers publish the documentation in any form and any language that they see fit and as a result, finding and using Web APIs can be quite challenging and requires extensive manual effort. API consumers need to search for suitable services, manually process and interpret the available documentation, which is sometimes in a different language, and produce custom implementation solutions that are rarely reusable.

In this paper, we focus in particular on supporting the Web API search and discovery tasks by enhancing the descriptions with: 1) information about the type of provided functionality (for example, a weather service or a shopping service) and 2) central concepts that can be used for determining the domain of the service or be taken directly for annotating service properties such as the inputs and outputs. For this purpose we present an approach that makes use of Cross-lingual Explicit Semantic Analysis [6] to classify and annotate APIs, given their textual description. As a result we are able to discover APIs with a particular functionality, or characterised by a set of keywords, across languages. Moreover, by including the computed classification and annotation details as part of the semantic Web API descriptions, we support service discovery as well as directly contribute to a Semantic Web that integrates Web APIs with multilingual documentation. We also validate the applicability of the devised approach by introducing a design of a system capable of supporting the creation of semantic Web API descriptions, enhanced with classification information and further annotations, and describe the implementation of its key components.

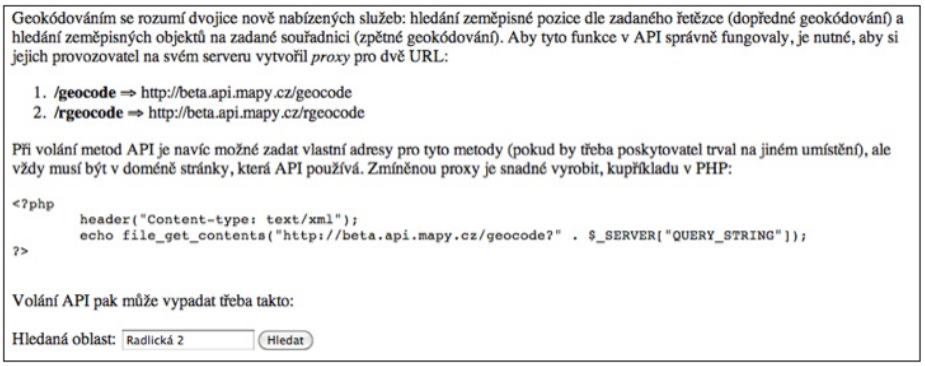
The remainder of this paper is structured as follows: Section 2 provides a motivating example that illustrates the challenges of searching for APIs with particular functionality or from a particular domain, while Section 3 lists related work and gives some background in the area of semantic Web API descriptions and details on the cross-lingual semantic relatedness approach. Our API classification and annotation approaches are given in Section 4. Section 5 describes in more detail the solution design and the implementation of the key components, and Section 6 concludes the paper.

2 Motivation

One of the most common service discovery tasks is discovery based on the functionality or the domain of the service (for example, “I am looking for an API that can map my travel route” or “I am looking for a shopping service”). Therefore,

in this paper we focus our work on supporting this basic but essential discovery type. Currently the search options for APIs are very limited. One possibility is to use conventional search engines such as Google or Yahoo and do keyword search and hope that one of the returned matches is a Web API description. It is important to point out that so far there is no way of automatically distinguishing between webpages that describe Web APIs and webpages that simply mention an API, such as a news article, so this differentiation has to be done manually. Another way is searching in Web API directories, such as ProgrammableWeb (`http://www.programmableweb.com`), which are based on manually collecting and registering APIs. A final option is looking in developer forums and asking other users for suitable APIs, i.e. the "word of mouth" approach.

Figure 1 visualises a simple example, which demonstrates the necessity of supporting cross-language Web API search. The presented API provides capabilities for geocoding and reverse geocoding. If we use Google to search for a geocoding API, the query will be language specific; therefore, we would either find a service such as the popular GeoNames (`http://www.geonames.org/export/web-services.html`), which is in English, or the example description in Czech (`http://ondras.zarovi.cz/smap/geokodovani/`). However, it would not be possible to find both descriptions with one and the same search keywords. Similarly, existing Web API directories are language specific, in particular restricted to English, as are developer sites and forums too.



Geokódováním se rozumí dvojice nově nabízených služeb: hledání zeměpisné pozice dle zadaného řetězce (dopředné geokódování) a hledání zeměpisných objektů na zadané souřadnici (zpětné geokódování). Aby tyto funkce v API správně fungovaly, je nutné, aby si jejich provozovatel na svém serveru vytvořil *proxy* pro dvě URL:

1. **/geocode** ⇒ http://beta.api.mapy.cz/geocode
2. **/rgeocode** ⇒ http://beta.api.mapy.cz/rgeocode

Při volání metod API je navíc možné zadat vlastní adresy pro tyto metody (pokud by třeba poskytovatel trval na jiném umístění), ale vždy musí být v doméně stránky, která API používá. Zmíněnou proxy je snadné vyrobit, kupříkladu v PHP:

```
<?php
        header("Content-type: text/xml");
        echo file_get_contents("http://beta.api.mapy.cz/geocode?" . $_SERVER["QUERY_STRING"]);
?>
```

Volání API pak může vypadat třeba takto:

Hledaná oblast: Radlická 2 Hledat

Fig. 1. Example Web API Description in Czech

In summary, even though there are at least two geocoding Web APIs, given the existing search possibilities, we would find either one or the other, depending on which language we use to conduct the search. Therefore, we propose to employ a cross-language classification approach and to enhance the API descriptions with metadata about the service functionality. Furthermore, we propose to determine the key concepts, characterising the textual documentation, and to use those directly as tags or even use them to determine the domain of the service and specific annotations for individual service properties, such as inputs and outputs. In particular, we follow a lightweight semantic approach for enhancing existing API description with metadata, which supports the completion of tasks such as discovery, but also composition and invocation, on the level of semantics,

abstracting away from syntactic specifics, including the original language of the documentation [3, 7]. We provide more detail to the proposed approach in the following sections.

3 Background and Related Work

In this section we provide some background on the use of lightweight semantics for describing Web APIs, list existing annotation and tagging tools, and focus on providing details on common classification approaches and, in particular, on classification based on cross-lingual semantic relatedness.

3.1 Lightweight Semantic Web API Descriptions

Since the advent of Web service technologies, research on semantic Web services (SWS) has been devoted to reduce the extensive manual effort required for manipulating Web services. The main idea behind this research is that tasks such as discovery, negotiation, composition and invocation can have a higher level of automation, when services are enhanced with semantic descriptions of their properties. Similarly to "classical" Web services based on WSDL/SOAP, Web API-related tasks also require a lot of developer involvement and face even further difficulties, since there is no established common formalism for describing Web APIs. In order to address this, lightweight annotations over API descriptions have been proposed as means for achieving a higher-level of automation.

Currently, there are two main contributions aiming at using semantics to support the automation of common Web API service-related tasks. Both approaches rely on marking service properties within the HTML description and subsequently linking these to semantic entities. MicroWSMO [7] is a formalism for the semantic description of Web APIs, which is based on adapting the SAWSDL [8] approach for enhancing service properties with semantic information. MicroWSMO uses microformats for adding semantic information on top of HTML service documentation, by relying on hRESTS [9] for marking service properties. Another formalism is SA-REST [10], which also applies the grounding principles of SAWSDL but instead of using hRESTS relies on RDFa [11] for marking service properties. Similarly to MicroWSMO, SA-REST enables the annotation of existing HTML service descriptions by identifying service elements and linking these to semantic entities. The main differences between the two approaches are not the underlying principles but rather the implementation techniques. For the here presented work, we have adopted hRESTS and MicroWSMO that are already implemented as part of SWEET [3], which is a tool that enables the semi-automated creation of semantic Web API descriptions.

Currently, there are quite a few tagging tools that enable the tagging of web pages but also support the user in choosing the correct tags. Some of the main ones include TagAssist [12], collaborative tagging [13] and user-based collaborative tagging [14]. In the context of our work, there are also a number of application that are especially developed for supporting Web service and API annotation [15, 16]. However, since we propose a general approach for classifying

APIs and determining further annotations, any of the existing tagging or service description tools can be extended to include the computed results and present them to the user. In this paper, we verify the applicability of our approach by enhancing SWEET through integration with the developed cross-language classification and central concepts deriving components.

3.2 Cross-lingual Text Classification

Text classification has been successfully applied to many real world problems including spam detection, plagiarism detection or newspaper content classification, and its importance grew quickly with the amount of information available on the Web. Along with the widespread use of text classification methods comes the need for automated classification of new documents or web pages into hierarchies. This can be demonstrated on the Web by the existence of large web directories, such as Open Directory Project or ProgrammableWeb.

Over the past 20 years, text classification largely benefitted from the advances in the field of machine learning [17]. The machine learning approach, which aims at inducing a classifier given a set of training examples, already dominates over the knowledge engineering approach, which consisted of manually constructing the classifier. A common way to address the problem is to represent a textual document using a Vector Space Model [18], i.e. as a weighted vector of terms, and to automatically build a classifier from a set of training examples. While this approach often produces good results when applied to monolingual texts, it is not directly applicable in a multilingual environment.

There are two common approaches to address this problem:

- *Machine translation approach* - involves machine translation of texts to a common language or interlingua and then represents the documents as vectors in that language.
- *Mapping to a shared conceptual space* - represents the documents as term vectors in their source language and then projects them into a shared conceptual space. This is typically done in practice with the help of ontologies/vocabularies or by applying the distributional hypothesis [19].

An approach, which received much attention in the recent, years is to use Wikipedia terms as a shared conceptual space. Texts can be mapped into this space by performing Explicit Semantic Analysis (ESA) [20], hence this method is called Cross-language Explicit Semantic Analysis (CL-ESA) [6]. While there has been significant research involvement in monolingual text classification, the multilingual context has been addressed only recently. The Cross-Language Evaluation Forum (CLEF) has been, over the last decade, the main conference specialising in this research field.

In this paper we describe a Web API classification and annotation method that uses CL-ESA to classify the textual description of a Web API, given a background collection of APIs. The form of CL-ESA that we utilise is equivalent to [6], and lies in finding the correct cross-lingual mapping of the ESA concepts from the Wikipedia. Since CL-ESA uses Wikipedia concepts to represent documents in a multilingual shared vector space, the approach is applicable to the majority of languages.

4 Supporting the Cross-lingual Web API Classification and Annotation

In this section we describe in detail our approach for classifying Web APIs based solely on their textual documentation. We provide the devised algorithm as well as a specific application example. We take the cross-lingual processing one step further and use it to determine the key concepts of the description, which can be used directly as tags or can serve as the basis for deriving further API annotations.

4.1 Cross-lingual Web API Classification

Our approach towards Web API classification is based on comparing the description of an API, which is to be classified, with a set of APIs already classified according to a given taxonomy. The specific implementation of our approach is based on the ProgrammableWeb taxonomy, which comprises of 54 classes (`http://www.programmableweb.com/apis/directory`). We refer to the set of pre-classified services as *Background Collection*. In particular, we determine a number of representative service descriptions for each class in the taxonomy. These service descriptions are used as service models for the classification process. Moreover, the actual classification process is not based on the textual descriptions in the background collection but rather on the pre-computed ESA vector representations, thus saving computation time at runtime.

In addition to the background collection, we also define a set of *stop words*. Web API documentation use very limited vocabulary for describing the format of data and also for describing the behaviour of the Web API. For this reason, a stop-word file must be built to prevent the Explicit Semantic Analysis from focusing on the features of Web API descriptions that do not differentiate the services into classes. Therefore, a sufficiently large document collection in each of the input languages must be acquired and used to build the stop-word list. The stop-word list serves as an input for the pre-processing step of the Explicit Semantic Analysis.

Algorithm 1 formally describes the proposed API classification approach. In particular, the devised method includes the following steps. First we determine the language, in which the Web API description is written. This is currently not an issue and can be done easily by comparing the word distribution of the Web API description to average word distributions of other languages, or using one of the Web Services[1]. Second, we remove the web-service specific stopwords and project the Web API description into the concept space given by the particular language version of Wikipedia. After that we project the vector into the English Wikipedia concept space, to facilitate its comparison with our Web API background. In the following step we iterate over each document in the background and record its similarity with the previously determined vector

[1] `http://code.google.com/apis/language/translate/v1/using_rest_langdetect.html`

of the input Web API description. Finally, for each category, we add up the acquired similarity measure and divide it by the number of examples for the given category. We do this in order to derive a normalised similarity measure, which is not influenced by the number of representative services. There are a number of further ways for determining the similarity measure (selecting the category with best service score, selecting the category with best median, etc.). The output is a list of categories, sorted according to their score.

Algorithm 1 Assigning Class Labels to a Web API Description

```
Require: webAPIDescription, backgroundCollection
Ensure: Scored class suggestions
  language ← recognize_language(webAPIDescription);
  esa_vector ← esa_analyze(language, webAPIDescription);
  esa_vector_en ← esa_map_vector(esa_vector, language, "en");
  category_score ← new Map();
  category_cnt ← new Map();
  for ⟨background_api_vector, category⟩ ∈ backgroundCollection do
    doc_score ← vector_similarity(esa_vector_en, background_api_vector);
    category_score[category] ← category_score[category] + doc_score;
    category_cnt[category] ← category_cnt[category] + 1;
  end for
  for category, score ∈ category_score do
    result[category] ← score / category_cnt[category];
  end for
  sort(result);
  return result
```

Coming back to the example introduced in Section 2, independently of whether we want to classify the GeoNames API or the Czech geocoding API, both descriptions will be converted to English ESA vectors. Based on each vector a list (ideally, an identical list) of sorted categories will be produced. Therefore, independently of the language, both descriptions would in the end be mapped to the same category. We do not consider the case where a new category needs to be created but simply map the API to the closest of the existing categories. Previous approaches base classification on word matches or word stemming/similarity, therefore, they are not applicable to a multi-lingual context.

4.2 Cross-Lingual Web API Annotation

Central Concepts Detection We assume that two APIs can be described with the same central concepts if their descriptions are semantically similar (their semantic relatedness measure is above some threshold). Our approach towards detecting the Central Concepts of a non-english Web API description is to find similar descriptions in a repository of English-based APIs (in this approach serving as background collection), and re-use its central concepts.

The Central Concepts for API descriptions in the repository. i.e. background collection, can be assigned either manually (e.g. by letting users assign keywords

to services), using a concept extraction method or a concept extraction Web service. We will use the concept extraction Web service AlchemyAPI[2]. It would be possible to extract concepts from the non-English WebAPI description directly using the aforementioned Web service, but from our experience the concept detection from an English text yields much better results.

Algorithm 2 Determining the Central Concepts for a Web API Description

```
Require: webAPIDescription, backgroundCollection
  language ← recognize_language(webAPIDescription);
  esa_vector ← esa_analyze(language, webAPIDescription);
  esa_vector_en ← esa_map_vector(esa_vector, language, "en");
  for ⟨background_api_vector, central_concepts⟩ ∈ backgroundCollection do
    score ← cosine_similarity(esa_vector_en, background_api_vector);
    results[score] ← central_concepts;
  end for
  return max(results)
```

Algorithm 2 represents the pseudo-code for our central concepts detection method. First, the language of the input API description is determined, and the description is projected into the ESA concept space of the particular language. Then, the ESA vector is mapped into the English concept space to facilitate its comparison with the ESA vectors of the services in the background API collection. The best matching service from the background collection is chosen and its central concepts are suggested as central concepts for the input API description.

If we use the algorithm to process the examples introduced in Section 2, the central concepts for the GeoNames API can be determined directly by using the AlchemyAPI. However, calculation of the central concepts for the Czech geocoding API is more challenging and is based on computing the cross-lingual similarity between its description and the descriptions in the background collection. The results, however, are comparable for both APIs.

The benefits of determining the central concepts for an API description are multifold. First, they can be used directly as tags for the Web API. These tags can be employed to enhance search within directories or as complementary information presented to the user as part of the API description. However, with some further processing, the central concepts can serve as the basis for determining semantic annotations for separate service parts, such as inputs and outputs, or for extrapolating the domain of the service. In particular, we propose to input the computed words into Watson [21] or Sindice (`http://sindice.com`) and to use the results as suggestions for semantic entities suitable for annotating the API. In our example, two of the central concepts are "latitude" and "longitude", which when posted in Watson return `http://www.w3.org/2003/01/geo/wgs84_pos#long` and `http://www.w3.org/2003/01/geo/wgs84_pos#lat`. These properties can directly be used to semantically describe the inputs of the API. Furthermore, the central concepts can be processed in order to determine

[2] `http://www.alchemyapi.com/`

the domain of the service and extrapolate a set of relevant domain ontologies. However, this work is beyond the scope of the paper but is envisioned as part of our future work.

4.3 Supporting Web API Search and Discovery

The here described methods for cross-lingual classification and determining central concepts can be employed for supporting Web API search and discovery, overcoming language boundaries. In particular, the benefits of enhancing Web API descriptions with classification information and specific key words can be implemented both directly on the level of the API documentation as well as on the semantic level. For instance, existing Web API directories could be extended with search functionality about the type of service or based on keywords describing the service, which in contrast to current solutions, would be language independent. This is a simple, yet effective way for enabling cross-language Web API search.

Furthermore, our work supports enhanced discovery by following the general approach outlined by semantic Web service technologies that aims to reduce the extensive manual effort required for performing tasks such as discovery, negotiation, composition and invocation by enriching services with semantic descriptions of their properties. In particular, the computed classification type and annotations can directly be included as part of lightweight semantic Web API descriptions given in MicroWSMO or SA-REST. These, in turn serve as a basis for applying automated discovery approaches. In the following section we describe the implementation of a system that enables precisely the semi-automatic creation of semantic API descriptions in MicroWSMO, where the user is presented with a list of suitable categories and annotations to choose from.

5 System Design and Implementation

In this section we validate our approach by presenting a system design and giving an implementation solution realised by extending the Semantic Web sErvice Editing Tool – SWEET [3]. SWEET[3] is a Web application developed using JavaScript and ExtGWT, which is started in a Web browser by calling the host URL. It takes as input an HTML Web page describing a Web API and offers functionalities, which enable users to annotate the service properties and to associate semantic information with them. As it can be seen in Figure 2, the architecture of SWEET consists of three main components, including the visualisation component, the data preprocessing component and the annotations recommender. In order to integrate the here presented work, we have extended the interface of the *Annotations Recommender*, to receive input from the *Cross-lingual Classification* and *Central Concept Detection* components.

The implementation of our cross-lingual Web API classification and annotation approach consists of three parts. The first one is the background builder,

[3] `http://sweet.kmi.open.ac.uk/`

which prepares the background collection for further classification, the second one proceeds with the actual classification, and the third one detects the central concepts. As background for the Explicit Semantic Analysis, we use different language versions of Wikipedia, in particular, English and Czech. The text analysis and its projection into ESA concepts space is done by our Java library, created by adapting the code from Wikiprep ESA implementation[4].

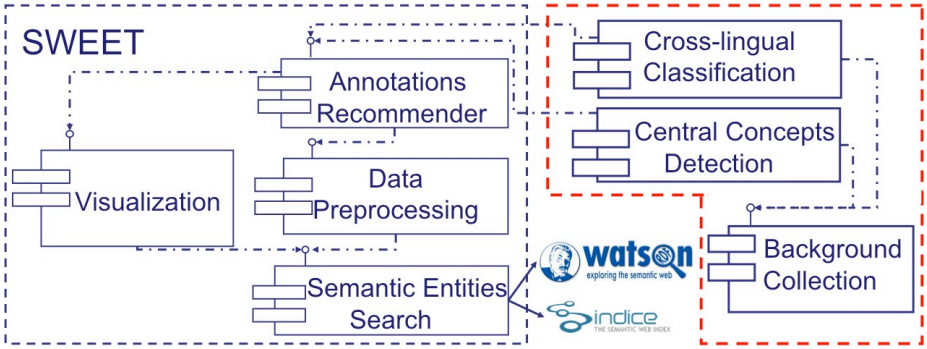


Fig. 2. SWEET Extended Architecture

The Web API background collection is built by getting APIs and categories from `http://www.programmableweb.com`. Five APIs are taken as an example for each category. Information about each API is saved to a database and after that the web pages describing each API are harvested. Subsequently, the HTML mark-up is removed and the text is normalised by removing stop-words and stemming. Then, the ESA vector is computed and stored in the database. Additionally, central concepts for each API in the background collection can be automatically determined by the AlchemyAPI. Before putting the Web API description into the AlchemyAPI engine, we remove the service-specific stop-words to get the Web API specific concepts.

Both, classification and central concept detection operate similarly, and differ only in the last step. They start with projecting the input API description into the Czech Wikipedia concept space. Then, the resulting Czech ESA vector is mapped into English ESA vector, using the concept mapping from Wikipedia. Afterwords, the ESA vector is compared with each API description ESA vector from the API background collection. The last step is the following:

– In case of classification, the results are aggregated and the best categories are suggested as candidates.

– Concept detection does not summarise the results but rather suggests the central concepts of the first few most semantically similar Web APIs as concept suggestions.

The so computed results can be represented to the user as annotation suggestions, aiding the process of creating the semantic Web API description. In the case of the classification of the service functionality, the top 3 results, for ex-

[4] `http://github.com/faraday/wikiprep-esa`

ample, can be automatically assigned to the API and the annotator would only need to validate them.

We also ran some preliminary evaluation and tests. In particular, we ran the concept detection system on the APIs from the geocoding domain. The first phase, which identifies the most similar service worked quite well, and was able to locate relevant similar Web APIs. Therefore the classification task was completed successfully. This evaluation needs to be extended to cover further domains, in order to be able to make statements about the precision of the classification approach in general. Our previous experiments with CL-ESA reported in [22] suggest that the method is able to detect semantically comparable text across languages with high precision (about 0.7 precision at top_{50}) from a 3.5 million large corpus. Given the fact that the size of ProgrammableWeb is smaller and we are classifying only into 54 classes, significantly better results can be expected.

In contrast, the concept extraction phase must be further refined because the returned central concepts were not always relevant. We discovered that the results greatly depend on the quality of the background collection. In particular, we are using the Web APIs from the ProgrammableWeb directory, where Web APIs are sometimes assigned to the wrong category or the link to the API documentation is inaccurate. We can overcome these limitations by hand-picking the APIs per category or by ensuring that the URLs pointing to the API documentation are correct. Even if improvements still remain to be done, the initial results show that the approach, especially in the context of the classification task, is quite promising.

6 Conclusions and Future Work

Nowadays, finding, interpreting and invoking Web APIs requires extensive human involvement due to the fact that the majority of the APIs have only textual documentation, not conforming to any particular standards and guidelines. Moreover, providers publish API description in any language that they see fit, making the discovery of suitable services a challenging tasks. In this paper, we present a cross-lingual approach, based on calculating semantic similarity, for classifying APIs and determining the central concepts of their descriptions, thus enabling language-independent search and discovery. We validate the applicability of the proposed method by implementing it as part of an extension to SWEET [3], which support users in creating semantic Web API descriptions. We also give some preliminary test results. Future work will mainly focus on extensively evaluating the system, staring off with improving the quality of the background collection and covering further domains in addition to the geocoding/mapping one.

Acknowledgment The work presented in this paper is partially supported by funding from the EC FP7, under grant agreement number 270001 – Decipher

References

1. M. J. Hadley: Web Application Description Language (WADL). Technical report, Sun Microsystems, November 2006. Available at `https://wadl.dev.java.net`.

2. Web Services Description Language (WSDL) Version 2.0. Recommendation, W3C, June 2007. Available at http://www.w3.org/TR/wsdl20/.
3. M. Maleshkova, C. Pedrinaci, J. Domingue: Semantic annotation of Web APIs with SWEET. 6th Workshop on Scripting and Development for the Semantic Web at ESWC, 2010.
4. L. Richardson, S. Ruby: RESTful Web Services. O'Reilly Media, May 2007.
5. R. T. Fielding: Architectural styles and the design of network-based software architectures. PhD thesis, University of California, 2000.
6. P. Sorg, P. Cimiano: Cross-lingual information retrieval with explicit semantic analysis. In Working Notes for the CLEF Workshop, 2008.
7. J. Kopecký, T. Vitvar, D. Fensel, K. Gomadam: hRESTS & MicroWSMO. Technical report, available at http://cms-wg.sti2.org/TR/d12/, 2009.
8. J. Kopecký, T. Vitvar, C. Bournez, J. Farrel. SAWSDL: Semantic Annotations for WSDL and XML Schema. IEEE Internet Computing, 11(6):60-67, 2007.
9. J. Kopecký , K. Gomadam, T.Vitvar: hRESTS: an HTML Microformat for Describing RESTful Web Services. In Proc of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence (WI-08), 2008.
10. A. P. Sheth, K. Gomadam, J. Lathem: SA-REST: Semantically Interoperable and Easier-to-Use Services and Mashups. In IEEE Internet Computing, 2007.
11. RDFa in XHTML: Syntax and Processing. Proposed Recommendation, W3C, September 2008. Available at http://www.w3.org/TR/rdfa-syntax/.
12. S. C. Sood, K. J. Hammond. TagAssist: Automatic tag suggestion for blog posts. In Proc of International Conference on Weblogs and Social, 2007.
13. S. Lee, A. Chun: Automatic tag recommendation for the web 2.0 blogosphere using collaborative tagging and hybrid ANN semantic structures. 6th Conference on WSEAS International Conference on Applied Computer Science, 2007.
14. R. Jaeschke, R. Marinho, A. Hotho, L. Schmidt-Thieme, G. Stumme: Tag recommendations in folksonomies. In PKDD, pages 506514, Springer, 2007.
15. A. Hess, E. Johnston, N. Kushmerick: ASSAM: A tool for semiautomatically annotating semantic web services. In Proc of the 3rd International Semantic Web Conference (ISWC), 2004.
16. A. Patil, S. Oundhakar, A. Sheth, K. Verma: METEOR-S web service annotation framework. pages 553562. ACM Press, 2004.
17. F. Sebastiani: Machine learning in automated text categorization. ACM Computing Surveys, 34(1):1-47, 2002.
18. C. D. Manning, P. Raghavan, H. Schutze: Introduction to Information Retrieval. Cambridge Press, 2008.
19. C. D. Manning, H. Schutze: Foundations of Statistical Natural Language Processing. The MIT Press, 1999
20. E. Gabrilovich, S. Markovitch: Computing semantic relatedness using Wikipedia-based explicit semantic analysis. Proceedings of IJCAI, 1606-1611, 2007.
21. Watson - The Semantic Web Gateway: Ontology Editor Plugins. http://watson.kmi.open.ac.uk. Online November 2008.
22. P. Knoth, L. Zilka, Z. Zdrahal: Using Explicit Semantic Analysis for Cross-Lingual Link Discovery. Workshop: 5th International Workshop on Cross Lingual Information Access: Computational Linguistics and the Information Need of Multilingual Societies (CLIA) at The 5th International Joint Conference on Natural Language Processing (IJC-NLP 2011), Chiang Mai, Thailand, 2011.

OntoVerbal-M: a Multilingual Verbaliser for SNOMED CT

Shao Fen Liang, Robert Stevens and Alan Rector

School of Computer Science, The University of Manchester, Oxford Road, Manchester, UK M13 9PL
{Fennie.Liang, Robert.Stevens, Rector}@cs.man.ac.uk

Abstract. OntoVerbal-M is an ontology verbaliser that transforms OWL into fluent natural language paragraphs in multiple languages. We describe the application of OntoVerbal-M to SNOMED CT, whereby SNOMED CT classes are presented as textual paragraphs in both English and Mandarin through the use of natural language generation. SNOMED CT is a large description logic based terminology for recording in electronic health records. Often, neither the labels nor the description logic definitions in SNOMED CT are easy for users to understand. Furthermore, information is increasingly being recorded, not just using individual SNOMED CT concepts, but using dynamically created description logic expressions ("post-coordinated" concepts). Such post-coordinated expressions can have no pre-assigned labels. In this context automatic verbalisation into multiple languages will be useful both for understanding and quality assurance of SNOMED CT definitions, and for helping different language-speaking-users to understand and share post-coordinated expressions.

Keywords: Multilingual Generation, Ontology Verbalisation, Ontology verbaliser, SNOMED verbalisation.

1. Introduction

We present OntoVerbal-M, a multi-lingual verbaliser for ontologies tailored to be used with SNOMED CT, a large medical terminology. Such ontologies and terminologies are increasingly authored in description logics, such as the W3C recommendation, the Web Ontology Language, OWL [2]. Expressions in Description Logics and OWL are often difficult for domain experts to understand [17]. Even using the human readable Manchester Syntax [10], expressions can have multiple levels of nesting and many inter-related axioms.

Verbalising these expressions in natural language is therefore attractive as a means to communicate with users [3; 4; 6]. Verbalisation has the added advantage that it should be possible to re-use some of the same language generation components in the generation of verbalisations in multiple languages.

SNOMED CT [19; 21] (Systematized Nomenclature of Medicine Clinical Terms) is big and potentially widely used OWL based terminology in any field. It attempts to provide a comprehensive terminology for use in medical records across all of

medicine, including diseases, diagnoses, procedures, anatomy, microorganisms and pharmaceuticals. It is maintained by the International Health Terminology Standards Development Organisation (IHTSDO)[1], and has been mandated or advocated for use in more than 50 countries. Today SNOMED CT is available in US English, UK English and Spanish. Translations to several other languages are currently taking place.

We have taken SNOMED CT as an example to demonstrate our techniques for verbalisation. In its OWL form, SNOMED CT is often awkward and even obscure. For example, the rendering of even just the definition of a simple concept such as *heart disease* in the raw OWL version of SNOMED CT is several lines long:

Class: Heart disease
EquivalentTo: Disorder of cardiovascular system
and RoleGroup *some* (Finding site *some* Heart structure)

By contrast, an English "verbalisation" of this definition in natural language as shown below will be easier for domain experts to understand, although it still seems somewhat stilted:

A heart disease is a disorder of the cardiovascular system that is found in the structure of the heart.

The verbalisation also omits the technically necessary, but to the domain expert mysterious, expression "RoleGroup".

When we attempt to present, not just the definition, but the information present in the ontology about a concept – *e.g.* Heart disease – the OWL expressions become more complex. Worse, they may not all be located together in the ontology. Hence the advantage of a verbaliser that presents the entire description of a concept in a single natural language paragraph, according to the discourse rules expressed in Rhetorical Structure Theory [14].

Using Rhetorical Structure Theory, furthermore, gives us a major component that appears to be re-usable across languages. The same mechanisms that produced the English above can generate Mandarin as:

心臟 病 是 由於 心臟 結構 異常 導致的 心血管
heart disease is from heart structure disorder caused cardiovascular
系統 失調
system disorder

Such verbalisations could be produced manually, but this is time consuming and, as mentioned, not possible for the dynamically created "post-coordinated" expressions for concepts.

OntoVerbal-M provides natural text descriptions with the aim of helping non-ontology experts understand the concepts in SNOMED CT. Currently, we have produced an English version using the official SNOMED CT labels and an experimental Mandarin version using ad hoc translations by a native speaker. The Mandarin must be taken with caution, as the translations of the individual labels are ad hoc and the validation has so far been only opportunistic. Nonetheless, the results have been sufficiently well received that we are strongly encouraged to extend the

[1] http://www.ihtsdo.com

study to a more formal analysis. In future, we hope to extend this to other languages and to compare verbalisations from OntoVerbal-M with manual translations.

It must be emphasised that OntoVerbal-M is not a machine translation system from one string to another. Rather it generates texts in multiple languages from the same underlying conceptual structure – ultimately a set of expressions in a description logic and the lexicon associated with those concepts in a particular language, as in other multilingual Natural Language Generation Systems [13; 18].

2. The OntoVerbal-M system

OntoVerbal-M is an extension of OntoVerbal. OntoVerbal was initially built for verbalising ontologies into English text [11], and has motivated us to test its top level rhetoric structure schema as a multilingual generator. Although there is no official SNOMED CT mandarin labels, we have tried our best using a mandarin native speaker's medical knowledge and consulting with an English SNOMED CT expert to produce mandarin labels as a test bed.

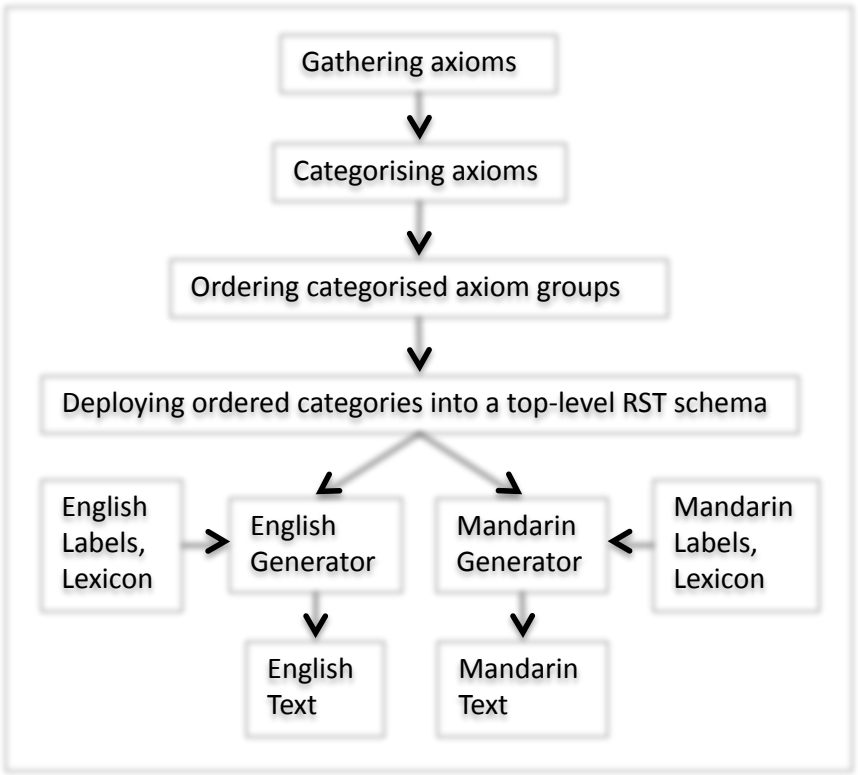


Fig. 1 The system architecture of OntoVerbal-M

OntoVerbal-M utilises the intuitive correlation between axioms and sentences to produce paragraphs that are more than simple collections of individual sentences. Instead, the sentences are structured and ordered [12]. This is achieved through five main operations: (a) gathering axioms together based on a shared focus class; (b) categorising the gathered axioms into different groups; (c) ordering categorised axiom groups; (d) deploying the ordered categories into a top-level discourse structure using Rhetorical Structure Theory (RST) [14]; and (e) using language generators to make the text hang together in a meaningful and organised manner.

OntoVerbal-M currently has two language generators as shown in Fig. 1, but they share a single discourse structure. Each generator has its own input labels as well as its own lexicons. Table 1 shows some examples of class labels in both languages.

Table 1 Example of class labels in both languages

SCT ID	English	Mandarin
302215000	thrombocytopenic disorder	血小板减少失調
107671003	vascular sclerosis	血管硬化
206596003	neonatal hypertension	新生兒的高血壓
10725009	benign hypertension	良性高血壓
113331007	structure of endocrine system	內分泌系統結構

Axioms in different notions are also transformed into sentences respectively according to the role of the focused class in the axiom as shown in Table 2. So, for example, a focus class "X" is to be expressed as a sub class of Y in an axiom, then this axiom is to be transformed into English as an X is a kind of Y, and in Mandarin as X屬於Y.

Table 2 Axiom transformation templates

Axiom notion	English template	Mandarin template
X sub class of …	An X is a kind of ..	X 屬於 …
X super class of …	A more specialised kind of X is..	X 包含了…

2.1. Applying natural language techniques

There are several natural language (NL) techniques that have been embedded in OntoVerbal-M. The first one is aggregation [8; 16]. For example "an X is a kind of an O", "an X is a kind of a P" and "an X is a kind of a Q". The three sentences are aggregated as "an X is a kind of an O, a P and a Q". The same technique is also applied to Mandarin to have the sentence as "X屬於O 、P和Q".

The second NL technique used is topic-maintenance-device [15]; This is used to avoid introducing a disfluency through the sudden shift of topic from one to another [22], and thus placing an additional cognitive load on the reader [7]. In general, axioms are expressed in one direction – from-child-to-parent – such as X sub class of Y, Y sub class of Z. However, there is often the chance that the focus class is in a parent position in an axiom. Therefore, in order to keep a topic consistent in a

generated text, instead of saying “an X is a kind of Y”, we need to say “a more specialised kind of Y is X” in English and “Y包含了X” in Mandarin to maintain a consistent topic for Y.

The third NL technique is the use of discourse markers [5; 20]. Discourse markers are applied when a focus class contains several axioms to be verbalised. Using discourse markers ensures the maintenance of fluency and coherence in a paragraph. For example, to connect an additional sentence from the above example, we use “additionally” in English and “而且” in Mandarin to produce the following paragraphs: “an X is a Y. A more specialised kind of X is Z. Additionally, an X is defined as a P that …”, and “X屬於Y 。它也包含了Z。而且X被定義為 P…中…”.

The fourth NL technique uses a set of key phrases to signal a change of topic in the generated text. Without such signalling, the text will lack coherence and fluency and be harder to understand. In cases where extra information should be given to the focus class, we introduce “Another relevant aspect of” or “Other relevant aspects of ” as key phrases in English and “其他與...相關的資訊” in mandarin to signal the topic change.

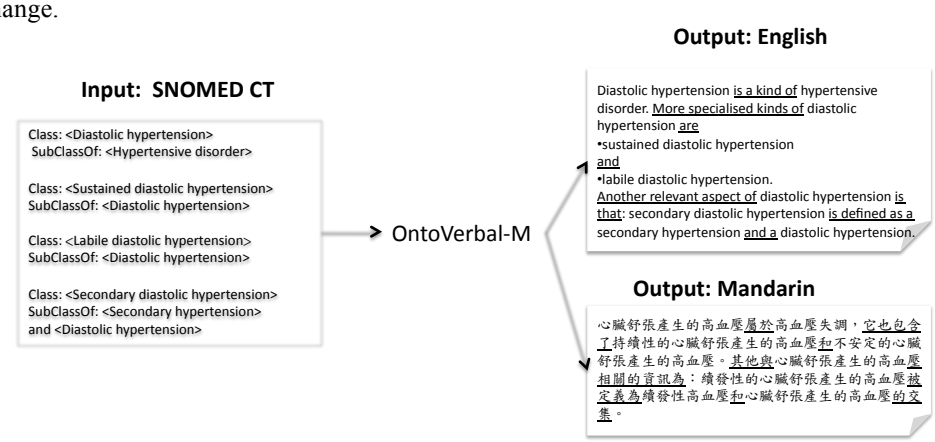


Fig. 2 Input and Output of OntoVerBal-M

Fig. 2 shows an example of an actual SNOMED CT concept input, and its English and Mandarin outputs in natural language. The non-underlined words are SNOMED CT labels, and the underlined words are system-selected words for text fluency purposes.

2.2. Results

Our primary goal is to provide text that not only has the structure of SNOMED CT concepts, but also to have them made clear. The textual output of OntoVerbal-M is thus faithful to the ontological input; a more idiomatic verbalisation is not our current goal. We show here some typical output in two languages; the input is the module

extracted[2] using as a signature *Hypertension* and all of its inferred subclasses from SNOMED CT full. The underlined words are system-generated words, and the non-underlined words are formal SNOMED CT terms. The following are five outputs from OntoVerbal-M. Each of them can be a verbalisation from one simple axiom to several complex axioms.

(a) Goldblatt hypertension is a kind of renovascular hypertension.
腎血管阻塞性高血壓屬於腎血管性高血壓。

(b) Disorder of the pelvis is defined as a disorder of the trunk that has a finding site in the pelvis.
骨盆失調被定義為軀幹失調中在骨盆的結構上有病灶。

(c) The cell is a kind of anatomical structure. More specialised kinds of the cell are
- entire cell

and
- subcellular structure.

細胞結構屬於解剖的結構，它也包含了全部的細胞和亞細胞組成的結構。

(d) Disorder of pregnancy is defined as a finding related to pregnancy and a disease. Another relevant aspect of a disorder of pregnancy is that: complication related to pregnancy is defined as a complication and a disorder of pregnancy.
妊娠失調被定義為妊娠相關的發現和疾病的交集。其他與妊娠失調相關的資訊為：妊娠相關的併發症被定義為併發症和妊娠失調的交集。

(e) Kidney disease is defined as a disorder of the genitourinary system that has a finding site in a kidney.
Other relevant aspects of a kidney disease include the following:
- renal impairment is a kind of kidney disease that has a finding site in a kidney;
- uremia is a kind of metabolic disease that is a kidney disease, and has a finding site in a kidney;
- renal vascular disorder is defined as a vascular disease of the abdomen that is a kidney disease, and has a finding site in a vessel of kidney;
- toxic nephropathy is defined as a kidney disease that has a causative agent in a substance;
- renal hypertension is defined as a secondary hypertension that is associated with a kidney disease;
- hypertensive renal disease is defined as a hypertensive disorder that is a kidney disease, and has a finding site in a kidney.

腎臟病被定義為生殖泌尿系統失調中在腎臟結構上有病灶。其他與腎臟病相關的資訊為：

一、腎臟損傷是一種腎臟病中在腎臟結構上有病灶。
二、尿毒症是一種新陳代謝疾病中的腎臟病和在腎臟結構上有病灶。
三、腎血管失調被定義為腹部血管的疾病中的腎臟病和在腎臟的血管結構上有病灶。
四、腎毒症被定義為腎臟病中在物質上有導致的藥物。
五、腎臟高血壓被定義為續發性高血壓中在腎臟病上有關聯。
六、腎性高血壓疾病被定義為高血壓失調中的腎臟病和在腎臟結構上有病灶。

[2] http://owl.cs.manchester.ac.uk/snomed/

3. SNOMED CT Challenges for English and Mandarin Text Generation

Most SNOMED CT class IDs have several associated terms [1], such as a "preferred term" (that is expected to be used most commonly in medical records and interfaces), and a "fully specified term" (that is intended to be completely unique and self explanatory). We have tried different SNOMED CT labels with OntoVerbal-M. So we get, for example, "Disorder of pelvic region is defined as a disorder of trunk that has a finding site in pelvic structure" from fully specified terms. We also get "Disorder of pelvis is defined as a disorder of trunk that has a finding site in pelvis" from SNOMED CT preferred terms. Neither of them has articles in the sentences as we have in the result section 2.2 (b).

3.1. The difference in using articles and plurality

The use of SNOMED CT supplied terms directly causes articles to be missed in the generated text, especially when definite articles are needed in the text. For example, if we used only the SNOMED CT supplied terms, we would have just "pelvis" and "trunk" in the output rather than "the pelvis" and "the trunk". The ontology alone does not provide sufficient information to deal with articles and plurals completely. For example, the use of singular or plurals in anatomy depends on whether the body normally has just one or more than one of a particular kind of part. The naming convention is, however, to take an unadorned singular form, such as "heart", rather than "the heart". Therefore, instead of naming a class "the heart structure", SNOMED CT actually names this class "heart structure". Our approach on dealing with the definite article and plurals is to look up online resources that can provide examples of usage of articles in human anatomical terms, so that we could replace the SNOMED CT labels with English expressions; Table 3 shows some examples from this approach.

Table 3 Relabelling anatomical terms

Original SNOMED CT label	New label
Procedure on thorax	Procedure on the thorax
Procedure on mediastinum	Procedure on the mediastinum
Procedure on abdomen	Procedure on the abdomen
Procedure on pelvis	Procedure on the pelvis
Procedure on heart	Procedure on the heart
Disorder of soft tissue of thoracic cavity	Disorder of soft tissue of the thoracic cavity
Branch of abdominal aorta	Branch of the abdominal aorta
Finding of cellular component of blood	Finding of cellular component of the blood

This approach is, however, not perfect, as some terms in SNOMED CT are either missing or in different phrasing order from our looked-up resources. For example, we are able to change "pelvis" to "the pelvis" but unable to change "pelvic structure" to "the pelvic structure". A further process, such as the use of the Unified Medical

Language System (UMLS)[3], would be needed to improve this problem to change the adjectival form for this case.

In Mandarin, articles and plurality are not as a concern, as they are in generating English text. The difference between these two languages is that English has count nouns in a singular or plural form, or even to have mass nouns, while Mandarin can only modify a noun by adding an adjective or numeral in front of the noun. For example: "你的(your)心臟(heart)", "我的(my)心臟(heart)", "一個(one)心臟(heart)" or "二個(two)心臟(heart)". Wherever the "心臟" appears in a text, it never becomes "the心臟" or "心臟s". This means the fixes applied for articles in English NLG of SNOMED CT are not needed in the Mandarin version.

3.2. The difference in generality

Logical conjunction is one of the important methods used in designing SNOMED CT terms. It reveals the semantic relationship between the clinical medical concepts and their logical triple structure in order to present clinical information. For example:

Class: Acute metabolic disorder
EquivalentTo: Acute disease *and* Metabolic disease

The *Acute metabolic disorder* is an intersection between *Acute disease* and *Metabolic disease*. This axiom is transformed into English as "Acute metabolic disorder is defined as both an acute disease and a metabolic disease", where the intersection in this axiom is not shown directly in the English verbalisation. This verbalisation can be ambiguous for non-native English speakers in understanding that *Acute metabolic disorder = Acute disease* ∪ *Metabolic disease* rather than *Acute metabolic disorder = Acute disease* ∩ *Metabolic disease*.

Why not just simply transformed the above axiom as "Acute metabolic disorder is defined as an intersection between an acute disease and a metabolic disease"? The reason is that the word "intersection" is a mathematical word, which is not commonly used outside Mathematics. Also, translating "intersection" unambiguously into English is awkward at best.

In comparison, the same axiom is transformed into Mandarin as "急性(Acute)新陳代謝(metabolic)失調(disorder)是(is)急性(acute)疾病(disease)和(and)新陳代謝(metabolic)疾病(disease)的(apostrophe)交集(intersection)。", where 交集(intersection) is the word – intersection, and is just simply to be used to express its role. In fact, the phrase 交集 is not awkward in Mandarin's daily conversation. For example, if A decides to break a relationship with B. A can say to B "your life has no 交集(intersection) with me", or A complains about B and says "my conversation with B has no 交集(intersection)".

[3] http://www.nlm.nih.gov/research/umls/

3.3. The different role of properties in the translation

Properties are one of the problems in ontology verbalisation due to the lack of NLG orientated guidelines for labelling properties [9]. Morphological features of the first word in a property have an impact on producing fluent text automatically. For example, the first word of a property could be a noun, an adjective, a verb or a preposition as shown in Table 4.

Table 4 Morphological features of SNOMED CT properties

Morphological feature of the first word	Frequency	Example
Noun	33	Procedure site – direct
Adjective	12	Clinical course
Verb in its present tense and 3rd person singular form	8	Has focus
Verb in its present participle	4	Using device
Verb in its past participle	4	Associated with
Preposition	1	After

With each different initial morphology, the verbaliser needs a Part Of Speech (POS) checking in order to choose a correct verb for generating a sentence such as using ”is” or “has” or without adding verbs. Our experience has suggested that a standardised term modelling approach will represent ontologies well, and will also save much time when building ontology verbalisers. The quality of the verbalisation will improve dramatically if ontology classes and properties are well phrased or annotated such that the linguistic behaviour of the concept or property is apparent. For example: class labels are noun phrases; property labels start with a third personal singular verb and end with a preposition. In this case, a property would act as a nice predicate between its subject and object classes. This way would free a verbaliser from concerning itself with or without a verb and the text fluency while transforming properties without appropriate prepositions.

In SNOMED CT the property “after” is a good example. If it were lexicalised as “has an after effect in” then the *postoperative complication* concept can be verbalised as “postoperative complication is defined as a complication of a surgical procedure that has an after affect in a surgical procedure”. In this case the verbaliser only needs to concern itself about the article in this sentence.

The issues on phrasing class and property labels in verbalising English SNOMED CT has led us to be more careful in translating Mandarin labels. We adapt the standardised term modelling approach suggested from English labels to translate Mandarin labels. So every class is translated into a noun phrase, and every property starts with a verb in the Mandarin labels.

Table 5 shows examples of our manually annotated property labels in both English and Mandarin according to the suggested standardised term modelling approach. However, because of the different sentence structure between English and Mandarin, instead of ending each property with a preposition, Mandarin properties are ended with nouns. Therefore each axiom can be transformed to start with a noun subject

class, and what ever happens in between, then it ends with a property in a Mandarin sentence.

Table 5 Example of property labels in suggested standardised form

SCT ID	English	Mandarin
255234002	has an after affect in	有(has)後遺症(after affect)
246454002	has an occurrence in	出現(appears)異常(difference)
116676008	has an associated morphology in	有(has)關聯的(associated)形態(morphology)
363698007	has a finding site in	有(has)病灶(finding site)
263502005	has a clinical course in	有(has)臨床的(clinical)療程(course)

The following text is an example that shows English text ending with a verbalised class but Mandarin ends with a verbalised property:

English: Renal arterial hypertension is a kind of renovascular hypertension that has a finding site in a kidney

Mandarin: 腎臟(renal)動脈(artery)的(apostrophe)高血壓(hypertension)是(is)一種(a kind of)腎血管性(renovascular)高血壓(hypertension)中(among)在(at)腎臟(kidney)結構(structure)上(upon)有(has)病灶(finding site)。

4. Discussion

OntoVerbal-M currently produces well-structured English and Mandarin natural languages for the fragment of SNOMED CT so far studied. Natural language texts are easier to understand than DL based terminologies and ontologies, especially for non-DL users. This motivates the need for automatically generated verbalisations. When users in multiple languages for which there are no full translations want at least limited access to the content, then multilingual generation becomes important.

It is striking that the same rhetorical structure schema appears to be applicable across two such different languages, despite marked differences in grammar and syntax. This significantly reduces the effort required to produce verbalisers in different languages as significant portions of the verbalisation machinery can be re-used.

Clearly, there are dangers of erroneous verbalisations, particularly in "new" languages such as Mandarin. Our Mandarin labels were not developed by a team of Mandarin speaking medical experts, but purely from our own team knowledge. At this time they must be regarded as experimental and used for OntoVerbal-M's purpose only. However, a small survey from Mandarin speaking doctors without SNOMED CT knowledge has indicated that the non-underlined medical terms of OntoVerbal's output are appropriate to express the meaning of the output texts. The survey also indicates that the underlined words we have chosen are generally suitable for text fluency purposes.

Although OntoVerbal-M's output needs some linguistic polish, especially in plurality and articles in English, its design in, first: organising information into super,

sub and equivalent classes, second: transforming OWL classes and properties into text, and third: the use of discourse structure for generating text, would apply to ontologies from any domain. Specific to SNOMED CT, OntoVerbal-M has annotated classes and properties for text fluency purposes, and particularly to deal with human anatomical phrasing. Its experiences have raised the issues in phrasing ontology terms in English and Mandarin.

In the future, we plan to evaluate the text generated by OntoVerbal-M in two aspects: a) whether the text is faithful to the ontology, so the subjects need to be ontologist to be able to read text and regenerate ontology axioms; b) whether the translation into Mandarin is equally faithful. The participants in evaluation a), and ideally b), need to be SNOMED CT experts so that they understand both the axiomatic descriptions and a natural language text. Failing this, we expect to ask a broad team of domain experts to identify definitions that appear questionable and then consult a more limited team of SNOMED CT experts about those identified as questionable. We will also explore if OntoVerbal-M's system architecture can adopt more language generators such as French, Spanish and German.

Acknowledgments. This work is part of the Semantic Web Authoring Tool (SWAT) project (see www.swatproject.org), which is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) grant EP/G032459/1, to the University of Manchester, the University of Sussex, and the Open University. We are extremely grateful to Professor Donia Scott for her critical feedback on an earlier draft of this article. The Mandarin survey has been supported by doctors from Chang Gung Medical Foundation LinKo Taiwan. We also thank Dr. Wu from Kaohsiung to give his useful comments on the medical terms translation.

References

1. SNOMED-CT User Guide, http://www.ihtsdo.org/fileadmin/user_upload/Docs_01/SNOMED_CT/About_SNOMED_CT/Use_of_SNOMED_CT/SNOMED_CT_User_Guide_20090731.pdf
2. Baader, F., Horrocks, I., and Sattler, U.: Description logics as ontology languages for the semantic web. Lecture Notes in Artificial Intelligence. 2605, 228-248. (2005)
3. Baud, R., Lovis, C., Alpay, L. *et al*.: Modelling for natural language understanding. In: 17th Annual Symposium on Computer Applications in Medical Care (SCAMC-93), pp. 289-293. McGraw Hill, (1993)
4. Baud, R.H., Rodrigues, J.-M., Wagner, J.C. *et al*.: Validation of concept representation using natural language generation. Journal of the American Medical Informatics Association, 841. (1997)
5. Callaway, C.B.: Integrating discourse markers into a pipelined natural language generation architecture. 41st Annual Meeting on Association for Computational Linguistics. 1, 264-271. (2003)
6. Ceusters, W., and Spyns, P.: From natural language to formal language: when MultiTALE meets GALEN. In: Medical Informatics Europe '97, pp. 396-400. (1997)
7. Clark, H.H.: Psycholinguistics. MIT Press, (1999)
8. Dalianis, H.: Aggregation as a subtask of text and sentence planning. In: Florida AI Research Symposium, FLAIRS-96, pp. 1-5. J.H.Stewman, (1996)
9. Fliedl, G.n., Kop, C., and Vo¨hringer, J.r.: Guideline based evaluation and verbalization of OWL class and property labels. Data & Knowledge Engineering. 69, 331-342. (2010)

10. Horridge, M., Drummond, N., Goodwin, J. *et al*.: The Manchester OWL syntax. In: 2006 OWL: Experiences and Directions (OWLED'06). (2006)
11. Liang, S.F., Stevens, R., Scott, D. *et al*.: Automatic Verbalisation of SNOMED Classes Using OntoVerbal. In: 13th Conference on Artificial Intelligence in Medicine, AIME 2011, pp. 338-342. (2011)
12. Liang, S.F., Scott, D., Stevens, R. *et al*.: Unlocking Medical Ontologies for Non-Ontology Experts. In: 2011 Workshop on Biomedical Natural Language Processing, ACL-HLT 2011, pp. 174-181. (2011)
13. Linden, K.V., and Scott, D.: Raising the Interlingual Ceiling in Multilingual Text Generation. Proceedings of the Multilingual Natural Language Generation Workshop. In: International Joint Conference in Artificial Intelligence (IJCAI'95)t, pp. 95-109. (1995)
14. Mann, W.C., and Thompson, S.A.: Rhetorical Structure Theory: toward a functional theory of text organisation. Text. 8, 243-281. (1988)
15. McKeown, K.R.: Discourse Strategies for Generating Natural Language Text. Artificial Intelligence. 27, 1-41. (1985)
16. Reape, M., and Mellish, C.: Just what is aggregation, anyway? In: European Workshop on Natural Language Generation. (1999)
17. Schulz, S., Stenzhorn, H., Boeker, M. *et al*.: Strengths and limitations of formal ontologies in the biomedical domain. Electronic Journal of Communication Information & Innovation in Health. 3, 31-45. (2009)
18. Scott, D., Bouayad-Agha, N., Power, R. *et al*.: PILLS: A Multilingual Authoring System for Patient Information. In: the 2001 Meeting of the American Medical Informatics Association (AMAI'01). (2001)
19. Spackman, K.A., and Campbell, K.E.: Compositional concept representation using SNOMED: Towards further convergence of clinical terminologies. Journal of the American Medical Informatics Association, 740-744. (1998)
20. Sporleder, C., and Lascarides, A.: Using automatically labelled examples to classify rhetorical relations: an assessment. Natural Language Engineering. 14, 369-416. (2008)
21. Stearns, M.Q., Price, C., Spackman, K.A. *et al*.: SNOMED clinical terms: overview of the development process and project status. In: AMIA Fall Symposium (AMIA-2001), pp. 662-666. Henley & Belfus, (2001)
22. Walker, M.A., Joshi, A.K., and Prince, E.F.: Centering Theory in Discourse. Oxford University Press, (1998)

Representing Translations on the Semantic Web

Elena Montiel-Ponsoda, Jorge Gracia, Guadalupe Aguado-de-Cea, and Asunción Gómez-Pérez

Ontology Engineering Group, Dpto. Inteligencia Artificial
Facultad de Informática, Universidad Politécnica de Madrid
Campus de Montegancedo s/n, 28660 Boadilla del Monte, Madrid, Spain
jgracia, emontiel, lupe, asun@fi.upm.es
http://www.oeg-upm.net/

Abstract. The increase of ontologies and data sets published in the Web in languages other than English raises some issues related to the representation of linguistic (multilingual) information in ontologies. Such linguistic descriptions can contribute to the establishment of links between ontologies and data sets described in multiple natural languages in the Linked Open Data cloud. For these reasons, several models have been proposed recently to enable richer linguistic descriptions in ontologies. Among them, we find *lemon*, an RDF ontology-lexicon model that defines specific modules for different types of linguistic descriptions. In this contribution we propose a new module to represent translation relations between lexicons in different natural languages associated to the same ontology or belonging to different ontologies. This module can enable the representation of different types of translation relations, as well as translation metadata such as provenance or the reliability score of translations.

Keywords: multilingual Semantic Web, multilingual Linked Data, *lemon* model, translation relations

1 Introduction

The Linked Open Data [1, 2] initiative has triggered the publication and linking of data sets in the RDF [13] format, contributing in this way to semantically structuring huge amounts of data on the Web. Thanks to the representation format propounded by Linked Data, concepts are connected across resources, breaking down the barriers imposed by data silos, and enabling machines to smartly navigate the Web as a big data set. Currently, more than 250 data sets containing more than 30 billion triples are available in the Linked Open Data (LOD) cloud[1], ranging from domains as far apart as biomedicine, music or geography. Governmental institutions, enterprises and the private sector have realized the benefits and potential of such an initiative and have made their data sets available for linking and exploitation by third parties.

[1] http://www4.wiwiss.fu-berlin.de/lodcloud/state/

The launching phase of the LOD was led by English speaking countries, but in recent years, the LOD cloud has also seen an increase in resources documented in languages other than English. By having a quick look at the CKAN[2] catalogue of data sets, we come across the *data.bnf.fr* data set from the French National Library, the *GeoLinkedData.es* data set of Spanish geographical data, *Rechtspraak.nl* from the Netherlands Council of the Judiciary, or the *FAO geopolitical ontology* with labels in English, French, Spanish, Arabic, Chinese, Russian and Italian.

This proliferation of semantic data described in several natural languages evidences the need for accounting for the linguistic information relative to ontologies and linked data because of several reasons. One of the main reasons is that the linguistic descriptions of these resources help in finding and establishing mappings between concepts and individuals of different ontologies and data sets [22]. Another evident reason is that such descriptions contribute to a better exploitation of the data sets by tasks such as information extraction [19], natural language generation [3], or multilingual data access [7], to mention but a few.

Several formats and annotation properties have been developed in the Semantic Web to represent natural language descriptions associated to ontologies and linked data, such as the *rdfs:label* [13] or *skos:prefLabel* [15] properties. Their limitations have been discussed in several fora [5, 18, 14], and extensions or new models have been proposed in the last years for the representation of linguistic descriptions relative to ontologies and linked data in more principled ways. Some of these models are SKOS-XL [16], LexInfo [5], LIR [18], or the recently appeared *lemon* model [14]. Most of these models also provide some mechanisms to allow for the representation of multilingual descriptions associated to the same ontological representation. However, we argue that explicit relations between descriptions in different languages, i.e., translation relations, as well as translation descriptive metadata, would help in a more efficient exploitation of these multilingual annotations. Moreover, they would also contribute to the establishment of principled links between ontologies and data sets described in multiple natural languages in the LOD cloud.

In this paper, we propose a representation mechanism of translations between labels in different languages associated to ontology terms. To that end, we propose a metamodel in OWL which extends the *lemon* ontology, and which is offered as a module of the *lemon* model. *lemon* is a linguistic model developed in the framework of the Monnet[3] project to represent lexical and terminological descriptions relative to an ontology. The *lemon* extension we propose in this paper enables the representation of translations in a separate linguistic layer, thus leaving the original ontologies or data sources untouched. It also contributes to the linking of ontologies and data sets described in different natural languages in the Web of Data.

The rest of the paper is organized as follows. Section 2 summarizes the mechanisms that some Semantic Web formats or models have for linking linguistic

[2] http://ckan.net/
[3] http://www.monnet-project.eu/

descriptions in several natural languages. In section 3, we analyze the problem of translation relations in the context of the Semantic Web. After that, in section 4, we briefly present the *lemon* model. Thanks to the modular conception of this model, we are now able to propose a translation module, i.e., a module to explicitly represent translations in *lemon*. Section 5 will be devoted to a detailed description of the translation module, and some examples will be provided to illustrate the use of this module. Finally, we conclude the paper in section 6.

2 Related work

As it is well known, RDFS [13] and SKOS [15] rely on limited annotation properties to represent labels or linguistic descriptions associated to ontologies and linked data. They also enable a simple form of multilingual labeling by using language tags to restrict the scope of a label to a particular language (e.g., *skos:prefLabel "bank"@en*). This representation allows for *indirect* or *non-explicit links* between or among multilingual labels, when associated to the same resource in the data set.

Conscious of these limitations, SKOS developers worked on an extension of SKOS called SKOS-XL [16] , that allows to make links explicit between labels associated to the same concept. This extension introduces a *skosxl:Label* class that allows labels to be treated as first-order RDF resources, and a *skosxl:labelRelation* property that provides links between the instances of *skosxl:Label* classes. In this way, we can specialize the *skosxl:labelRelation* into a translation relation and explicitly link *skosxl:Label* instances in different natural languages.

The LIR [18] model also focuses on the representation of links between labels within and across natural languages. This model was created with the purpose of keeping the ontology and the linguistic information independent from each other, so that lexical and terminological properties of labels could be further described (e.g., part-of-speech, gender, terminological variants). The relations provided by LIR to labels within the same natural language have lexical (*hasSynonym*, *hasAntonym*) or terminological nature (*hasVariant*, *hasAbbreviation*, *hasTransliteration*, etc.). And the ones between labels across different natural languages have a translational nature (*hasTranslation* or *hasScientificName*).

Now, the relations provided by the SKOS-XL and LIR models, though being useful for certain applications because of the explicitness of the *hasTranslation* relation between labels in different natural languages, do not allow to account for some aspects of the translation process that may also be relevant for certain applications. For instance, the difference between original and target label. This may be interesting in the case that we have an ontology documented in four natural languages, and we want to specify which labels (or which linguistic descriptions) have been taken as the source in the translation process. Another aspect to be considered could be the type of translation relation existing between labels (we will come back to this in section 3). Moreover, the provenance, i.e., the resource from which translations have been obtained may also be the kind of metadata that enriches the information about translation. Finally, it is

important to account for the adequacy and reliability of the translation in the specific context of the ontology. An extension of these models would be required to represent further translation metadata. However, we have chosen the *lemon* model for this purpose, because its design principles make it specially appropriate for the Web of Data scenario. Firstly, *lemon* introduces a 'well-defined lexical-conceptual' path between linguistic descriptions and ontology elements. Secondly, *lemon* has been designed as a concise RDF model that captures complex linguistic descriptions by dereferencing resources that contain them. And thirdly, it is an extensible and modular model, which allows the use or inclusion of certain modules if so required by the final application. These and other features of the model will be further detailed in section 4.

Finally, we will refer to the *LOD in Translation work*[4], in which a model has been created to describe and retrieve translations in the LOD cloud relying on resources that contain labels in different natural languages. This model takes advantage of multilingual labels associated to resources by means of language tags (as in *rdfs:label "bank"@en*, *rdfs:label "Bank"@de*, *rdfs:label "banco"@es*) and retrieves available translations. Our purpose, on the other hand, is to contribute to the creation of explicit translation links within the same data source and across data sources, so that this and other systems can benefit from the multilingual data in the LOD cloud.

3 Translation relations in the Semantic Web

Ontology localization [21, 8, 6] has been defined as the activity of adapting an ontology to the needs of a particular (linguistic and cultural) community. Methodological guidelines, tools and models have been developed to support the ontology localization activity, which normally results in an ontology in which labels are documented in multiple natural languages, what is the same, a multilingual ontology [6]. Since the different linguistic versions are assumed to be pointing to the same ontology concepts, it could be derived that they are all translations of each other. However, if we have several terms in each language (synonyms or term variants), we may want to unambiguously express which term variant in language A is translation of which term variant in language B. At this point, translation relations acquire significance.

Let us illustrate this with a simple example. In the FAO geopolitical ontology mentioned in the introduction, one ontology term may describe the organization as such and have the labels "Food and Agriculture Organization" and "FAO". Translations of full form and acronym will be provided in the rest of languages, and, ideally, explicit links will be created between the full forms and the acronyms, respectively.

However, translation relations are not always so direct and simple. As claimed in [8, 6], depending on the type of conceptualization represented in the ontology, direct translations in the target language will be available or not. A distinction

[4] http://sites.google.com/site/pierreyvesvandenbussche/apps/lod-in-translation

is made between the so-called *internationalized or standardized conceptualizations*, and conceptualizations more prone to reproduce the vision of the world of a certain community, the so-called, *culturally-influenced domains*. When localizing ontologies of these two types, translation relations may also need to be of different types. To put it in other words, when dealing with internationalized domains, i.e., *technical or specialized domains of knowledge such as engineering or medicine that have standards for processes and descriptions, and whose categorizations usually reflect the common view of different cultures* [17], we may find translations for all terms describing the concepts in the ontology, since the same conceptualization is shared among the languages represented in the ontology. Contrary to that, when localizing ontologies representing culturally-influenced domains, in which the granularity level of some concepts may differ from culture to culture, we may come across mismatches that need to be solved to provide adequate translations. Under this group we include domains such as law, geography or the political and administrative organization of countries, universities, and so on.

Imagine an ontology of financial institutions in Germany. One of the concepts represented in the ontology may be *Sparkasse* (which we could generally translate as *savings bank* in English). However, there may be differences between these concepts concerning business purpose, ownership or governance of the institution. So, maybe, a more adequate translation of *Sparkasse* could be *German savings institution*, although we usually tend to look for the *closest equivalent concept* in the target language and get the term used to refer to it, i.e., *savings bank* in this case. This simple example aims at illustrating the difference between 'literal or documentary translations', and 'functional translations'[5]. The first type usually describes the concept in the target language, because there is no *exact equivalence* in the target language. The second type looks for the *closest equivalence* -though being conscious of the existence of disparities- because it may be convenient for practical reasons. For instance, when aiming at interoperability (at a European or international level), near-equivalents are assumed to match although a complete overlap between them does not exist.

According to this, we make a distinction between *literal translations* and *cultural equivalences*. In the context of the Semantic Web, this distinction may be quite simple to make. The literal translation would be pointing to the same ontology concept, whereas the cultural equivalent would most probably belong to an equivalent ontology documented in the target language. See figure 1 for an illustration of this. Ontology A is an ontology of German credit institutions in which labels have been translated into English, whereas Ontology B conceptualizes the structuring of British credit institutions in English. It would be highly interesting to specify the links between these terms in a multilingual scenario. For these reasons, we claim that further specifications of the translation relation would contribute to envisage a true Multilingual Semantic Web.

[5] Many practitioners and translation theorists agree on this difference and speak about *overt* vs. *covert* translation [11], or *documentary* vs. *instrumental* or *functional* translation [20], respectively.

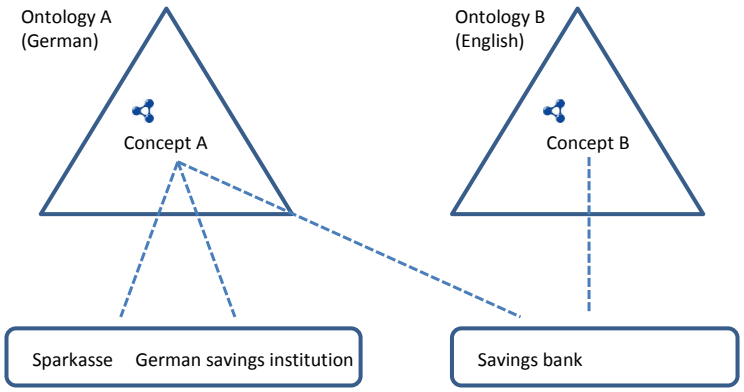


Fig. 1. Oversimplified example of literal translation and cultural equivalence links

4 *lemon*, an interchange model for the Multilingual Semantic Web

The *lemon* model (*le*xicon *m*odel for *on*tologies) [14] is an RDF model of linguistic descriptions that has been designed to a) be published with ontologies, b) extend their lexical layer with as much linguistic information as needed, and c) exchange the resulting lexical resources on the Web. Technical details and usage of the model can be found at http://lexinfo.net/lemon-cookbook.pdf The main features of the model can be summarized as follows:

- Linguistic descriptions are kept separated from the ontology, but their semantics are defined by pointing to the corresponding semantic objects in the ontology (what has been called 'semantics by reference' [4]).
- The model consists of a core set of classes (as described below) and several modules capturing different types of lexical and terminological descriptions.
- Rich lexical and terminological descriptions are grouped into five modules: linguistic properties (part-of-speech, gender, number...), lexical and terminological variation, decompositions of phrase structures (representation of multi-word expressions), syntactic frames and their mappings to the logical predicates in the ontology, and morphological decomposition of lexical forms.
- Linguistic annotations (data categories or linguistic descriptors) are not captured in the model, but have to be specified for each lexicon by dereferencing their URIs as defined in the repositories that contain them (for instance, the ISOcat repository [12]).

The different types of linguistic descriptions captured by the model and its main classes can be seen in figure 2. The core classes of the model are the ones that form the main path between the *Ontology* and the lexical variants represented in the *LexicalEntry* class. The *LexicalSense* class provides a principled link between an ontology concept and its lexical materialization (*LexicalEntry*).

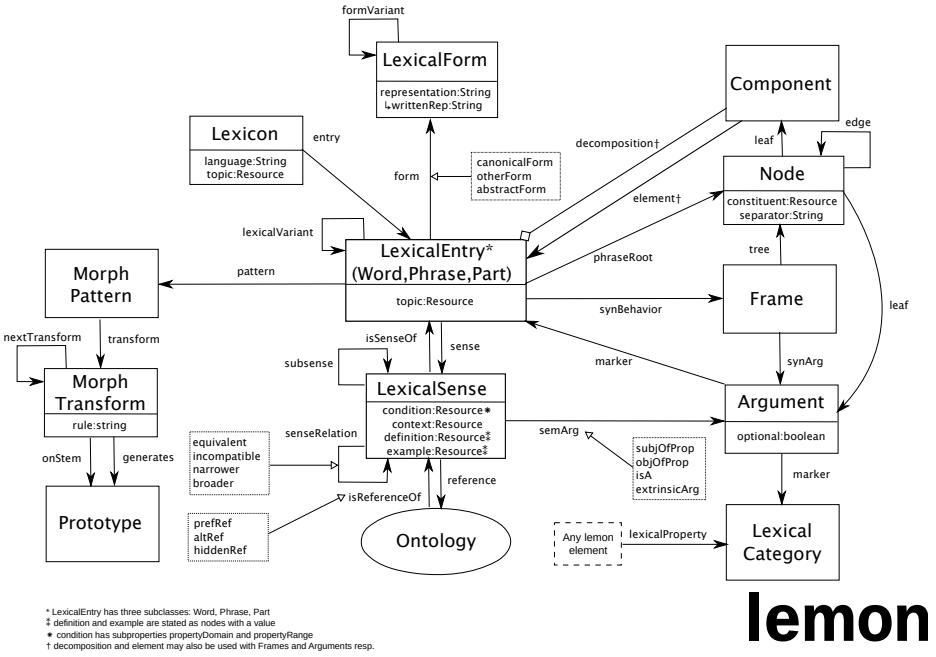


Fig. 2. Core classes and modules of the lemon model

Since 'concepts', as defined in ontologies, and 'lexical entries', as defined in lexicons, cannot be said to overlap [10], the *LexicalSense* class provides the adequate restrictions (usage, context, register, etc.) that make a certain lexical entry appropriate for naming a certain concept in the specific context of the ontology being lexicalized.

LexicalSense is also the class that is foreseen to provide the links between lexical entries within and across languages. Four specializations of this relation are provided: equivalent, incompatible, narrower and broader, as illustrated in figure 2. As the *lemon* model defines one lexicon per language, translation relations could be inferred as lexical entries in different languages would be all pointing to the same ontology reference. However, it is also foreseen to make this type of relation explicit between lexical senses, in the case that, for instance, lexical entries are not pointing to the same ontology reference, but belong to the linguistic descriptions associated to other ontologies.

As such, the translation relation between lexical senses is a powerful mechanism to represent translations. Nevertheless, and as already pointed out in section 1, when dealing with translations, additional properties of the translation relation need to be made explicit, such as reliability score, provenance, or type of translation relation, as already introduced in section 2. In this sense, the flexibility provided by the *lemon* model by means of modules allows us to propose a so-called 'translation module', by reifying a translation relation be-

tween lexical senses into a class. The use of such a module could be exploited by applications that require multilingual ontologies and want to keep track of the relations between the lexical entries in different languages. This information would be very valuable if translations have been automatically generated via an ontology localization system (e.g., LabelTranslator NeOn Toolkit plug-in [9]).

5 *lemon* module for translations

In this section we describe the entities of the translation module in *lemon*[6] and illustrate its use by means of some examples. Figure 3 shows the class diagram of the translation ontology. Some classes are imported from the core of the *lemon* ontology, namely *Lexicon*, *LexicalEntry*, *Form*, and *LexicalSense*.

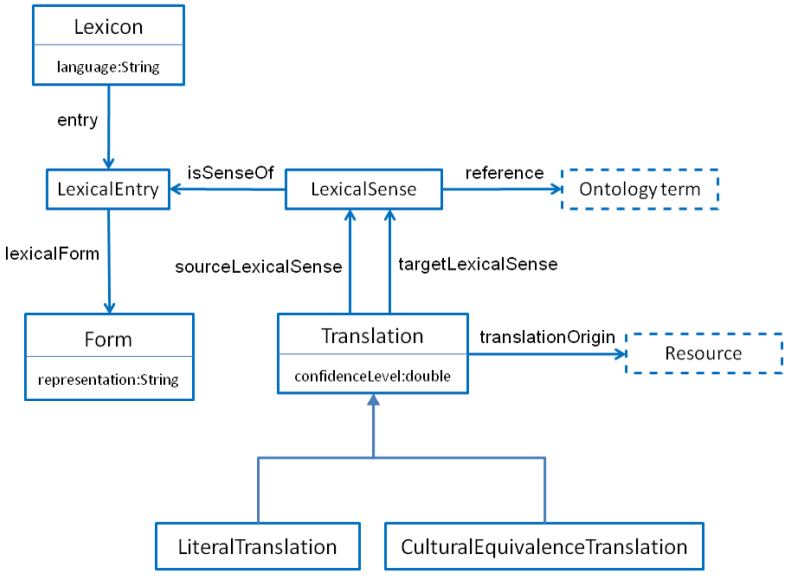


Fig. 3. *lemon* Translation module

– *Translation*. This is the central class of the translation module. It mediates the translation relation between lexical senses, and contains also information that characterizes the translation process, such as a *confidence level*. This confidence level will ultimately depend on the translation tools and translation resources employed to obtain translations. We do not deal here with the algorithms used for its computation, but it will typically combine different features such as probabilities of translation systems, reliability of translations resources, scores of disambiguation methods, etc.

[6] It will be available at http://www.monnet-project.eu/lemon_translation.owl

- *Literal Translation*. It is a subtype of the translation class that corresponds to the idea of literal translations mentioned above.
- *Cultural Equivalence Translation*. A subtype of the translation class that covers translations that are not literal, but close cultural equivalences between the languages considered.
- *Resource*. It represents resources from which translations have been obtained.
- *Lexical Sense*. A sense links a lexical entry to the reference (ontology term) used to represent its meaning.
- *Lexical Entry*. It is a container of the different forms and meanings of a lexeme.
- *Form*. An inflectional form of an entry. It admits several representations (written, phonetic, etc.).
- *Lexicon*. This class represents the whole lexicon. It has a language associated, so it is assumed to be monolingual. Translations will typically connect entries between different monolingual lexicons.

5.1 Examples of use of the *lemon* translation module

In order to illustrate the usage of the translation module, in this section we provide some examples of the financial and politics domains.

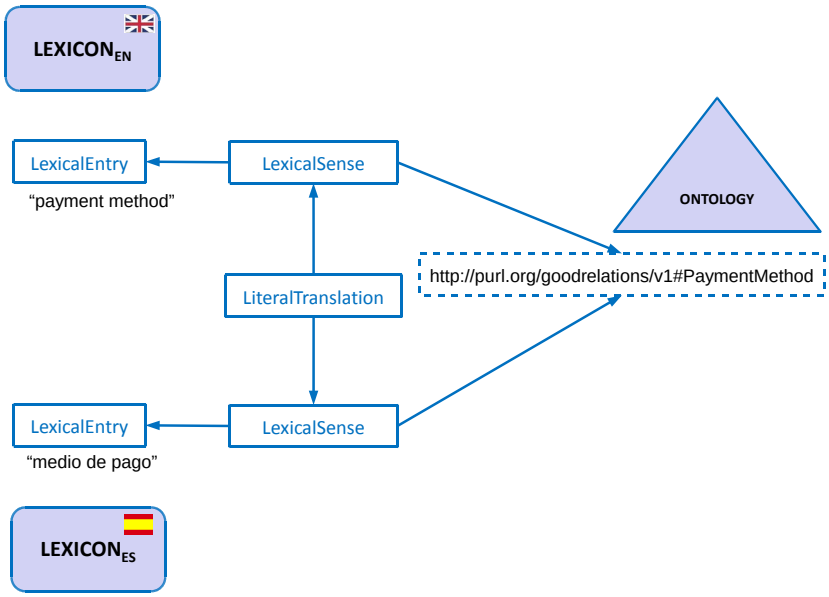


Fig. 4. Example of literal translation

Figure 4 represents an ontology term extracted from the GoodRelations ontology[7]. In *lemon* we would be able to associate as many lexicons in different languages to the ontology as wished. In the figure, we show two lexicons that have been associated to the ontology: one lexicon with English descriptions and the other with Spanish descriptions. Both lexicalize in different languages the same ontology concept, namely, *http://purl.org/goodrelations/v1#PaymentMethod*. Each lexicon contains a lexical entry and a lexical sense representing the ontology concept in each language. The lexical sense belonging to the English lexicon would be the *sourceLexicalSense*, and the one of the Spanish lexicon would be the *targetLexicalSense*, since the ontology was conceived in English. The provenance of the translation would be specified at the *Resource* class. It could be an on-line resource (machine translation service), a lexicon or terminology of the domain, or even a human translator. A confidence value could also be assigned to the translation by means of the *confidenceLevel* property of the *Translation* class. Finally, we would relate these two translations by means of the *LiteralTranslation*, subclass of the *Translation* class. **This would mean that in the specific context of the ontology being lexicalized and localized, the target lexical sense provides a description or literal translation of the term, which is to be used in the context of the original ontology.** It is highly probable that the Spanish translation "medio de pago" is also its cultural equivalent, which would mean that the same concept exists in the Spanish financial system and has been termed as the literal translation. So in this case, both translation relations would be valid.

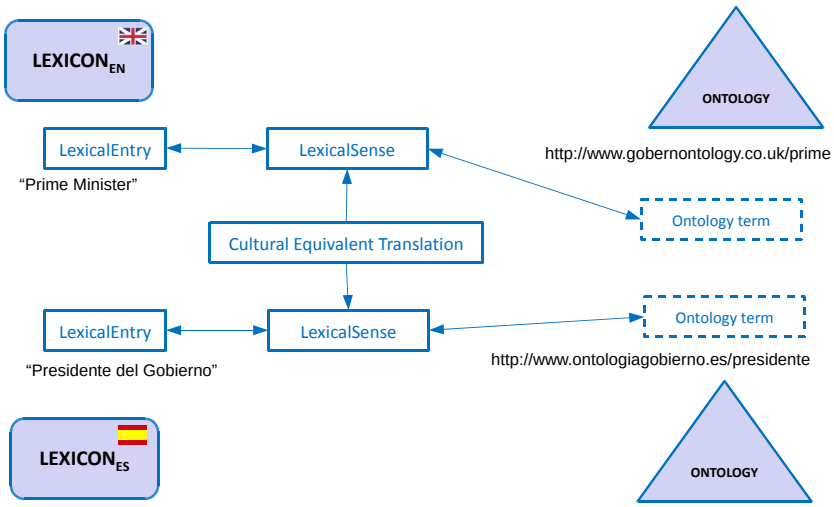


Fig. 5. Example of Cultural Equivalence

[7] http://www.heppnetz.de/ontologies/goodrelations/v1

Now, let us have a look at figure 5. This aims at illustrating cultural equivalents between political systems. Here we have two ontologies, each one representing a different political system, and each one documented in a different natural language. The concept of "Prime Minister" in the British political system and the concept of "Presidente del Gobierno" in the Spanish political system are not exact equivalents, but can be considered the closest equivalents in the respective cultures. This is why we would use the class *CulturalEquivalenceTranslation* to relate the two lexical senses that we assume would belong to two lexicons associated to two different ontologies. **Such a relation would indicate that these two terms are substitutable or translations of each other, when looking for interoperability and referring to (close) equivalents in different languages and cultures, whose extention may not completely overlap.** In this case, we could also include literal translations of each lexical entry in the respective lexicons. In the English lexicon we could include the Spanish lexical entry "Primer Ministro Británico", which would be a literal translation in Spanish. In the same way, we could also add the lexical entry "Spanish President" or "Spanish President of the Government" in the Spanish lexicon. These translations would be related to each other by the *LiteralTranslation* class.

6 Conclusions

The publication of ontologies and data sets in multiple natural languages has raised some issues related to the representation of the linguistic descriptions relative to ontologies. In the context of Linked Data, this takes on more importance since ontologies and data sets described in different natural languages have to be linked to each other. Moreover, such natural language descriptions have proven essential in enabling the exploitation of semantically structured knowledge by language-based tasks. With the purpose of establishing explicit links between the linguistic descriptions associated to ontologies and linked data in several natural languages, in this paper we propose an extension of the *lemon* model to represent translation relations. This translation module allows us to differentiate between *literal* and *cultural equivalence* translations. In addition to that, we can provide metadata relevant to the localization process that may be of great interest when relying on the automatic translation of ontologies.

As future work we plan to carry out some experiments to provide statistics on the impact of such translation relations in the Multilingual Semantic Web, specifically the distinction between literal translations and cultural equivalences. We also aim at investigating the implementation of algorithms that would automatize this process.

Acknowledgments. This work is supported by the EU project Monnet (FP7-248458), and by the Spanish national project BabeLData (TIN2010-17550).

References

1. Bizer, C., Heath, T., Berbers-Lee, T.: Linked data - the story so far. International Journal on Semantic Web and Information Systems. 5, 1–22 (2009)
2. Bizer, C., Heath, T., Idehen, K., Berners-Lee, T.: Linked data on the web (LDOW2008). In: 17th international conference on World Wide Web, pp. 1265–1266 (2008)
3. Bontcheva, K.; The Semantic Web: Research and Applications. In: Generating tailored textual summaries from ontologies Springer, pp. 531–545 (2005)
4. Buitelaar, P.: Ontology-based Semantic Lexicons: Mapping between Terms and Object Descriptions. In: Ontology and the Lexicon, pp. 212–223. Cambridge University Press (2010)
5. Buitelaar, P., Cimiano, P., Haase, P., Sintek, M.: Towards Linguistically Grounded Ontologies In: 6th European Semantic Web Conference (ESWC09), pp. 111–125 (2009)
6. Cimiano, P., Montiel-Ponsoda, E., Buitelaar, P., Espinoza, M., Gómez-Pérez, A.: A Note on Ontology Localization. Journal of Applied Ontology, 5(2), pp. 127–137 (2010)
7. Declerck, T., Krieger, H.-U., Thomas, S. M., Buitelaar, P., ORiain, S., Wunner, T., Maguet, G., McCrae, J., Spohr, D., Montiel-Ponsoda, E.: Ontology-based Multilingual Access to Financial Reports for Sharing Business Knowledge across Europe. In Rooz, J., Ivanyos, J. (eds.) Internal Financial Control Assessment Applying Multilingual Ontology Framework, HVG Press Kft, pp. 67-76 (2010)
8. Espinoza, M., Montiel-Ponsoda, E., Gómez-Pérez, A.: Ontology Localization. In: 5th International Conference on Knowledge Capture (KCAP09), pp.33–40 (2009)
9. Espinoza, M., Gómez-Pérez, A., Mena, E.: Enriching an Ontology with Multilingual Information. In: 5th Annual of the European Semantic Web Conference (ESWC08), pp. 333–347 (2008)
10. Hirst, G.: Ontology and the Lexicon. In: Staab, S., Studer, R. (eds.) Handbook on Ontologies, International Handbooks on Information Systems, Springer, pp. 209–230 (2004)
11. House, J.: A Model for Translation Quality Assessment, Narr, (1977)
12. Kemps-Snijders M., Windhouwer M., Wittenburg P., Wright S.: ISOcat: Corralling data categories in the wild. In: International Conference on Language Resource and Evaluation (LREC) (2008)
13. Manola, F., Miller, E.: RDF Primer. Technical report, W3C Recommendation World Wide Web Consortium (W3C) (2004)
14. McCrae, J., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., Wunner, T.: Interchanging Lexical Resources in the Semantic Web. Language Resources and Evaluation, in press (2011)
15. Miles, A., Bechhofer, S.: SKOS-Simple Knowledge Organization System Reference, W3C, Retrieved April 11, 2011, from http://www.w3.org/TR/skos-reference/ (2009)
16. Miles, A., Bechhofer, S.: SKOS Simple Knowledge Organization System eXtension for Labels (SKOS-XL) Namespace Document - HTML Variant , W3C, Retrieved June 21, 2011, from http://www.w3.org/TR/skos-reference/skos-xl.html (2009)
17. Montiel-Ponsoda, E.: Multilingualism in Ontologies. Building Patterns and Representation Models. LAP LAMBERT Academic Publishing, Germany (2011)

18. Montiel-Ponsoda, E., Aguado de Cea, G., Gómez-Pérez, A., Peters, W.: Enriching Ontologies with Multilingual Information. Journal of Natural Language Engineering, 17 (3), 283–309 (2010)
19. Müller, H.-M., Kenny, E. E., Sternberg, P. W.: Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature. PLoS Biol, 2, e309 (2004)
20. Nord, Ch.: Loyalitt statt Treue. Lebende Sprachen, 34, pp. 100–105 (1989)
21. Surez-Figueroa, M. C., Gómez-Pérez, A.: Towards a Glossary of Activities in the Ontology Engineering Field In: 6th Language Resources and Evaluation Conference (LREC08) (2008)
22. Svab-Zamazal, O., Svatek, V.: Analysing Ontological Structures through Name Pattern Tracking. In: EKAW 2008 - 16th International Conference on Knowledge Engineering and Knowledge Management, pp. 213–228 (2008)
23. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Silk A Link Discovery Framework for the Web of Data. In: 2nd Workshop about Linked Data on the Web (LDOW2009) (2009)

A Semantic Model for Integrated Content Management, Localisation and Language Technology Processing

Dominic Jones[1], Alexander O'Connor[1], Yalemisew M. Abgaz[2], David Lewis[1]

[1 & 2] Centre for Next Generation Localisation
[1]Knowledge and Data Engineering Group,
[1]School of Computer Science and Statistics, Trinity College Dublin, Ireland
{Dominic.Jones, Alex.OConnor, Dave.Lewis}@scss.tcd.ie

[2]School of Computing, Dublin City University, Dublin, Ireland
[2] Yabgaz@computing.dcu.ie

Abstract. Providers of products and services are faced with the dual challenge of supporting the languages and individual needs of the global customer while also accommodating the increasing relevance of user-generated content. As a result, the content and localisation industries must now evolve rapidly from manually processing predicable content which arrives in large jobs to the highly automated processing of streams of fast moving, heterogeneous and unpredictable content. This requires a new generation of digital content management technologies that combine the agile flow of content from developers to localisers and consumers with the data-driven language technologies needed to handle the volume of content required to feed the demands of global markets. Data-driven technologies such as statistical machine translation, cross-lingual information retrieval, sentiment analysis and automatic speech recognition, all rely on high quality training content, which in turn must be continually harvested based on the human quality judgments made across the end-to-end content processing flow. This paper presents the motivation, approach and initial semantic models of a collection of research demonstrators where they represent a part of, or a step towards, documenting in a semantic model the multi-lingual semantic web.

Keywords: Multilingual Web, Content Management, Localisation, Language Technology and Interoperability

1. Introduction

To engage successfully with global markets, enterprises increasingly need to manage fast moving streams of multi-lingual content from both within the enterprise and from the wider user communities with which they wish to engage. For example, modern software markets are increasingly dominated by larger numbers of fine-grained applications, e.g. smartphone and social network 'apps' or Software-as-a-Service (SaaS) offerings, that feature high frequency/ "perpetual beta" release cycles. For these products, technical documentation increasingly adopts the form of FAQs,

blogs and wikis that grow with significant input from users and customer support staff as features and problems are discovered and solutions documented. Even when technical manuals are provided (and localised into different target markets), users increasingly find more direct solutions to problems on user-generated question & answer sites, which may then themselves, merit localisation based on demand.

Managing this multilingual content stream requires seamless integration of content management systems, the localisation chain and data-driven natural language technologies. Cost, timeliness and quality trade-offs need to be actively managed for different content flows with a far greater level of flexibility and automation than that that has been achieved previously to cover the range of multilingual content generation and consumption paths, e.g. from online documentation, to Q-A fora and micro-blog and RSS feeds.

The Centre for Next Generation Localisation (CNGL) specifically consists of over 100 researchers from academia and industry who conduct research into the integration of natural language technologies, localisation and digital content management required in addressing these challenges. Researchers work collaboratively in CNGL to produce a range of technical demonstrators that integrate multiple forms of multilingual content. A key challenge associated with such large-scale research systems integration is the need for researchers to collaborate and for software to interoperate. However, the components are derived from a number of different research and industrial communities, where either meta-data was not formally defined or was specified from a fragmented set of standards or industry specifications. CNGL therefore established a meta-data group (MDG) to concentrate and integrate the meta-data expertise from these different research areas, including statistical machine translation and text analytics research, adaptive content and personalisation research and localisation workflow and interoperability. To address the universal trend of content to be web based and to offer a well-supported, community-neutral approach to semantic modelling, the standardised languages of the W3C Semantic Web initiative were used. This allowed multiple existing meta-data standards and component meta-data requirements to be incorporated into a single model thereby demonstrating the interrelation and utility of such interlinked meta-data. This paper presents the initial process for the development of such semantic models for interoperation within a large software research project, such as the CNGL. Future research will see these models are deployed and evaluated in multiple industrial settings to establish their broader applicability in the field of localisation and digital content management. This paper only provides a discussion around forming these initial models.

2. Approach

There is an increasing focus on the interoperability of Content Management Systems (CMS), the tools in the localisation chains and the emerging array of language technologies offered as services. In the localisation industry there are many different platforms, favoured by various Language Service Providers (LSPs) for translation projects. Some are in-house deployments, some open-source (such as GlobalSight) and some purchasable from third parties (such as SDL WorldServer). However each platform comes with its own nuances, meta-data specification and

translation workflow management tools. For example within Digital Content Management multiple approaches to storing meta-data apply ranging from storing simple XML content as mark-up (from both authors and consumers) through to complete Ontological models. In Machine Translation meta-data appears in the many forms from terminological mark-up, workflow management meta-data and Translation Memory (TM) that allows previous translations to be recycled in the Machine Translation (MT) process. As has been shown complex meta-data standards occur in many places across both the multi-lingual semantic web and localisation industry.

CNGL develops a large number of distinct "demo" systems integrating different aspects of content management, localization and language technology integration.

Figure 1 summarises how the current range of integrated systems in these categories map onto the NGL Process Map. Semantic models, consisting of content and service models are developed by synthesizing a seed model from an analysis conducted across the range of these integrated systems. This model has then been mapped back onto revisions of the integrated system to assess the degree to which the model can accommodate the natural evolution of these systems as they evolve in response to specific industry requirements and technological advances.

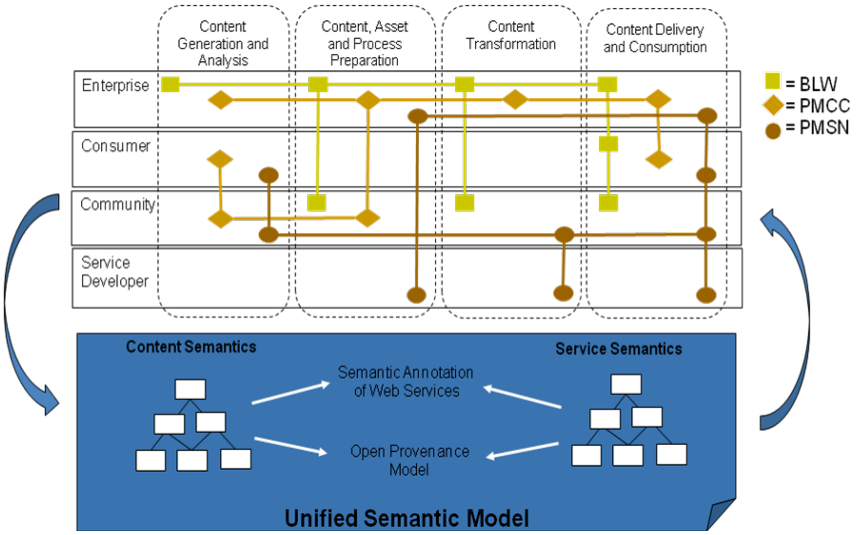


Figure 1: NGL Process Map used for Organizing Semantic Modeling

The availability of such a large, motivated cohort of collaborating researchers presented a range of distinct views and expertise and allowed us to gather input for and define common models of both the content and services offered across Localisation, Content-management and Integrated Language Technologies. The demonstration systems used address use scenarios in the areas of: Bulk Localisation Workflow (BLW); Personalised Multilingual Customer Care (PMCC) and Personalised Multilingual Social Networking (PMSN). To develop a shared understanding of this modelling, demonstration systems workflows and the components implementing specific activities are overlaid onto a Next Generation

Localisation (NGL) Process Map (Figure 1) that identified different stakeholders and abstractions of the major process areas involved in content generation, its preparation for manual and machine processing, the translation, such as translation and personalisation, resulting from this processing and the delivery and consumption (including rating and feedback) by consumers. Semantic models, the content being processed and the services offered by the components doing these processing were then developed by synthesizing a seed model from an analysis conducted across the range of integrated systems and making continual iterations to this model.

However, the breadth of input also put into sharp focus the degree of change these models would be subject to, resulting from continual advances in the research components used, we therefore applied Ontological Evolution techniques to create a common evolving model that is representative of this set of changing viewpoints and that is extensible for use by others in the future. This semantic model has therefore been mapped back onto revisions of the integrated demonstration systems to assess the degree to which the model can accommodate the natural evolution of these systems (which have undergone four coordinated iterations over the last three years) as they change in response to specific industry requirements and technological advances.

3. Challenges for Multilingual Web Meta-data Interoperability

The purpose of building common semantic models is driven by the fact that standardisation efforts in support of content management and localisation processes and tool interoperability are somewhat fragmented between the Organization for the Advancement of Structured Information Standards (OASIS), the Open Standards for Container/Content Allowing Re-use (OSCAR) operated by the (now defunct) Localisation Industry Standards Association (LISA) and the W3C. Most current standards are in the form of XML document schema designed to support hand-over of specific content and meta-data between different tools through import and export functions using these different file formats. This occurs primarily when a translation job is handed off from a content developer to an Language Service Provider (LSP) via their respective Translation Management System (TMS) tools and onwards to an individual translator's CAT tool, possibly via a translation agency [1].

For example LISA's Translation Memory Exchange (TMX) [15] standard is one of the most widely used and supports the hand-off of the key linguistic assets currently used in the translation process. This is typically from Translation Memory (TM) repositories to Computer Aided Translation (CAT) tools, where content is leveraged, and back again as quality-check translations that are added to the TM. Similarly, LISA defined Term Base Exchange (TBX) [16] which is an XML format for exchanging terminological information, aligned with description categories defined by ISO/TC37 on Terminology and other language and content resources.

The XML Localization Interchange File Format (XLIFF) [14] from OASIS offers a standard way of passing a live translation job (combined in an envelope of content and meta-data between different tools) handling: the extraction and segmentation of translatable text; its translation by human, TM leverage or Machine Translation (MT); translation review and re-integration of target language content with formatting

skeleton files for publication. The XLIFF schema suffers however from a high level of optional elements as well as common use of user-generated extensions resulting in difficulties in its use for blind interchange of localisation meta-data between tools [2]. In addition to this XLIFF does not support the passage of process meta-data from either content authoring processes (including the application of controlled language) to the localisation process, nor from the localisation process to the TM and terminology maintenance processes, so that meta-data potentially useful for future TM leverage or TM and terminology use in Statistical Machine Translation (SMT) training is lost.

Another OASIS TC, on an Open Architecture for XML Authoring and Localization (OAXAL) [18], has articulated a reference model that proposes using the XML Text Memory schema defined by LISA (xml:tm) [19] to provide better end-to-end management of localised content. By managing the segmentation of source content within the content management system, it provides both author memory to assist changes in the source termed 'translation memory', and maintains the link between source and target content such that changes to either can be managed and also used to more effectively build new TMs. However, only a high-level integration framework has been outlined to date and considerable additional work toward alignment with other standards would be required for a workable solution. In summary the challenges faced in supporting integration of content processing across the content management and localisation industries are seen as:

- The tendency towards tool vendor lock-in, in terms of both Translation Management Systems and exchange formation, and the associated lack of a strong independent tool vendor market supported by widely adopted common standards.
- The need to migrate localisation processes to integrate with increasingly web based, multi-media and multi-modal content that can be personalised to individual user and community needs.
- The need to apply localisation and personalisation processes to user generated content and live social media while also actively leveraging the wisdom of the crowd in accelerating those processes. More and more content is originating from rapidly changing sources of content, involving the user in the translation and post-editing of such content is beneficial to the user (being provided access to content in their native tongue) and to the producer in terms of reduced costs associated with translation. Where reduced costs, increased speed, free marketing and delivering of quality assurance from the users themselves form a new approach for companies to translate content.
- The lack of interoperability standards that span the Next Generation Localisation (NGL) problem space and the lack of maturity of many of the standards that are presented in various sub-domains.

In order to begin to address these challenges integration in CNGL is based around the adoption of Service Oriented Architectures (SOA). However, the wide and diverse scope of CNGL requires strong common models so that meta-data can be easily shared and services integrated without expensive mismatches in assumptions about data and state and with clear processing expectations for the use of services.

Although localisation tool vendors are already attempting to support interoperability through exposing web service APIs to the functionality they offer, these interfaces are still largely proprietary, though there have been attempts to use XLIFF and TMX file format as input/output message payloads. Service-oriented integration of language technology has received some attention also by the research community, including: the Japanese-funded LanguageGrid project that provided a platform for the integration of arbitrary compositions in language resources [3], the CNGL which is applying service oriented techniques to integrating language and digital content management technologies with localisation workflows [4] and more recently the Panacea FP7 [5] project which is investigating service compositions for language resource processing pipelines.

However the document-centric nature of existing localisation standards means that their use in web service interfaces for localisation has been ad hoc and lacked any shared conceptual or common conformance framework needed to support seamless integration. In contrast, our semantic model forms a shared conceptual view of services and content that has been derived and validated through a set of software integrations. By providing a core common data type model with a well defined conceptual basis for service, developers can define their interfaces more precisely as contracts, with an easily reached shared understanding of precondition states and processing expectation involved in invoking a given service. At the same time, however, we do not aim to supplant existing interoperability standards where they exist, nor do we intend to be a source of new interoperability standards. Instead we aim to provide a minimal common model for data types that can be exchanged via NGL services, while leaving it to individual service developers to use this type set to define their service interfaces in their particular exchange format.

4. Semantic Model

As the common focus of integration was the processing of digital content, a content-based semantic model was developed that incorporated a content processing service model **Figure 2** and a processed (i.e. managed) content data model Figure 3. Both of these models have been developed using the Resource Description Framework (RDF) [6] and ontological principles from the Semantic Web to provide:

- Flexibility in defining types.
- Extensibility through class specialisation.
- Operational persistence with explicit meta-data in the form of triple stores.
- Future support for open linked data approaches to content processing.

This approach has provided a more flexible and extensible mechanism for defining data types that can be exchanged between what we expect to be a large and dynamic set of NGL services. In addition the semantic models also help to simplify the design and refinement of content processing workflows in the CNGL research space by interlinking concepts from existing models and existing standards including Content management including DocBook [12], DITA[13], HTML, XML and Localisation standards including: XLIFF [14], TMX [15], TBX [16], LCX [17], OAXAL [18] and

xml:tm [19]. The problem with such standards is that they often are deployed or utilised in separate deployments or installations crossing over one another without a common theme to their use where different platforms utilise different standards making their interoperation hard to achieve. In the CNGL the aim is to make the semantic model open, since to reduce industry interoperability costs, a common model must be widely adopted by other system integrators and tool vendors and this model must grow and evolve to adapt as new opportunities are exposed by third party usage and their resulting feedback. We aim to improve the quality of the semantic model by treating CNGL's broad range of demos as a unique Interoperability Laboratory providing, revising and reviewing semantic models sourced from a seed model and synthesised from an analysis of early iterations of integrated systems. Future research will see this model deployed in an industrial setting where the users of the localisation standards presented can evaluate how well such a common model applies in their particular workflow.

The seed Semantic Model is broken down into two core parts, these are: **Service:** High level classifications of the different services covered based on their content processing features **and Content:** The core processed content model that we use to record the content transformations of various types (including use of content as MT training data) delivered by either activities from the business model or services. These models provide an open mechanism for defining common meta-data from existing software development, modelling tools and persistent data-stores. It therefore offers a practical mechanism for defining a minimal core set of data types that can then be composed or extended as new services are defined and new application requirements arise. A benefit of such an approach towards semantic modelling is that it supports the development of content processing management logs, which due to the emerging nature of the services and applications involved, must be more exploratory in nature, while operational, configuration and tuning requirements are established for data-driven NLP technologies.

During the process of model evolution a parent-child relationship is formed between classes of content, so for example a sub-class of "GenerateContent" is "AuthorText" following a more specific or less specific formation of relationships. Properties are assigned to classes as data-type relationships for example "AuthorText" may have a property "WordCount = int" where users are free in their adaptation of the models to add as many sub-classes, super-classes or class structure changes as they wish. In the Content Model users are able to add as many properties as they wish. All of the changes to all of the models are integrated in one single collaborative session where the meta-data group debates, integrates and records the changes made to the initial seed models. The versions presented in this paper are the initial seed models of both content and services with version 1 soon to be released.

4.1. Service Model

The Service Semantic model, shown in **Figure 2**, is based around the assumption that all services or process, create or consume content in some form or another. As RDF allows multiple inheritances defining new class types, the schema is defined to allow integration of fundamental aspects of content processing services to define a wide range of services. The core upper level service types include:

- **GenerateContent**: the creation of content by human users.
- **TransformContent**: the transformation of content from one human understandable form to another, including translation, text to speech and content personalised or adaption for delivery to a particular user.
- **AnnotateContent**: where additional meta-data is associated with content.
- **ProcessGroupContent**: where operations on **sets** of content are represented.
- **CreateService**: allowing the creation of a service to result from a processing chain of other services, therefore allowing the configuration and training of an SMT engine or other data-driven components, or adaptive composition of services to be captured.

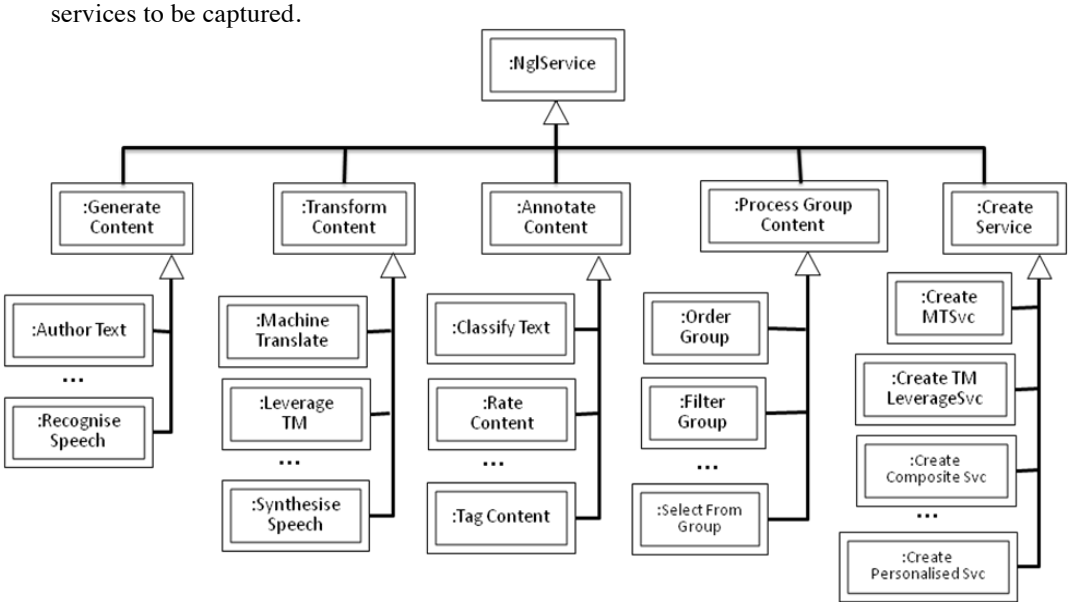


Figure 2: Current "Seed" Upper Layer Taxonomy of Service Types

4.2. Content Model

Similar to the Service Model, the Content Model, shown in Figure 3, aims to provide an extensible fine-grained schema of types from which the input and output of services that can be modelled. The core type in the content model is ManagedContent, which indicates that we are only interested in content that is subject to some form of management or monitoring. The following key content processing content types are then identified:

- **GeneratedContent**: content produced by a person.
- **AnalysedContent**: content that has been analysed prior to making some decisions on further processing, such as user adaption.
- **PreparedContent**: content that has been altered for further processing.
- **LocalisedContent**: content that has been subject to the localisation processes.

Other seed subclasses have been identified to differentiate content that has been personalised, published or presented to a user or been discovered via information

retrieval. Further orthogonal subclasses differentiate: the manual and automatic processing of content, content that is managed as an asset, e.g. a linguistic resource, as well as utility types indicating the content is grouped, time-stamped and serialised as a file, has had its elements counted or some intellectual property right asserted over it. In this seed structure it is important to note that "PreparedForLocalisation" and "PreparedForAdapation" are different in the following way: Localised content is content which is translated and personalised for delivery to a user, Adapted content is only content which is personalised and may not specifically have been translated before delivery to a user. Also important to note is that in terms of both the content and service models these change as the crowd of users adapt them to include the services and content types offered through their research. For example the content model does not include any user generated or live social media content types, however once deployed and adapted by users using a crowd-sourced approach such additions are deemed more likely.

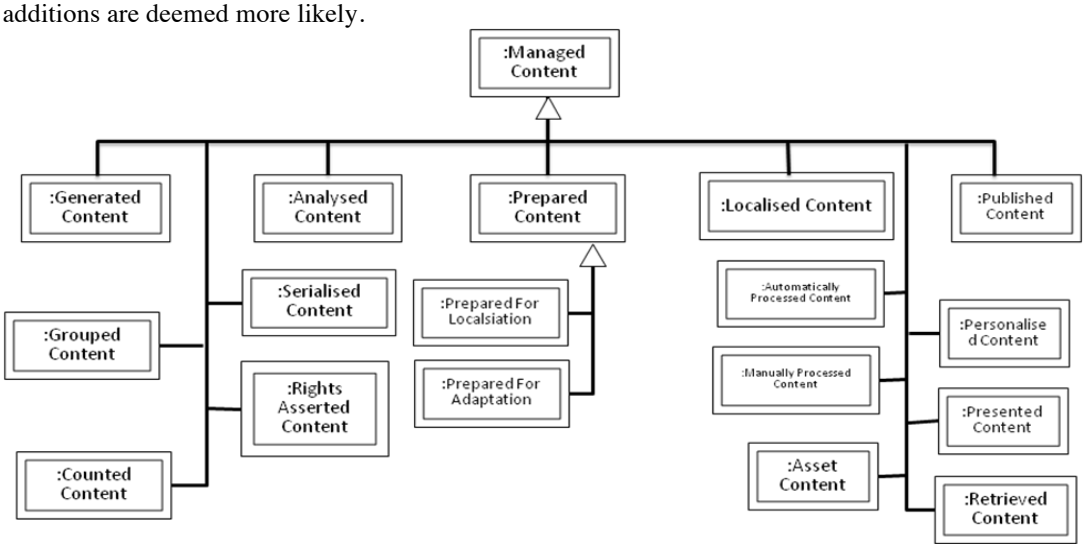


Figure 3: Current "Seed" Upper Layer Taxonomy of Content Types

5. Ontology and Semantic Evolution

The models presented here have been captured as an ontology using RDF. An Ontology or semantic model is a specification of a shared conceptualization of a domain [7] and serve to capture domain knowledge in a generic way providing a commonly agreed understanding and conceptualization of knowledge elements within that domain. They provide common ground for understanding, conceptualization, representation and interpretation of domain concepts uniformly across different systems, languages and formats. However, as discussed above, the dynamics of the given ontology often demand changes to application requirements that may be fulfilled only by changing/evolving these principal ontologies [8].

In [9 and 10] ontology evolution is defined as: “The timely adaptation of an ontology to changed patterns of usage in the ontology-based application, as well as the consistent management and propagation of these changes to dependent elements.” Ontology evolution takes place when an ontology management system facilitates the modification of an ontology by applying the proposed changes to the model and by ensuring its subsequent consistency. The need for ontology evolution is directly related to the changes that occur in the domain area or in the business environment and reasons for change include conceptualization, representation or specification of the domain. State-of-the-art ontology evolution process has six defined phases [8,9,10]. The first phase, “change capture”, focuses on investigation and capturing changes to the domain such as new concepts, out-dated concepts or updated concepts. The second phase is “change representation”. Captured changes are represented using atomic or composite change operations. The “semantics of change” phase deals with the effects of the changes that have been requested, checking the effects of the changes if implemented in the ontology [11]. The main task of the next phase, “change propagation”, is to confirm that dependent ontologies and tools are still functioning properly and that no inconsistencies are introduced by changes to the ontology. The “change implementation” phase focuses on implementing changes to the ontology and keeping a record of these changes for undo or redo purposes. Finally the change validation phase validates the change and makes it publicly available after resolving inconsistencies if there are any.

In a collaborative environment such as CNGL, achieving a common semantic model requires a continuous evolution of models. Users expect the model to support their changing requirements and thus leave the model in a continuous evolution. Researchers applying their technical demonstrators to the current model of services and content adapt the models presented in this paper and these changes are incorporated into the seed models producing iterative versions of common, CNGL wide, agreed semantic models, the first full versions soon to be published.

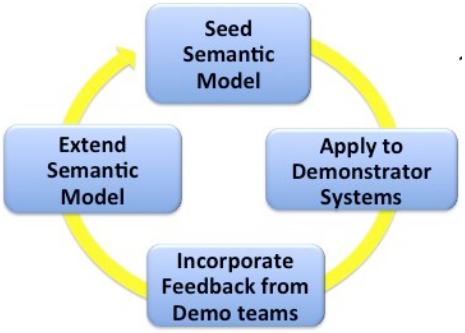


Figure 4: Approach taken to developing semantic models

The approach taken, shown in Figure 4, to populating the semantic models, from the collection of demonstrator systems, across the CNGL, involved producing a seed semantic model applying the model to each individual demo team. Incorporating feedback from the collection of demo teams into the seed models and extending the semantic model through a process of continual evolution using a panel of domain

experts (the meta-data group) to enforce chances. In this approach, the need for evolving the semantic model comes directly from the demonstrators.

In the process of building the model, new services and content types emerged from the demo teams. To incorporate these new services and content types, it is essential for the models to evolve. This evolution occurs whilst incorporating feedback from seventeen demo teams through an interview approach to data collection. A total of 43 change requests were made to the NGL service model (12 changes) and NGL content model (31 changes). These requests included the addition of new content types which were not captured in the seed model, specialization of existing concepts into two or more types, generalization of categories to a single type and renaming of existing types. There were no requests made for deletion of a given content or service type, however in the process of implementing the above changes, for example renaming, there were deletions introduced. In addition to working on the semantic models of services and content properties were incorporated into the content model 44 of which were identified. Some example of additions made in populating the models include:

Requested content changes:

- Add subclass (user feedback, managed content)
- Add subclass(user opinion, user feedback)
- Add subclass(user input, user feedback)

Requested service changes:

- Add subclass(rank TM, process group content)
- Add subclass(compare content, process Group content)
- Add subclass(classify audio, annotate content)

Requested properties:

- Add Property (source Language)
- Add Property (input Language)
- Add property (number Of Sentences)

6. Conclusions

Localization is moving more and more towards Content Management in terms of the process pipeline apparent in taking a piece of content and turning it from source language to destination. As the breadth and depth of the content being localized increases there becomes a more apparent need to move towards a common semantic model of the content being processed. This model needs to be both human understandable but also normative, and for this reason it is argued that RDF is a suitable candidate for building models of a rapidly expanding content set. This paper has presented an approach currently being utilized for building common semantic models of the services and content types from a number of demonstrator systems focusing on localization, translation, natural language processing technologies and digital content management. It is argued that such a collaborative approach to gathering and defining semantic models of services and content provision is vital in the future of the multi-lingual semantic web as it attempts to span the different related research and industrial communities.

7. Acknowledgements

In acknowledgement this research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Trinity College Dublin and Dublin City University.

8. References

1. Gough, J. "A troubled relationship: the compatibility of CAT tools", TAUS technology article, downloaded from http://www.translationautomation.com/technology/a-troubled-relationship-the-compatibility-of-cat-tools.html 28 Dec (2010)
2. Mateos, J " A Web Services Approach to Software Localisation. Bringing Software Localisation Tools from the Desktop to the Cloud", Trinity College Dublin, Computer Science Technical Report TCD-CS-2010-25 (MSc Dissertation), 20 October 2010, https://www.scss.tcd.ie/publications/tech-reports/reports.10/TCD-CS-2010-25.pdf -01/11
3. Inaba, R., Murakami, Y., Nadamoto, A., Ishida T. "Multilingual Communication Support Using the Language Grid" Intercultural Collaboration, LNCS 4568, Springer, Aug 2007
4. David Lewis, Stephen Curran, Dominic Jones, John Moran, Kevin Feeney (2010). An Open Service Framework for Next Generation Localisation. Web Services and Processing Pipelines in HLT: Tool Evaluation, LR Production and Validation, Malta, May 2010.
5. A. Toral, "Towards Using Web-Crawled Data for Domain Adaptation in Statistical Machine Translation." (2011)
6. O. Lassila and R. R. Swick, "Resource description framework (RDF) model and syntax," *World Wide Web Consortium, http://www.w3.org/TR/WD-rdf-syntax*
7. Gruber, T.R.: A translation approach to portable ontology specifications. Knowledge. Acquisition. 5(2) (1993) 199–220
8. A., Stojanovic, N., Studer, R.: Ontology evolution as reconfiguration design problem solving. Proceedings of the 2nd international conference on Knowledge capture (2003)
9. Zablith, F.: Dynamic ontology evolution. International Semantic Web Conference (ISWC) Doctoral Consortium, Karlsruhe, Germany (2008)
10. Stojanovic, L., Maedche, A., Motik, B., Stojanovic, N.: User-driven ontology evolution management. Lecture Notes in Computer Science. 6(4) (2002) 285–300
11. Abgaz, Y., Javed, M., Pahl, C.: Empirical analysis of impacts of instance-driven changes in ontologies. In: On the Move to Meaningful Internet Systems: OTM 2010.
12. DockBook, http://www.docbook.org/
13. Darwin Information Typing Architecture (DITA) http://dita.xml.org/ Accessed - 09/11
14. XML Localisation Interchange File Format (XLIFF) http://www.oasis-open.org/committees/xliff/ Accessed - 09/11
15. Translation Memory eXchange (TMX) http://www.lisa.org/fileadminstandards/tmx1.4/tmx.htm Accessed - 09/11
16. TermBase eXchange (TBX) http://www.ttt.org/tbx/ Accessed - 09/11
17. Localisation Content Exchange File Format (LCX) http://www.localisation.ie/xliff/resources/presentations/lcx-xliff.pdf Accessed - 09/11
18. Open Architecture for XML Authoring and Localization (OAXAL) http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=oaxal Accessed - 09/11
19. XML-based Text Memory (Xml:tm) http://www.infomanagementcenter.com/enewsletter/200608/fifth.htm Accessed - 09/11

ULiS: An Expert System on Linguistics to Support Multilingual Management of Interlingual Semantic Web Knowledge bases

Maxime Lefrançois, Fabien Gandon

EPI Edelweiss – INRIA Sophia Antipolis
2004 rt des Lucioles, BP93, Sophia Antipolis, 06902, France
Maxime.Lefrancois@inria.fr | Fabien.Gandon@inria.fr

Abstract. We are interested in bridging the world of natural language and the world of the semantic web in particular to support multilingual access to the web of data. In this paper we introduce the ULiS project, that aims at designing a pivot-based NLP technique called *Universal Linguistic System*, 100% using the semantic web formalisms, and being compliant with the Meaning-Text theory. Through the ULiS, a user could interact with an *interlingual knowledge base* (IKB) in controlled natural language. Linguistic resources themselves are part of a specific IK*B: The Universal Lexical Knowledge base* (ULK), so that actors may enhance their controlled natural language, through requests in controlled natural language. We describe a basic interaction scenario at the system level, and provide an overview of the architecture of ULiS. We then introduce the core of the ULiS: the *interlingual lexical ontology* (ILexicOn), in which each interlingual lexical unit class (ILU^c) supports the projection of its semantic decomposition on itself. We validate our model with a standalone ILexicOn, and introduce and explain a concise human-readable notation for it.

Keywords. Semantic Web; Explanatory Combinatorial Lexicology; Interlingual Lexical Ontology; Semantic decomposition; Interlingual Lexical Primitives, Meaning Text Theory.

1 Introduction

In this paper we introduce and illustrate the recently begun ULiS project, which aims at redesigning a pivot-based NLP technique, 100% using the semantic web formalisms, and being compliant with the Meaning-Text theory. ULiS stands for *Universal Linguistic System*, and is a system through which multiple actors could interact with *interlingual semantic web knowledge bases* in multiple controlled (i.e., restricted and formal) natural languages. Each controlled natural language (dictionary, grammar rules) would be described in a part of a *universal linguistic knowledge base* (ULK). Besides this, the ULK consists in one specific interlingual knowledge base. Actors could then enhance their controlled natural language through different actions in controlled natural language (e.g., create, describe, modify, merge, or delete lexical units in the dictionaries and grammar rules; connect situational lexical units to interlingual lexical units; add linguistic attributes with their associated rules, etc.).

The aim of this paper is to overview our proposal for the architecture of ULiS, and to introduce and validate the cornerstone of the universal linguistic knowledge base: the *interlingual lexical ontology* (ILexicOn).

2 Related Work

The Meaning-Text Theory (MTT). The MTT is a theoretical linguistic framework for the construction of models of natural language. As such, its goal is to write systems of explicit rules that express the correspondence between meanings and texts (or sounds) in various languages (Kahane, 2003). Seven different levels of linguistic representation are supposed for each set of synonymous utterances: a semantic representation that is a network; the deep and surface syntactic representations (DSynR and SSynR) that are trees; the deep and surface morphological representations (DMorphR and SMorphR) that are lists of annotated tokens; and the the deep and surface phonological representations (DPhonR and SPhonR) that are also lists of annotated tokens. (Mel'čuk, 1998).

Thus, twelve modules containing transformation rules are used to transcribe representations of a level into representations of an adjacent level. The main constituent of the MTT is the dictionary model where lexical units are described, which is called the *Explanatory Combinatorial Dictionary* (ECD), and has been the object of many works on lexical functions, e.g., (Mel'čuk et. al., 1995).

Lexical ontologies and meaning representation languages. Lexical ontologies are ontologies of lexicalized concepts, widely used to model lexical semantics. Some have broad coverage but shallow treatment (i.e., with no or little axiomatization) such as Princeton WordNet (e.g., Miller et al., 1990), and some have small coverage but are highly axiomatized such as FrameNet (Baker et al. 1998). They use different theories of lexical semantics but most of them do not describe phrasemes nor lexical collocations. The French Lexical Network (Lux-Pogodalla & Polguère, 2011) is a growing ECD-compliant lexical resource, but it does not use the semantic web formalisms, and the definitions of the lexical units are not fully formalized.

On the other hand, the Universal Networking Language (UNL) is a meaning representation language, originally designed for pivot techniques Machine Translation. Its dictionary is an interlingual lexical ontology based on so-called Universal Words ++, but the lack of argument frames and lexical functions in the UNL dictionary was pointed out in (Bogulsavsky, 2002; Bogulsavsky, 2005). This is when the idea of an ECD-compliant interlingual lexical ontology was first mentioned. After the semantic web formalisms were introduced at the W3C, an attempt to port the UNL to semantic web formalisms was the topic of the W3C Common Web Language Incubator Group (XGR-CWL, 2008), but no improvement was made to the lexical ontology.

SPARQL Inferencing Notation (SPIN). Grammar rules are not part of the Common Web Language (CWL) framework, in fact, the construction of grammar modules may

be done in any programming language. Knublauch et. al. (2011) introduced SPIN: an RDFS schema to represent SPARQL rules and constraints.

Positioning of the ULiS project. The lexical resource we propose to develop is an interlingual lexical ontology coupled with a situational (i.e., a generalization of language-specific) lexical ontology, both using semantic web formalisms, and that together form an ECD-compliant dictionary. Benefits of using semantic web formalisms are high as it enables us to construct an axiomatized graph-representation of a lexical ontology, with validation and inference rules. Using SPIN, we propose to include transformation rules directly in an RDF format, on top of the ECD-compliant lexical ontologies, thus obtaining an expert system on linguistics.

The ULiS model is somehow similar to the FunGramKB (Periñán-Pascual & Arcas-Túnez, 2010) which is a lexico-conceptual knowledge base for NLP. However, the two projects have different inspiring influence. We choose to comply with the Meaning-Text theory, which gives a thorough understanding of lexical functions that are ubiquitous in every natural language. We also choose to describe the whole ULiS with the semantic web formalisms. This thus potentially enables the enhancement of the system itself through controlled natural language interactions.

3 Basic Interaction Scenarios with the ULiS

The three basic scenarios of ULiS are illustrated on Figure 1 below.

An actor in a situation *c* inputs some utterance (e.g., in English: "Who killed Mary?") that is first transformed into an RDF situational representation, which undergoes different language-specific process, and which is finally transformed into a CWL-like interlingual representation.

Machine translation. At this stage, depending on the context, the interlingual representation of the utterance may be translated into another utterance in situation *d* (e.g., in the French situation: "Qui a tué Mary?") through a situational representation (Output1^{TEXT} on Figure 1.

Management of Interlingual Knowledge Bases. Another possibility is that the interlingual representation of the utterance is transformed in a SPARQL request that is applied on an *interlingual knowledge base* (IKB), which eventually produces an RDF output (e.g., `ex:John01`). This RDF output is then first transformed into an interlingual representation, then into a situational representation and finally into an output utterance: Output2^{TEXT} on Figure 1 (e.g., "John killed Mary").

Management of the Universal Linguistic Knowledge base. Finally, the third scenario is the human-computing scenario: the SPARQL request is applied on the Universal Linguistic Knowledge base, which is the Interlingual Knowledge Base where the

whole ULiS is described. Human actors may thus enhance the controlled natural languages through actions stated in controlled natural language.

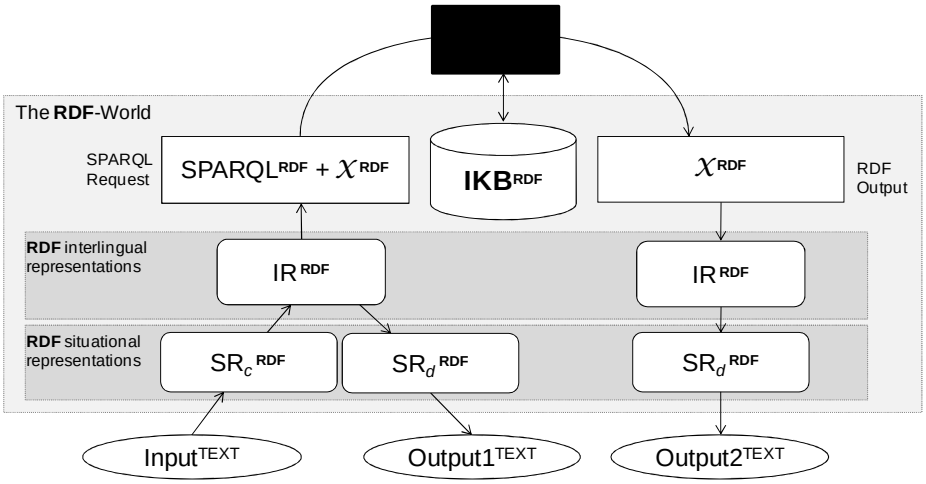


Fig. 1. ULiS: The basic interaction scenario with an interlingual knowledge base.

Thus the interlingual representation format acts as a pivot not only for natural languages, but any interlingual representation may be translated into a SPARQL request, and any RDF graph may be translated to an interlingual representation.

4 The ULiS components

4.1 Overview

Figure 2 below illustrates the ULiS, with its three different layers:

The second row represents interlingual layer (section 4.2), with a meta-ontology that describes the *interlingual lexical ontology* (ILexicOn): the cornerstone of the whole *Universal Linguistic Knowledge base*. The ILexicOn enables inference in *interlingual semantic representations* (ISemRs, on the right).

The first row represents the *interlingual knowledge base* (IKB) layer, with facts (on the right) and an ontology or thesaurus (on the left), augmented with anchors and transformation rules (section 4.4), that enable the transformation of facts into ISemRs, and vice versa. The IKB enables situation-independent inference on utterance representation.

The third row represents the situational layer (section 4.3), with a meta-ontology that describes the *situational lexical ontology* (SLexicOn), that itself enables situation-dependent linguistic inference on utterances' situation-dependent representations (*Situational representations*, SRs, on the right). Situation-annotated links and transformation rules define transformation of utterances among SRs.

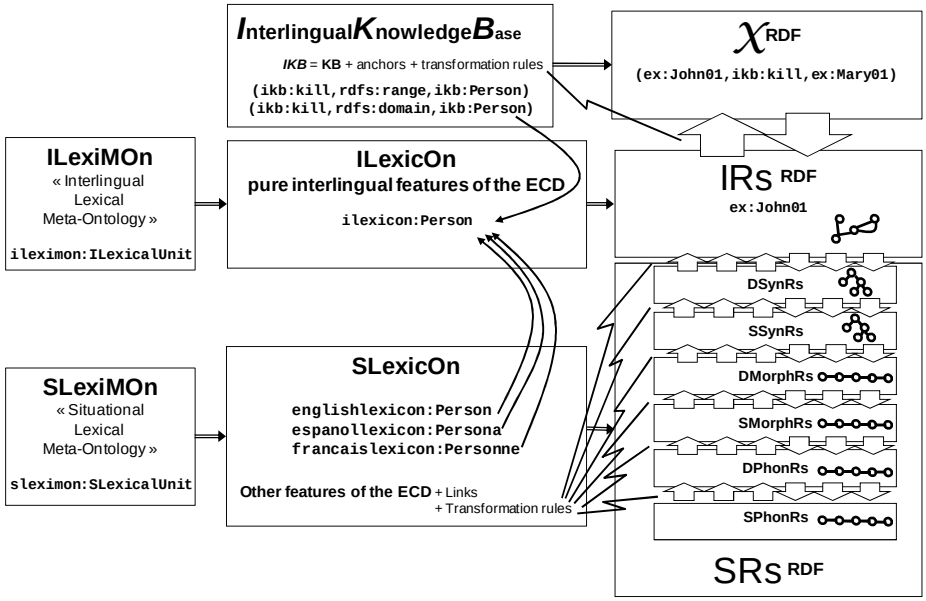


Fig. 2. Overview of the architecture of the ULiS.
From top to bottom: the interlingual layer, the interlingual layer, the situational layer.
From left to right: meta-ontologies; ontologies; facts and different representations.

4.2 Architecture in the interlingual layer

The interlingual layer of ULiS is divided in three components:

The meta-ontology. The *interlingual lexical meta-ontology* (ILexiMOn) is the schema that the ILexicOn must satisfy to be compliant with the pure semantic features of the Explanatory Combinatorial Dictionary (ECD). It defines meta-classes, uses RDFS and some of OWL full's axioms, and contains *ad hoc* SPIN validation and inference rules for the ILexicOn and the *interlingual semantic representations* (ISemRs).

The ontology. The *interlingual lexical ontology* (ILexicOn) is the interlingual dictionary where *interlingual lexical unit classes* (ILU^cs) are formally defined as instances of the ILexicalUnit meta-class from the ILexiMOn. The ILexicOn contains all the pure semantic features of the *Explanatory Combinatorial Dictionary* (ECD). Any concept expressible in a natural language or a jargon is defined in the ILexicOn that contains:

- The formal definitions of the ILU^cs (described in section 5.2)
- The definitions of *interlingual attribute classes* (IAtts) (e.g., plural, future, 1st person, indefinite, etc.);

- The definitions of the *interlingual semantic relations* (ISemRels), that are used in the formal definitions of the ILU^{c}s and to construct *interlingual semantic representations* (ISemRs);
- Interlingual lexical functions: every purely-semantic lexical links such as synonymy, and purely-semantic generic constructions such as the lexical function Centr(X), i.e., (the center of X), or Fin(X), i.e., (stop being X).

The interlingual semantic representations. ISemRs are RDF graphs with nodes being *interlingual lexical unit instances* (ILU^{i}s), and arcs being ISemRels. ILU^{i}s may also be instances of IAtts. Arcs are *interlingual semantic relations* (ISemRels).

4.3 To and from Natural Language facts

Situations. Interlingual-based lexical resources consider connecting language specific dictionaries to some interlingual dictionary. We generalize this by using situations (i.e., the situations of understanding and use of some linguistic element).

The situation of a linguistic element is part of the pragmatics of its use: it represents not only the language used (e.g., EN, FR), but also sociolectal marks (e.g., biologists, architects, official, slang, reverential), topolectal marks (e.g., U.S., Canada), chronolectal marks (e.g., old, neologic), and even individual marks (e.g., a particular group of people). The intersection of situations is also a situation (EN-U.S.-slang), and so is the union of situations (FR-Canada OR FR-France-old).

Architecture of the situational layer. This architecture purposefully mirrors the interlingual layer:

A *situational lexical meta-ontology* (SLexiMOn) describes the SLexicOn,

A *situational lexical ontology* (SLexicOn), contains all non-purely semantic features of the ECD. A non-exhaustive list is the following:

- Definitions of *situational lexical unit classes*, called SLU^{c}s, by means of a link to an ILU^c, which is annotated by a specific situation.
- Situational lexical functions such as Instr(X), i.e., the preposition that governs the keyword X and means: (by means of).
- Situational attribute classes (e.g., invariable English nouns, French 1st verb group, German dative, etc.), their associated situations and rules.
- Situational relations: relations that link two instances of the SLU^{c}s, thus defining the dependency syntax of the utterance, or the order of the words in an utterance.

Situational representations (SRs). The data consist of *situational representations* (SRs): RDF graphs having *situational lexical unit instances* (SLU^{i}s) as nodes and situational relations as arcs. A SR thus represents the different representations of the Meaning-Text theory.

Transformation rules. Contrary to the Common Web Language (CWL), where no grammar rules representation is proposed, we plan to introduce *transformation rules*

in the SLexiMOn. Transformation rules form a subclass of the SPIN rules and are attached to a SLU[c] to define a correspondence between a generic pattern from a representation level, to another pattern at a deeper or to a higher representation level. Thus, each situation may define its own analysis and production grammar, both made of six sets of transformation rules.

Transformation rules may be sorted according to their level of genericity: transformation rules that are attached to ISemRels, or to IAtts, are less specific than rules that may be triggered only when a complex ISemR patterns is met; also, rules that may be triggered in generic situations are less specific than those that may only be triggered in more specific situations. The important point is that a rule must be triggered if and only if there is not a more specific rule that can be triggered instead. This implies that an algorithm different from the simple forward-chaining algorithm must be proposed. It will be very important to optimize the application of such an algorithm with a whole set of rules. We therefore plan to construct a Rete network (Forgy, 1982) on top of each set of transformation rules, which is eased by the SPIN framework as each rule is modeled as an RDF graph.

Finally, a set of generic transformation rules must be designed to ensure that for each situation, every SR is transformable to an ISemR, and that every ISemR is transformable to a SR. When a new situation is introduced (e.g., a new language), this criterion is *a priori* not met. This is the reason why we suggest the introduction of the universal situation, and transformation rules that produce Notation3-like output. We claim that a small set of rules will suffice to produce and analyze simple controlled natural languages.

4.4 To and from Interlingual Knowledge Bases facts.

Interlingual knowledge bases. The main criterion that an interlingual knowledge base must meet is that any RDF graph inside it must be transformable into an *interlingual semantic representation* (ISemR). We thus propose to form interlingual knowledge bases by augmenting classic knowledge bases with *anchors* and *transformation rules*:

- An anchor is a triple that links an RDF resource to an ILU[c]. For instance, the RDF resource `rdfs:Class` will be anchored to a specific ILU[c] `ilexicon:RdfClass` that formally defines the concept of an RDF class, and that is itself linked to an English SLU[c] that is a pluralizable noun realized by the string "class";
- The transformation rules are stored in the interlingual knowledge base and form two separated sets of rules: one for producing RDF from an ISemR, the other for producing an ISemR from RDF. Here again, transformation rules may be sorted according to their level of genericity, and the most generic rules must be inhibited when more specific ones can be triggered.

Augmenting classic semantic web formalisms. The output of an ISemR must be a valid SPARQL request, and the output of any RDF graph must be a valid ISemR. This criterion will be satisfied by the introduction of different anchors and generic trans-

formation rules in the classic semantic web vocabularies: RDF, then RDFS, OWL and SPIN, and finally SKOS. Thus an RDF class that has no anchor, e.g., `foaf:Person`, has a correspondence with an ISemR that itself has a correspondence to the textual representation for the EN situation: "The RDF class `foaf:Person`".

5 Modeling Choices in the Interlingual Layer

5.1 Overview

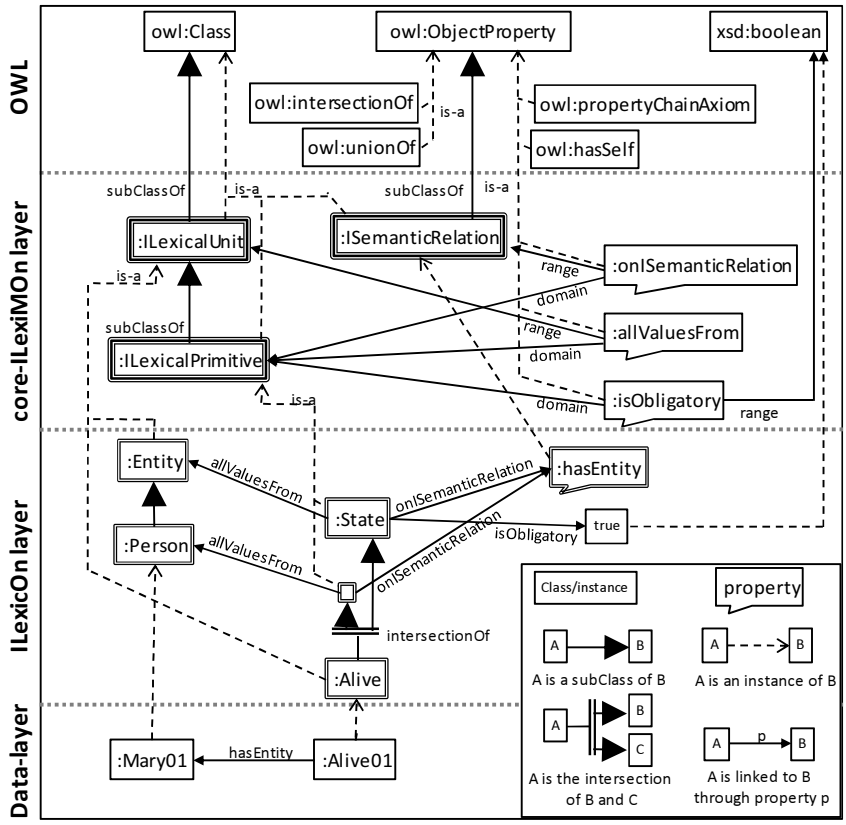


Fig. 3. The three compoents of the interlingual layer, with details of the whole core-ILexiMOn that we introduced, and overview of the light standalone ILexicOn and the data.

Figure 3 illustrates the architecture of our work, with its integration in the semantic web formalisms. To validate our approach, we designed a light core-ILexiMOn[1], a light standalone ILexicOn[2], and simple ISemRs[3].

[1] RDF/XML document available at URL: http://ns.inria.fr/ulk/2011/06/10/ileximon-core
[2] RDF/XML document available at URL: http://ns.inria.fr/ulk/2011/06/10/ilexicon-ex
[3] RDF/XML document available at URL: http://ns.inria.fr/ulk/2011/06/10/sems-ex

From top to bottom: 1) the semantic web formalisms, with a few OWL classes and properties that are useful for our work; 2) the detailed core-ILexiMOn; 3) an overview of the light standalone ILexicOn; and 4) an overview of data from the interlingual data component. Notice that: i) ILU^is from the data are instances of ILU^cs described in the ILexicOn, that are themselves instances of the ILexicalUnit meta-classes described in the ILexiMOn; and ii) properties used to link two resources in a layer are described in an upper layer.

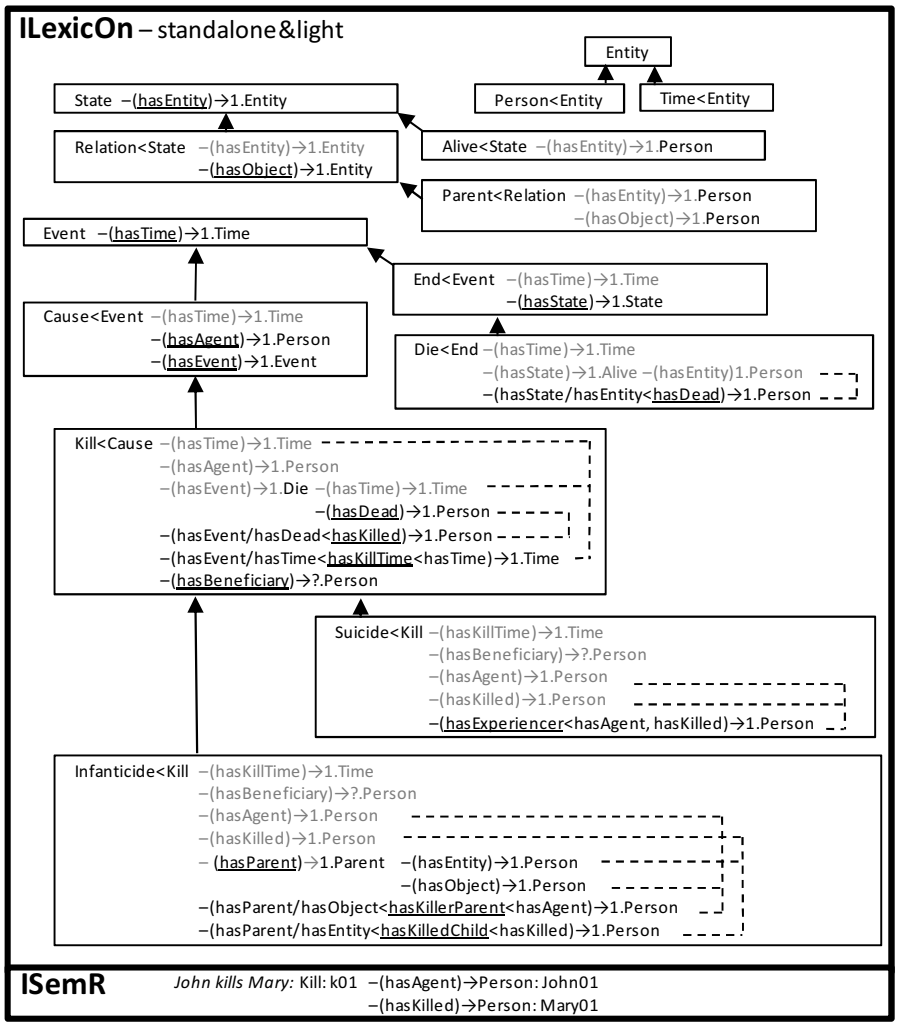


Fig. 4. The light standalone ILexicOn and one ISemR described with our notation.

Figure 4 above concisely describes the light standalone ILexicOn using a notation inspired from Sowa's conceptual graphs (Sowa, 1984). Each rectangle represents the definition of the ILUc that is written in its top-left corner.

5.2 The lexicographic definition of lexical units

In the ILexicOn is propose a novel approach to the lexicographic definition of an ILU^c that consists in projecting the minimal semantic decomposition of the ILU^c on the ILU^c using *Conceptual Participant slots* (ConP-slot): the implicit semantic link that exists between an ILU^c L and one of the participants of the minimal semantic decomposition of L (Mel'čuk, 2004 ; Lefrançois & Gandon, 2011).

Interlingual lexical units (classes and instances): ILU^cs are instances of the ILexicalUnit meta-class from the ILexiMOn (c.f., Figure 3). They are defined in the ILexicOn (c.f., Figure 4, e.g., Entity, Person, State, Alive, Event, Cause). In our notation, symbol < represents the rdfs:subClassOf axiom that may be used to state inheritance between ILU^cs (e.g., Person<Entity, Alive<State, Cause<Event). For instance, The ILU^c Person is a sub-class of the ILU^c class Entity, and the ILU^c Entity is the parent of the ILU^c Person. Complex ILU^cs may be constructed through owl:intersectionOf and owl:unionOf. Finally, *interlingual lexical unit instances* (ILU^is) are instances of ILU^cs and are used in the data component as nodes of the interlingual semantic representations.

Interlingual semantic relations: ISemRels are instances of the ISemRelation meta-class of the ILexiMOn, and thus instances of owl:ObjectProperties. They are introduced in the LexicOn and used in the data to link ILU^is (see Figure 3&4). In our notation, symbol < represents the rdfs:subPropertyOf axiom that may be used to define a new ISemRel as being a sub-ISemRel of one or more ISemRels (e.g., hasExperiencer<hasAgent, hasKilled). Symbol / represents the owl:propertyChainAxiom axiom that may also be used to state that a ISemRel is a super-ISemRel of the composition of two or more ISemRels (e.g., hasState/hasEntity<hasDead). These two axioms may be combined to define complex ISemRels (e.g., hasEvent/hasTime<hasKillTime<hasTime).

Interlingual lexical primitives: An ILU^c L is a ILP^c if and only if it derives from no other ILU^c but has at least one ConP-slot. Non- lexical primitives then derive from one or more lexical primitives following the *ConP-slot* inheritance and introduction principle:

> An ILU^c L inherits from its parents' ConP-slots, and may also introduce new ConP-slots;

One may thus consider only participants that are necessary and sufficient to the minimal projection of L. ILP^cs are defined as instances of the ILexicalPrimitive meta-class from the ILexiMOn (c.f., Figure 3). An ILP^c must be linked through: i) the onISemanticRelation property to exactly one ISemanticRelation; ii) the allValuesFrom property to exactly one ILexicalUnit; and iii) the isObligatory property to exactly one xsd:boolean.

Conceptual participant slots: In Figure 4, each line with an arrow in the definition of an ILU^c represents a conceptual participant slot (ConP-slot) that restricts the use of a specific ISemRel for this ILU^c and its descendants. Actually, such a line means that the defined ILU^c is a sub-class of an ILP^c. For instance, the line State–(hasEntity)→1.Entity states that any instance of the State class is linked exactly once

through the hasEntity relation to an instance of the Entity class. Let us focus on the notation used on Figure 4:

- **Inheritance**. ConP-slots may be newly defined (black font, e.g., State–(hasEntity)→1.Entity), fully inherited (grey font, e.g., Relation<State–(hasEntity)→1.Entity) or partially inherited (grey font for the inherited part, e.g., Alive<State–(hasEntity)→1.Person). The ILU^c on the right hand side of the line is called the *current range of the ConP-slot*.
- **Obligatory vs. optional.** A ConP-slot may be obligatory (symbol 1, e.g., Alive<State–(hasEntity)→1.Person) or optional (symbol ?, e.g., Kill<Cause–(hasBeneficiary)→?.Person). When an optional ConP-slot is inherited, it may be restricted to being obligatory.
- **Domain/range of the ISemRel.** As an ISemRel is an rdf:Property, it may restrict its domain and its range i.e., what ILU^c the subject (resp. the object) of a triple that involves this ISemRel does belong to. When an ISemRel is underlined, it means that its domain is set to the defined ILU^c, and that its range is set to the current ILU^c range of the ConP-slot. (e.g., State–(hasEntity)→1.Entity).
- **ISemRel subproperty and composition axioms.** As we stated in section 4.2.2, complex ISemRel may be defined thanks to inheritance and composition. There are benefits in using such ISemRel to qualify a new ConP-slot. In fact, this combined with the maximum cardinality of ConP-slots restricted to 1, imposes the equality of ILU^i in the data. We illustrate these inferable equalities by dotted lines on the right of ConP-slots.

The ISemRel inheritance and composition is what enables the projection not only of trees, but also graphs, onto one node. Thus, each ILU^c described in the ILexicOn contains the projection of its semantic decomposition graph. We illustrated this on Figure 4 with complex ILU^c such as ilexicon:Suicide (the killer is the person killed) and ilexicon:Infanticide (the killer is the parent of the person killed).

6 Conclusions and discussions

We introduced a *universal linguistic system* (ULiS) through which multiple actors could interact with an *interlingual knowledge base* (IKB) in controlled natural language. We explained an interaction scenario with ULiS, which can serve for machine translation and for multilingual management of interlingual knowledge bases. We then gave an overview of the layers ULiS is made of: the interlingual layer; the situational layer; and an interlingual knowledge base.

The main novelty of our proposal is that the characteristics of each controlled natural language are stored in a specific interlingual knowledge base. Thus, actors could enhance their controlled natural language through requests expressed in controlled natural language.

We introduced and illustrated a novel approach to formally define ILU^cs: we make ILU^cs support a projection of their semantic decomposition. We introduced a human-readable notation to represent ILexicOn, and we used this notation to validate our

approach with a simple standalone ILexicOn. We thus showed that simple and complex ILU^{c}s may be formally defined with our novel approach.

We are currently working on the formalization of lexical functions in the ILexicOn and of the SLexicOn, and we are to partly populate our lexical resources with lexical units from other lexical resources such as the French Lexical Network. We finally plan to validate our results by the design and the experimentation of a web-based prototype with a simple interlingual knowledge base (e.g., the "interlingual-augmented" wine ontology), and a few situations based on English and French.

References

1. Baker, C.F.,Fillmore, C.J., and Lowe, J.B.: The berkeley framenet project. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1, 16-90 (1998)
2. Boguslavsky, I.: Some Lexical Issues of UNL. Proceedings of the First International Workshop on UNL, other interlinguas and their applications, Las Palmas, 19-22 (2002)
3. Boguslavsky, I.: Some controversial issues of UNL: Linguistic aspects. Research on Computer Science, 12:77-100 (2005)
4. Blay-Fornarino, M., Pinna-Dery, A.-M., Schmidt, K., and Zaraté, P.: Cooperative Systems Design: A Challenge of the Mobility Age. In: Proceedings of COOP 2002, Saint-Raphaël, France, 4-7 June 2002, IOS Press, pp.23–37 (2002)
5. Forgy, C.: Rete: A Fast Algorithm for the Many Pattern/Many Object Pattern Match Problem. In: Artificial Intelligence, vol. 19, pp. 17–37 (1982)
6. Kahane, S.: The Meaning-Text Theory, Dependency and Valency. In: Handbooks of Linguistics and Communication Sciences 25 : 1-2, Berlin/NY: De Gruyter, 32 p. (2003)
7. Lefrançois, M., and Gandon, F.: ILexicOn: toward an ECD-compliant interlingual lexical ontology described with semantic web formalisms. In Proceedings of the 5th Meaning-Text Theory, Barcelona (Spain), 155-164 (2011)
8. Lux-Pogodalla, V., and Polguère, A.: Construction of a French Lexical Network: Methodological Issues. International Workshop on Lexical Resources (2011)
9. Mel'čuk, I.A., Clas, A., and Polguère, A.: Introduction à la lexicologie explicative et combinatoire (1995)
10. Mel'čuk, I.A.: The Meaning-Text Approach to the Study of Natural Language and Linguistic Functional Models. [Invited lecture.] In S. Embleton (ed.): LACUS Forum 24, Chapel Hill: LACUS, pp. 3–20 (1998)
11. Mel'čuk, I.A.: Actants in semantics and syntax I: Actants in semantics. Linguistics. 42(1):1-66 (2004)
12. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K.: Introduction to wordnet: An on-line lexical database. Int. Journal of lexicography, 3(4):235-344 (1990).
13. Periñán Pascual, C. and Arcas Túnez, F.: The architecture of FunGramKB. In Proceedings of the Seventh International Conference on Language Resources and Evaluation, Valeta (Malta), 2667-2674 (2010)
14. Sowa, J.F.: Conceptual structures: information processing in mind and machine, System programming series, Addison-Wesley (1984)
15. XGR-CWL: Report of W3C Incubator Group on Common Web Language, http://www.w3.org/2005/Incubator/cwl/XGR-cwl-20080331/ (2008)

Direct and Indirect Linking of Lexical Objects for Evolving Lexical Linked Data

Yoshihiko Hayashi

Graduate School of Language and Culture, Osaka University
1-8 Machikaneyama, 5600043 Toyonaka, Japan
hayashi@lang.osaka-u.ac.jp

Abstract. *Servicization* of language resources in a Web-based environment has opened up the potential for dynamically combined virtual lexical resources. *Evolving lexical linked data* could be realized, provided being recovered/discovered links among lexical resources are properly organized and maintained. This position paper examines a scenario, in which lexical semantic resources are cross-linguistically enriched, and sketches how this scenario could come about while discussing necessary ingredients. The discussions naturally include how the existing lexicon modeling framework could be applied and should be extended.

Keywords: lexical linked data, lexicon models, multilingual lexical resources, cross-lingual semantic similarity

1 Introduction

Servicization of language resources provides the potential of a dynamic lexical resource [4], which realizes a virtual yet composite lexical resource by combining servicized resources with a service workflow. Furthermore, it is expected that the recovered/discovered relationships among lexical objects in existing language resources can be organized as a secondary language resource, and hence can be effectively reused [6]. This direction could harmonize with the recent trend of Linked Data, as the derived relationships are being overplayed as *links* on top of the primary lexical resources. We would call such a lexical space *evolving lexical linked data* as a whole.

This position paper argues that by opportunistically associating different lexical resources across a language barrier, relevant portion of the lexical resources can be gradually enriched and could be made public by standing on the Linked Data mechanism. This paper also argues more relationships could be acquired, when there exists a lexical semantic disparity.

2 Basic Lexicon Model

The presented work concentrates on WordNet-type semantic lexicons. Their fundamental information structures are represented by the following lexical class objects.

– A `Lexical Entry` comprises of `Form`s and `Sense`s.
– A `Form` can be a `Lemma` or a `Phrase`; the latter comprises of more than one `Lemma`s.
– A `Sense` denotes a `Synset`.
– A `Synset` is denoted by one or more `Sense`s.
– `Synset`s are linked by one of the predefined `Conceptual Relation`s.

3 Conceptual Framework of Evolving Lexical Linked Data

Below we introduce a motivating example, where an English query term *gadget* is issued to search for a set of corresponding Japanese translations, each hopefully grounded in a Japanese conceptual system. Suppose we get two translations, under the same sense division, for *gadget* by using an appropriate translation resource: *t1*:"ガジェット" (*gajetto*), which is the transliteration of *gadget*, and *t2*:"有用な機器" (*yuuyounakiki*), which actually is a two-word phrase.

3.1 Direct Linking of Lexical Objects

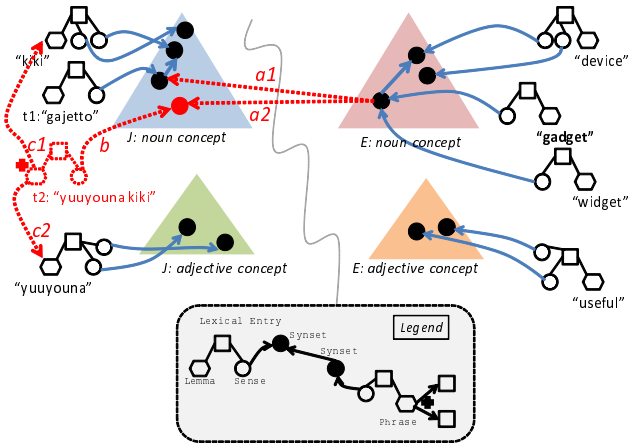


Fig. 1. Evolving Lexical Linked Data: direct linking has been conducted.

Figure 1 illustrates the relevant portion of the lexical linked data just after the query was entered, in which newly introduced lexical objects are indicated by dotted lines. First, cross-lingual synset-to-synset links *a1* and *a2* are introduced. Introduction of *a1* may require sense disambiguation, because *t1*, which is supposed to reside in the Japanese lexical space, could have more than one senses. A `Lexical Entry` node as well as a `Synset` node are, on the other hand, introduced for accommodating *t2*. As *t2* should be morpho-syntactically parsed

into [有用な (yuuyouna)/Adj, 機器 (kiki)/Noun], a `Phrase` node is introduced to associate this two-word phrase with its constituents by the *c1* and *c2* links.

These successive operations are invoked directly while handling the query; we thus call them *direct linking* of lexical objects. Note that the ad-hoc `Synset` node is yet to ground in the Japanese conceptual system at this time.

3.2 Indirect Linking of Lexical Objects

While the structure around *t1* has been settled in the current configuration, that of around the ad-hoc `Synset` node for *t2* can be further enriched, again by seeking cross-lingual correspondences. Figure 2 summarizes the outcomes.

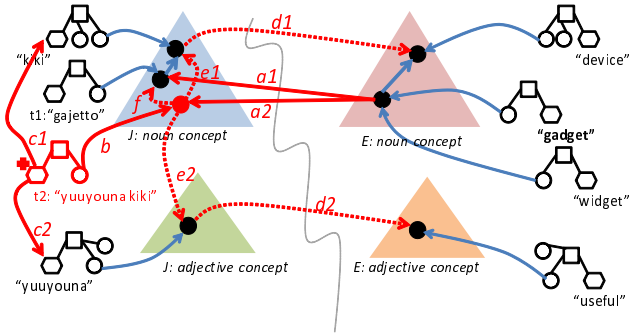


Fig. 2. Evolving lexical linked data: indirect links are introduced.

Two cross-lingual synset-to-synset links (*d1* and *d2*) are first introduced by associating a sense of "機器" (*kiki*) with a sense of *device* and a sense of "有用な" (*yuuyouna*) with a sense of *useful* respectively. By establishing *d1*, the semantic head of the ad-hoc synset for *t2* is then identified and represented by the link *e1*. The same story holds for the semantic modifier of *t2*, and the link *e2* is introduced to represent this semantic relationship. These operations also enable the introduction of the link *f*, which, in a sense, shows "ガジェット" (*gajetto*) can be rephrased as "有用な機器" (*yuuyounakiki*).

The evolving story so far signifies us the possibility of lexical knowledge enrichment that takes advantage of the opportunity to interrelate lexical objects across a language barrier. Let us remind that a semantic gap brought about by differences in the lexicalization would provide us a further opportunity to enrich relevant range of the existing lexical structures.

We could acquire more correspondences as illustrated in Figure 3 by further pursuing this strategy. In the figure, another ad-hoc `Synset` node in the English lexical space, and two semantic links (*g1* and *g2*) to label the semantic head/modifier of the ad-hoc synset are introduced. Besides, the ad-hoc `Synset` node is linked to that of *gadget* by the link *h*; this is in parallel with the link *f* in the Japanese lexical space. Notice again that almost instant introduction of

these links is originated from the cross-lingual synset-to-synset matching that is invoked for establishing the correspondences represented by *d1* and *d2*.

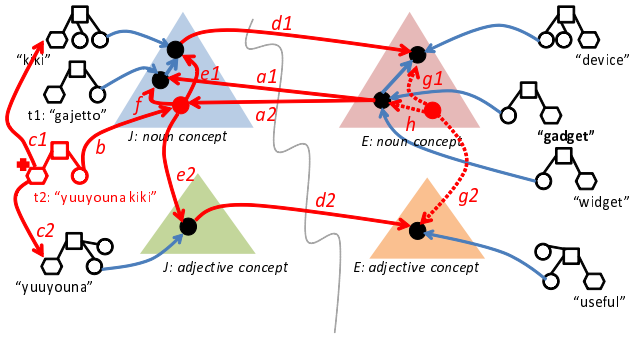


Fig. 3. Evolving lexical linked data: indirect links are further introduced.

We would call these secondary operations initiated after the direct linking as *indirect linking*. The lexical objects introduced in this motivating example are examined in more detail in the next section to sort the necessary elements to realize the scenario.

4 Enabling Direct and Indirect Linking

4.1 Modeling lexical information structure

The basic lexicon model described in section 2 has to be extended in some ways.

First, in the motivating example, two ad-hoc `Synset` nodes were introduced to accommodate the two-word translation phrase *t2*, and the corresponding virtual phrase (could be verbalized as *useful device*) in English. These nodes, in their nature, may be ad-hoc and represent a kind of complex concept that may lexicalize to a phrase rather than a single word in one language. Therefore an instance of the ad-hoc `Synset` class should have an attribute to indicate the instance is typed `complex`, and could have `Morpho-syntactic Head/Modifier` links (like *c1,c2*) as well as `Semantic Head/Modifier` links (like *e1,e2,g1,g2*).

Second, some of the introduced links should be typed differently from the existing lexicon model. Table 1 classifies the links introduced in the motivating example. The link type #1 is of intrinsic important in the presented framework. As the correspondence between synsets in different languages, in a sense, is rarely equivalent [7], it is necessary to label the relation type for each cross-lingual synset-to-synset link instance. We could develop a proper label inventory, presumably by basing on the one developed by EuroWordNet [9], while considering more bilingual characteristics. The link type #5, in a sense, is a variant of the link type #1; the difference is that the correspondence is cross-lingual or not. Therefore we can assume an upper class that subsumes these link types.

Table 1. Classification of the links introduced in the motivating example.

#	link instances	source node type	destination node type	relation type	computational process
1	*a1,a2,d1,d2*	`Synset`	`Synset`	cross-lingual correspondence	synset matching
2	*b*	`Sense`	`Synset`	denotation	–
3	*c1,c2*	`Phrase`	`Lemma`	morpho-syntactic decomposition	morpho-syntactic analysis
4	*e1,e2,g1,g2*	ad-hoc `Synset`	`Synset`	semantic decomposition	–
5	*f,h*	ad-hoc `Synset`	`Synset`	near-synonym	–

The link type #3 represents morpho-syntactic head/modifier relationships, whereas link type #4 represents semantic head/modifier relationships. As far as semantic compositionality holds, these two link types exhibit a kind of parallel structure as illustrated in the example: the semantic links (*e1* and *e2*; typed #4) were eventually introduced, corresponding to the already existing morpho-syntactic links (*c1* and *c2*; typed #3).

On the other hand, in cases where the semantic compositionality does not hold, we should demur the introduction of these semantic links, even each of the Japanese synsets could find their mates in the English lexical space. In such a case, we have to devise an independent method to check the semantic compositionality, or we should seek more semantic constraints to apply, probably from the English lexical space; but this issue largely remains as a future issue.

As for the actual modeling and representation of lexical resources, we can rest with the existing frameworks, including the ISO standard lexical markup framework (LMF) [5], and Lemon [3].

4.2 Matching synsets across a language

One of the most important elements is obviously a computational process for finding a synset mate in another language. We are now studying a method to calculate semantic similarity between synsets across a language, by simply employing bilingual translation resources and probability distributions acquired from a sense-tagged corpus in the target language.

We can also apply and/or combine previously proposed methods. For example, the method reported highly accurate [1] may be applicable with modifications, even it computes similarity between words rather than between synsets; the gloss-overlap-based method presented in [2] would also be readily applied, if we could translate the gloss in one language to another with a reasonable accuracy. However even with a highly promising method at hand, any synset-to-synset relation has to be established by choosing among computationally proposed candidates. The underlying process thus has to incorporate human intervention, where a collaborative operational environment plays a role.

4.3 Further issues

The following issues have to be considered in implementing an effective operating environment. First, we need to have a global mechanism to control the indirect linking operations. As shown in the example, indirect links can be introduced upon establishment of a direct link. However who/what should decide to initiate the indirect linking process is unclear. Moreover, to what extent the indirect linking should be propagated remains uncertain. Second, we are in need of having a proper vocabulary to annotate the lexical objects that participated in direct/indirect linking operations. For example, we would need to know when and how a particular link was established. We thus need to have a sort of ontology for describing linking events, which naturally includes references to the linguistic processes that were actually applied, as well as the human approvals.

5 Concluding Remarks

This position paper presented a notion of evolving linked data, in which recovered/discovered relationships among lexical objects would be published as *links*. It also argued that the associated lexical resources could be enriched further, in particular cases where a sort of lexical semantic disparity exists.

Acknowledgments. The presented work was supported by KAKENHI (21520401) provided by MEXT, Japan, and the Strategic Information and Communications R&D Promotion Programme (SCOPE) of the Ministry of Internal Affairs and Communications of Japan.

References

1. Agirre, E., Alfonseca, E., et al.: A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In: *NAACL-HLT2009*, pp.19–27 (2009)
2. Banerjee, S., and Pedersen, T.: Extended Gloss Overlaps as a Measure of Semantic Relatedness. In: *IJCAI 2003*, pp.805–810 (2003)
3. Buitelaar, P., Cimiano, P., et al.: Towards Linguistically Grounded Ontologies. In: *ESWC 2009*, pp.111–125 (2009)
4. Calzolari, N.: Approaches towards a 'Lexical Web': the Role of Interoperability. In: *ICGL 2008*, pp.34–42 (2008)
5. Francopoulo, G., Bel, N. et al.: Multilingual Resources for NLP in the Lexical Markup Framework (LMF). *Language Resources and Evaluation*, Vol.43, No.1, pp. 57–70 (2009)
6. Hayashi, Y.: A Representation Framework for Cross-lingual/Interlingual Lexical Semantic Correspondences. In: *IWCS 2011*, pp.155–164 (2011)
7. Hirst, G.: Ontology and the Lexicon. In: Staab, S., and Studer, R. (eds.): *Handbook of Ontologies, Second Edition*. Springer, pp.269–292. (2009)
8. Isahara, H., Bond, F., et al.: Development of the Japanese WordNet. In: *LREC 2008*, pp.2420–2423 (2008)
9. Vossen, P.: EuroWordNet: A Multilingual Database of Autonomous and Language-Specific Wordnets Connected via an Inter-lingual Index. *International Journal of Lexicography*, Vol.17, No.2, pp.161–173 (2004)

Linking Domain-Specific Knowledge to Encyclopedic Knowledge: an Initial Approach to Linked Data

Pilar León Araúz[1], Pamela Faber[1], Pedro J. Magaña Redondo[2]

[1] Department of Translation and Interpreting, University of Granada
Buensuceso, 11 18002 Granada, Spain
[2] Andalusian Centre for the Environment (CEAMA), University of Granada
Avda. Mediterráneo s/n, 18071 Granada, Spain
{pleon, pfaber, pmagana}@ugr.es

Abstract. Linked Data creates a shared information space by publishing and connecting resources in the Semantic Web. However, the specification of semantic relationships between data sources is still a stumbling block. One solution is to enrich ontologies with multilingual and concept-oriented information. Usefully linking entities in the Semantic Web is thus facilitated by a semantic-oriented cross-lingual ontology mapping framework in which knowledge representations are not restricted to a particular natural language. Accordingly, this paper describes a preliminary approach for integrating general encyclopedic knowledge in DBpedia with EcoLexicon, a multilingual terminological knowledge base on the environment.

Keywords: terminology, knowledge representation, linked data, multilinguality

1 Introduction

Knowledge bases play an increasingly important role in enhancing the intelligence of Web as well as in supporting information integration [1]. In this respect, the Semantic Web is an extension of the current Web in which information is given a well-defined meaning, better enabling computers and people to work in cooperation [2, 3]. This refers to people all over the world, who speak different languages. As Cimiano [4] states, the Semantic Web has the potential for dealing with cross-linguistic mappings since its content is structured much like a database and thus is language-independent.

The awareness of linguistic complexity has intensified over the last ten years as the number of Internet webpages in other languages has soared. This is a challenge for usefully linking entities in the Semantic Web because this process requires some sort of semantic-oriented cross-lingual ontology mapping framework in which knowledge representations are not restricted to the use of a particular natural language [5]. However, without a coherent description of concepts and terminological variants that take into account the categorization of real world entities by other language communities, the Semantic Web will never be truly multilingual. We thus propose a model for integrating general encyclopedic knowledge in DBpedia with our domain-specific resource, EcoLexicon (http://ecolexicon.ugr.es), a multilingual terminological knowledge base (TKB) on the environment.

2 EcoLexicon

EcoLexicon was initially implemented in Spanish, English, and German though more languages are currently being added (Modern Greek, Russian, Dutch and French). So far it has a total of 3,271 concepts and 14,646 terms. One of its main assets is its multilinguality, but as an added value, its user interface also includes semantic networks, graphical resources, definitions and contextual information that enhance the representation of conceptual and terminological knowledge [6]. Nevertheless, the focus here is on terminological and semantic information since these data sets are the ones used for linking EcoLexicon to DBpedia. For every environmental concept, multilingual choices are made available. The users can then click on any of them and obtain terminological information, such as whether a linguistic designation is a synonym, acronym, register, or stylistic variant. For instance, the concept EBB CURRENT has a total of 16 different designations in Spanish, English, German, and Greek since all registers and linguistic varieties are accounted for. In this entry, *ebb tide* appears as a non-technical variant for the concept EBB CURRENT. Even in large term bases, multilingual variety is rarely represented in an exhaustive way. However, this type of information is invaluable because not only does it provide users with multiple options for text comprehension and production, but it is also useful for conceptual disambiguation (section 3). Concepts are also displayed in dynamic semantic networks linked to other concepts. Nevertheless, problems can arise when it is a question of browsing networks of very general concepts, which carry an excessive load of information (Fig. 1).

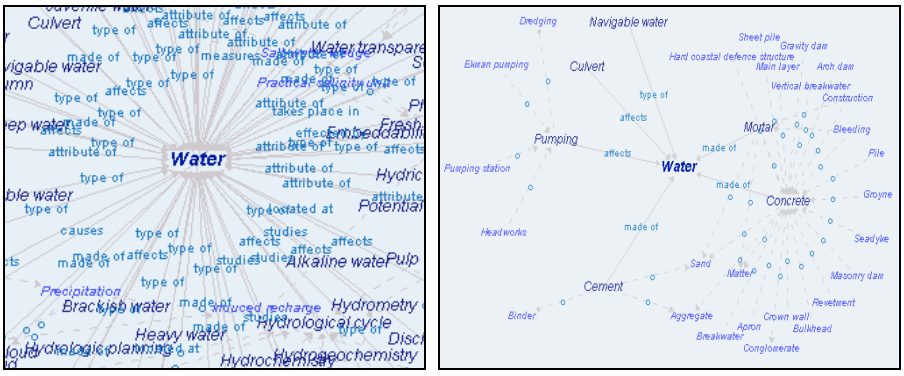


Fig. 1. and **2.** Context-free and context-based semantic network of WATER

Overloaded concepts, such as WATER, share multiple relations with many other concepts, but they rarely, if ever, activate all those relations at the same time since this would evoke completely different and incompatible scenarios. Our claim is that any specialized domain contains sub-domains in which conceptual dimensions become more or less salient, depending on the activation of specific contexts [7]. The area of environmental knowledge was thus divided into a set of contextual domains (e.g. HYDROLOGY, GEOGRAPHY, OCEANOGRAPHY, CIVIL ENGINEERING, ENVIRONMENTAL ENGINEERING, etc.) and the relational power of concepts was

constrained accordingly. Thus, when constraints are applied, the network of WATER within the CIVIL ENGINEERING domain is recontextualized and becomes more meaningful (Fig. 2).

EcoLexicon is primarily hosted in a relational database (RDB), but at the same time it is integrated in an ontological model. Semantic information is stored in the ontology, while leaving the rest in the relational database [7]. This is important because the linked data process not only involves the transformation of data to RDF format, but also includes the use of terminologies, controlled vocabularies, and ontologies to describe triples attributes in a systematic way and as reference conceptual models to support an integrated view of data and semantic interoperability between datasets [8]. As seen in Fig. 3, contextual domains have inspired the design of our ontology classes. The ontology is automatically retrieved from the data stored in our RDB, according to the following assumption: if a concept *c* is part of one or more propositions allocated to a contextual domain *C*, *c* will be an instance of the class *C*. EcoLexicon keeps multilingual terminological information and ontological information separate. Each terminological entry has different word forms linked to the same natural language definition, constrained by the knowledge represented in the ontology concept [9].

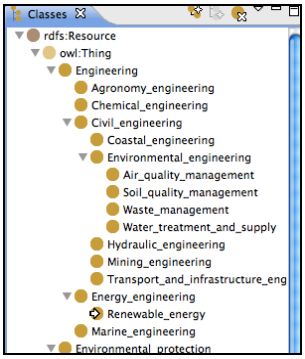


Fig. 3. Ontological classes

3 Linking EcoLexicon to DBpedia

Linked Data is an important initiative for creating a shared information space by publishing and connecting structured resources in the Semantic Web [10]. However, the specification of semantic relationships between data sources is still a stumbling block. Our initial proposal is to integrate EcoLexicon with DBpedia through the *sameAs* property, because: (1) DBpedia is at the core of the Linked Data initiative; (2) users can complete their knowledge acquisition process through a guided access to encyclopedic knowledge.

Linking data sources from DBpedia can be quite straightforward since different tools, such as the ontology editor TopBraid Composer, can automatically suggest the links. However, because of lexical variation and the lack of univocity in both general and specialized knowledge, automatic mappings are not always viable. Furthermore, although establishing an identity relation initially may appear to be a simple task,

matching two entities, both at the syntactic and the semantic levels, is often far from easy [11]. Problems with text searching and entity matching highlight the fact that a word is more than a mere string of characters. The following are basically the same problems that have plagued linguists over the years: polysemy, homonymy, synonymy, and different levels of specificity [12]. There are also other aspects of lexical meaning that lead to confusion, such as the fact that: (i) the meaning of a term can vary, depending on the context; (ii) meanings can change in time and space; (iii) different languages reflect different mappings of reality, which may coincide totally, partially, or not at all.

A solution to some of these problems can be found when ontologies are enriched with multilingual and concept-oriented information, as reflected in the field of environmental knowledge, but manual work is still necessary to a certain extent. Nevertheless, instead of mapping one-to-one manual correspondences, we can take advantage of the semantics contained in each resource. In our approach, the term strings of EcoLexicon are compared with those from DBpedia, enhanced by those data sets that include multilingual choices and variants as well as category membership. To illustrate our data linking proposal, we have chosen four concepts: GROIN, BANK, ACCRETION, WASTEWATER TREATMENT PLANT and the pseudocode of the general matching algorithm is shown in the following table:

```
for each w:word in ecolexicon
for each cp:concept in dbpedia
    w' = stem(w); cp' = stem(cp)
    if str_compare(v', cp') > word_threshold
      multi_e = multilingual_variants(v)
      multi_g = multilingual_variants(cp)
      if multingual_compare(multi_e, multi_g) > multilingual_threshold
        result.add(pair(v, cp))
        related_instances = instances_of(context(v))
        for each i:instance in related_instances
          if look_for_text(comment_properties(cp), i) > text_threshold
            result.add(pair(v, i))
```

The concept GROIN in DBpedia is not designated by its most frequent form but by a geographical variant (*groyne*). The fact that EcoLexicon stores all lexical variations of each concept allows us to identify the same entity in both resources by comparing the string of all our English monolingual variants with the entries in DBpedia. However, if the search was only performed for the string *groin*, DBpedia would redirect to a disambiguation page, since GROIN can also refer to a part of the human body. In this case, with the help of the English variant *groyne* and the French equivalent *épi*, the concept can be easily disambiguated.

The case of BANK is similar to that of GROIN. Nevertheless, it is necessary to add other parameters to the linking rule since *bank* is polysemic at a cross-linguistic level. For example, as in English, the Spanish term *banco* can refer to a geographic landform or a financial institution, and there are not many other common multilingual equivalents in DBpedia for disambiguation. In DBpedia, this domain-specific entry is named, and differentiated from others, as BANK (GEOGRAPHY). In order to match this entry and not any of the others, it is necessary to add a context-based rule. Therefore, this match will occur in the following situations: (1) when the word in brackets matches the string of any of our contextual classes or their linguistic variants; (2)

when any term, in any language, associated with any concept belonging to the same contextual class as the search concept appears in one or more of the values of the following properties: *dbpedia-owl:abstract*, *dcterms:subject*, *rdfs:comment,* or *dbpedia-owl:wikiPageRedirects*. In this case BANK in EcoLexicon belongs to the classes, GEOGRAPHY, GEOLOGY and OCEANOGRAPHY, as do many other concepts, such as SHORELINE, ESTUARY, RESERVOIR, SLOPE, RIVER, MARSH, etc, all of which are contained in the properties *dbpedia-owl:abstract*, *dcterms:subject* and *rdfs:comment*. Furthermore, since the disambiguating word in brackets coincides with the EcoLexicon class GEOGRAPHY, the second step is not even required in this case.

Nevertheless, there is a similar but even more complex example in the concept ACCRETION. ACCRETION is polysemic in different languages as well as within the environmental domain. This time disambiguation is not only performed in order to differentiate other domains from the environmental one. On the contrary, three different senses (concepts) in EcoLexicon, designated by the same terms in all languages and with no variants, have to be matched with three out of the five entries in DBpedia. In DBpedia, the term *accretion* may be related to the fields of FINANCE, ASTROPHYSICS, ATMOSPHERE, GEOLOGY, or COASTAL MANAGEMENT, of which only the last three are included in EcoLexicon. In EcoLexicon, the concepts belong to the classes of ATMOSPHERIC SCIENCES, GEOLOGY, and OCEANOGRAPHY, respectively. The concepts related to FINANCE and ASTROPHYSICS are ruled out through the same context-based rule as in BANK. However, this rule must be further specified in order to disambiguate the DBpedia entries of ACCRETION (ATMOSPHERE), ACCRETION (GEOLOGY) and ACCRETION (COASTAL MANAGEMENT). In this case, matching the concepts in common with those included in the property values and those that belong to the same contextual class as each of the concepts designated by *accretion* is insufficient, since all three concepts are closely interrelated. For instance, the key terms *ice* or *droplet*, only present in the property values of ACCRETION (ATMOSPHERE), could seem enough to disambiguate the concept. However, the concepts designated by these terms belong to both our ATMOSPHERIC SCIENCES and GEOLOGY classes. Apart from their obvious relation to the atmosphere, they are also related to geological concepts, such as AVALANCHE or EROSION. Therefore, at this point, disambiguation is still necessary between ACCRETION (GEOLOGY) and ACCRETION (ATMOSPHERE). As for the property values of ACCRETION (COASTAL MANAGEMENT), there are certain terms in the property values, such as *erosion*, *sediment, beach,* and *weather* that can point to all of the three classes (i.e. *weather* to ATMOSPHERIC SCIENCES, *erosion* and *sediment* to GEOLOGY, and *beach* to both GEOLOGY and OCEANOGRAPHY). Consequently, for these cases, one more variable is added to the matching algorithm: from all the contextual classes to which key concepts may belong, only the most frequent one will be used for disambiguation. This means that if most concepts included in the property values of ACCRETION (COASTAL MANAGEMENT) are mostly activated in propositions framed within the OCEANOGRAPHY class, then both concepts are equivalent.

Finally WASTEWATER TREATMENT PLANT does not show ambiguity problems because it is a very specialized concept. Nevertheless, this is a good example of how linking data does not always ensure knowledge acquisition since conceptual modeling does not necessarily follow a concrete pattern in all resources. There is thus no assurance that the content is well structured. The definition of *wastewater treatment plant* in DBpedia does not describe the concept at all. In fact, it is incorrectly assigned

to a disambiguation category, and it redirects users to different types of wastewater treatment. In fact, it does not even offer a proper definition of the plant itself. The Spanish version of Wikipedia has a good entry for its equivalent (*estación depuradora de aguas residuales*), but there is no link between them. In this sense, EcoLexicon could serve as a bridge between the multilingual environmental entries in DBpedia that are not correctly linked.

4 Conclusions

This paper has discussed the importance of multilinguality for the Semantic Web and the problems that can arise when knowledge representations in other languages are not taken into account in the linked data process. More specifically, we have compared the term strings of EcoLexicon's concepts GROIN, BANK, ACCRETION, and WASTEWATER TREATMENT PLANT with those from DBpedia, enhanced by multilingual choices and variants as well as category membership. The results show how valid correspondences can be obtained by taking advantage of the semantics contained in each resource.

References

1. Meij, E., Bron, M., Hollink, L., Huurnink, B., De Rijke, M. Mapping Queries to the Linking Open Data cloud: A Case Study Using DBpedia. Web Semantics: Science, Services and Agents on the World Wide Web (2011)
2. Berners-Lee, T., Hendler, J. and Lassila, O. The Semantic Web. Scientific American (2001)
3. Janev, V. and Vranes, S. Applicability Assessment of Semantic Web Technologies Information Processing and Management vol. 47 pp. 507–517 (2011)
4. Cimiano, P. Towards the Multilingual Semantic Web. Lecture given at the University of Granada, February18, 2011. (2011)
5. Fu, B., Brennan, R. and O'Sullivan, D. Cross-Lingual Ontology Mapping and Its Use on the Multilingual Semantic Web. 1st Workshop on the Multilingual Semantic Web (2010)
6. Faber, P. The Dynamics of Specialized Knowledge Representation: Simulational Reconstruction of the Perception–Action Interface. Terminology, vol. 17, pp. 9-29 (2011)
7. León Araúz, P., Magaña Redondo, P. and Faber, P. Managing Inner and Outer Overinformation in Ecolexicon: an Environmental Ontology. 8th International Conference on Terminology and Artificial Intelligence (2009)
8. Cordeiro, K., Marino., T., Campos, M.L., Borges, M.R.S. Use of Linked Data in the Design of Information Infrastructure for Collaborative Emergency Management System. Computer Supported Cooperative Work in Design (CSCWD) pp. 746-711 (2011).
9. Montiel-Ponsoda, E., Aguado de Cea, G., Gómez Pérez, A. and Peters, W. Enriching Ontologies with Multilingual Information. Natural Language Engineering, pp. 1-27 (2010)
10. Bizer, C., Heath, T. and Berners-Lee, T. Linked Data: Principles and State of the Art (2008)
11. Hassanzadeh, O., Kementsietsidis, A., Lim, L., Miller, R.J. and Wang, M. A Framework for Semantic Link Discovery over Relational Data. CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management. (2009)
12. Dostal, M. and Jezek, K. Automatic Tagging Based on Linked Data: Unsupervised Methods for the Extraction of Hidden Information. Service-Oriented Computing and Applications (SOCA), pp. 1-4 (2010)

Squeezing lemon with GATE

Brian Davis, Fadi Badra, Paul Buitelaar,
Tobias Wunner and Siegfried Handschuh,

Digital Enterprise Research Institute,
National University of Galway, Ireland
{brian.davis, fadi.badra, paul.buitelaar,tobias.wunner,
siegfried.handschuh}@deri.org

Abstract. An increasing number of enterprises are beginning to include ontologies into Text Analytics (TA) applications. This can be challenging for a TA group wishing to avail of such technologies due to the manual effort needed to map language resources within a TA system for a new domain. Ontology lexicalization offers a solution to this problem by seeking to automatically generate lexical resources in order to shrink the manual effort of this concept-to-text mapping process. However, conventional approaches are limited in that they often can only generate term mentions of proper noun, personal noun or fixed key phrases from concept labels in ontologies. Such approaches do not generalize to cope with more complex concept mentions such as nominal compounds or multi-word expressions. An alternative consideration is lemon - Lexicon Model for Ontologies which offers a more sophisticated solution to this problem. We describe a simple use case for exploiting lemon within a widely used open-source TA framework and demonstrate how lemon generated lexical resources are at least comparable in agreement to OntoRootGazeteer, a conventional ontology lexicalization approach.

Keywords: Ontology Lexicalization, NLP frameworks, Semantic Annotation

1 Introduction

An increasing number of enterprises are beginning to include semantic web ontologies into their Information Extraction and Text Analytics process regardless of whether this is to model the application domain or to model the internal data structures of text analytics system itself[1]. The Semantic Web/Linked Data community is also increasingly becoming aware of the need to encode linguistic knowledge concerning concepts directly into ontologies. In this paper we briefly describe lemon – Lexicon Model for Ontologies, which has been developed in the Monnet Project[2] in order to drive a standard for the sharing of lexical information across the semantic web. Furthermore we describe a simple experiment which uses an ontology based on

1 As demonstrated by the recent use of OntoText KIM for the BBC's 2010 World Cup.
2 http://www.monnet-project.eu/

food recipes. We generate a lemon lexicon model using existing services available from the Monnet website. Our goal is to demonstrate the ease of wrapping lemon API as a resource within widely used open-source framework – GATE[3][1]. Furthermore, we exploit lemon generated lexical resources for semantic annotation and provide a preliminary evaluation with promising results. The rest of this paper is structured as follows: Section 2 discusses the lemon model, the OntoRootGazeteer, which is an existing ontology lexicalization tool, distributed with GATE and key related work. Section 3 outlines our use case and implementation of a lemon resource in GATE, for the purpose of generating ontology aware lexical resources for semantic annotation. In Section 3, we compare the lemon approach with GATE's OntoRootGazeteer for observed agreement. Finally, Section 4 offers conclusions and future work.

2 Ontology Lexicalization – Tools and Related Work

Lemon – Lexicon Model for Ontologies

As mentioned earlier, lemon is a model sharing lexical information on the semantic web. Lemon is designed to be a:

- **Concise:** As small number of classes and definitions as needed.
- **Descriptive but not prescriptive**: it uses external sources for the majority of its definitions. A lemon based system can thus be extended in different ways for different tasks i.e. terminological variation, morpho-syntactic description, translation memory exchange.
- **Modular:** Lemon can be separated into a number of modules and it is not necessary to implement the entire lemon model to create a functional lexicon.
- **RDF-native:** Lemon is based on RDF for the purposes of interfacing and sharing across the semantic web. It also permits greater linking between different sections of the lexicon.

A simplest of a lemon entry is as follows:

```
@base <http://www.example.org/lexicon>

@prefix ontology: <http://www.example.org/ontology#>

@prefix lemon: <http://www.monnetproject.eu/lemon#>

:myLexicon a lemon:Lexicon ;

lemon:language "en" ;

lemon:entry :animal .
```

3 GATE - General Architecture for Text Engineering

```
:animal a lemon:LexicalEntry ;

lemon:form [ lemon:writtenRep "animal"@en ] ;

lemon:sense [ lemon:reference ontology:animal ] .
```

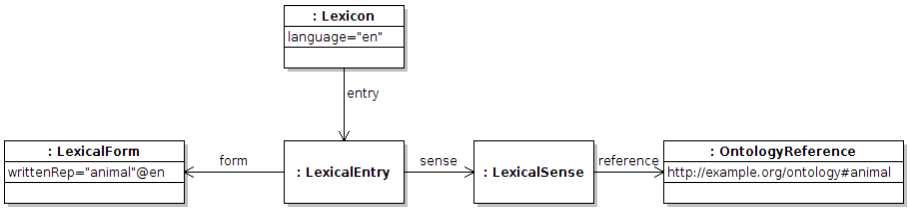


Fig. 1. Sample Lemon Entry visualized. (Extracted from lemon cookbook [4]).

Figure 1. defines the following entities:

1. `Lexicon` : This is the lexicon containing all elements in the lexicon. This approximately corresponds to a SKOS scheme.
2. `Lexical Entry` : This represents the given lexical entry.
3. `Lexical Sense` : Represents the relationship between the lexical entry and the ontology entity.
4. `Reference` : The reference to the resource that can be described by this lexical entry.
5. `Form` : A surface realization of a given lexical entry, typically a written representation

2.2 GATE OntoRootGazeteer

The goal of the GATE OntoRootGazeteer[5] is to produce ontology-based annotations i.e. annotations by pre-processing an ontology in order to extract human-understandable lexicalisations. The OntoRootGazeteer initially extracts all the names of ontology resources within a given ontology as well assigned property values for all ontology resources (e.g., label and data-type property values). Further processing involves replacing any name containing dash ("-") or underline ("_") character(s) with a blank space. In addition, built-in GATE lemmatizers and POS tagging resources are exploited in order to create the proper lemma for a given resource name. Finally an in-memory ontology aware gazetteer is created.

[4] http://www.monnet-project.eu/Monnet/resource/Monnet-Website/0000%20%20Library/0700%20-%20Downloads/lemon-cookbook.pdf. Accessed 15th August 2011

[5] http://gate.ac.uk/sale/tao/splitch13.html#x18-33900013.9

1.3 Additional Related Work

With respect to lemon, it is influenced strongly by Lexical Markup Framework- LMF [2] , which is part of the ISO TC37/SC4[6] working group on the management of Language Resources. LMF has its origins in language engineering standardization initiatives such as EAGLES[7] and ISLE[8] . With respect to in depth literature on lemon and its historical influences, we recommend [3] and [4]. The LIR (Linguistic Information Repository) model is similar in many respects, but focuses strongly on multilingualism [5]. Finally, there is OntoLing, which is a Protégé plug-in that allows for linguistic enrichment of ontologies[6].

3. Implementation and Experiment: Squeezing Lemon with GATE

3.1 Experimental Use Case

In our use case, we utilize an ontology of food recipes which contains a over four and half thousand classes of food ingredient. Currently it is unpopulated with instance data. We took the first one hundred concepts in the ontology for testing purposes. Our goal was not for scalability testing with respect to ontology storage but rather to test agreement between lemon generated lexical resources with those of the OntoRootGazeteer as well the ease of importing lemon as a resource into GATE.

3.2 LemonGazeteerGenerator PR

Using the online lemon generator[9], we uploaded our sample ontology to generate a a lexical model file in turtle[10] format. We wrote a small application using the lemon API, to iterate through all written representations for each given concept. For each unique concept in the lexicon mode, it creates a gazetteer list, which is a simple text file with lexical entries such as: *quail eggs and quail egg.* The application also writes an entry to a mapping definitions (See Figure 2) file which aligns ontology resources with gazetteer list entries. Finally, we wrapped the application as a GATE processing resource (PR) to promote language resource reuse (See Figure 3).

```
quail.lst:file:///home/bridav/Fadi/FoodOntologySmall.owl:Quail
avena.lst:file:///home/bridav/Fadi/FoodOntologySmall.owl:Avena
golden_raisin.lst:file:///home/bridav/Fadi/FoodOntologySmall.owl:Golden_raisin
foie_gras_entier.lst:file:///home/bridav/Fadi/FoodOntologySmall.owl:Foie_gras_e
```

6 http://www.tc37sc4.org/
7 http://www.ilc.cnr.it/EAGLES96/browse.html
8 http://www.mpi.nl/ISLE/
9 http://monnetproject.deri.ie/lemonsource
10 http://www.w3.org/TeamSubmission/turtle/

```
ntier
chanterelle.lst:file:///home/bridav/Fadi/FoodOntologySmall.owl:Chanterelle
```

Fig 2. *Mapping.def* : Contains alignments between gazetteer lists and ontology class URIs.

Once the following gazetteer list files and mapping file are created they can be exploited by the GATE OntoGazeteer resource, which is a hierarchical hash gazetteer for

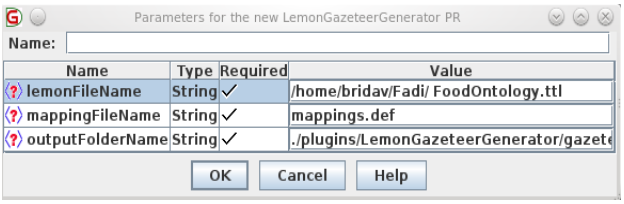


Fig 3. GATE LemonGazeteerGenerator PR, which takes a LemonModel turtle file as input and produces an ontology-aware gazetteer lists.

semantic annotation of concept mentions in text. However it differs from the OntoRootGazeteer in that it does not automatically lexicalize an ontology rather it follows traditional knowledge engineering approaches whereby the ontology must be manually aligned to lexical resources. In general it is used for small to medium sized ontologies where accuracy is critical, however for much larger ontologies it becomes unmanageable, hence the automatic approach using the OntoRootGazeteer. Note in our use case the OntoGazeteer is exploiting automatically generated lexical resources produced by our LemonGazeteerGenerator PR. Using existing GATE processing resources such as: a tokeniser, part of speech tagger and lemmatiser, we created a IE pipeline for semantic annotation. The same pipeline was reused with the OntoRootGazetteer to create annotations for agreement comparison. Recall the OntoRootGazeteer generates its own lexicalizations from the same food recipe ontology used by the LemonGenerator service.

3.3 Experimental Results

Using a small test corpus contain over 4650 lines of food recipes, we compared both the lemon generated OntoGazeteer and conventional OntoRootGazeteer.
As we do not at this time have a gold standard annotation set as a baseline, we only record observed agreement across both methods. Of the 798 annotations created by the OntoRootGazeteer, the LemonOntoGazeteer matched 74 % of annotations' spans,. Of those matches, 91% were in agreement with ontological concepts. Upon closer observation we noticed two issues:

1. The LemonGenerator web service appeared to have lexicalized only leaf node concepts in the food recipes ontology while the OntoRootGazeteer had traversed and lexicalized the entire graph. While this may seem a disadvantage. However, depending on the annotation task, it could be advantageous as an optional feature and thus benefit the user.

2. There were some unexpected errors in the LemonGazeteerGenerator PR in the form of some erroneous mappings. So for example, in addition to a mapping for `Fruit_Juice`, a mapping for a concept `Fruit` was also created. This may be a bug in either the PR or the LemonGenerator Service itself.
However despite these short comings, the lemon lexicalizations for the concepts mapped, were upon inspection correctly generated.

4 Conclusions and Future Work

In this paper, we have described initial experiments towards using lemon for ontology lexicalization. We demonstrated how easily lemon resources can be exploited by a well known TA framework. We compared a lemon based hierarchical gazetteer with the OntoRootGazeteer, a conventional ontology lexicalization tool available in GATE. We found that the results that results are at least comparable. The reader should note that we do not exploit the inherent multilingualism of lemon, nor its richer lexicalization features, which are not available to the OntoRootGazetteer. Future work will focus on exploiting the full power of the lemon model, improving the output of the lemon generator web-service as well as a more thorough evaluation.

Acknowledgments. : The work presented in this paper was supported (in part) by the European project MONNET No. (FP7/2007-2013) 248458 and (inpart) by the Lìon 2 project supported by Science Foundation Ireland under Grant No. SFI/08/CE/I1380

References

1. Cunningham, H., et al: Text Processing with GATE (Version 6). University of Sheffield Department of Computer Science. (2011). ISBN 0956599311.
2. Francopoulo, G. George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., Soria, C. : Lexical Markup Framework (LMF).In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC). Genoa, Italy, (2006).
3. McCrae, J, Spohr D, Cimiano P. : Linking Lexical Resources and Ontologies on the Semantic Web with lemon. Proceedings of the 8th Extended Semantic Web Conference (ESWC). Heraklion, Crete, (2011).
4. McCrae, J, Aguado-de-Cea G, Buitelaar P, Cimiano P, Declerck T, Gomez-Perez A, Gracia J, Hollink L, Montiel-Ponsoda E, Spohr D et al.: In Press. Interchanging lexical resources on the Semantic Web. Language Resources and Evaluation
5. Montiel-Ponsoda E, Aguado de Cea G, Gómez Pérez A, Peters W: Enriching Ontologies with Multilingual Information. Natural Language Engineering, (2010).
6. Pazienza,M., T., Stellat, A.: An Environment for Semi-automatic Annotation of Ontological Knowledge with Linguistic Content. In 3rd European Semantic Web Conference (ESWC 2006) Budva, Montenegro, (2006).

Accessing Multilingual Data on the Web for the Semantic Annotation of Cultural Heritage Texts

Karlheinz Moerth[1], Thierry Declerck[1,2], Piroska Lendvai[3], Tamás Váradi[3]

[1] ICLTT, Austrian Academy of Sciences, Sonnenfelsgasse 19/8,
1010 Wien, Austria

[2] DFKI GmbH, Stuhlsatzenhausweg, 3
66123 Saarbrücken, Germany

[3] HASRIL, Hungarian Academy of Sciences, Benczúr u. 33.
H-1068 Budapest, Hungary
Karlheinz.moerth@oeaw.ac.at,
declerck@dfki.de,
piroska@nytud.hu,varadi@nytud.hu

Abstract. Our study targets interoperable semantic annotation of Cultural Heritage or eHumanities texts in German and Hungarian. A semantic resource we focus on is the Thompson Motif-index of folk-literature (TMI), the labels of which are available only in English. We investigate the use lexical data on the Web in German and Hungarian for supporting semi-automatic translation of TMI: lexical resources offered by Wiktionary accessed via the Lexvo service, and discuss shortcomings of those resources. An approach for mapping the XML dump of Wiktionary onto a TEI and MAF compliant data is presented, whereby we discuss improvements in the representation of Wiktionary data for exploiting its multilingual value within the LOD framework.

Keywords: Multilinguality, LOD, Cultural Heritage, Semantic Annotation

1 Introduction

In the context of a cooperation between the Austrian and the Hungarian Academies of Sciences we investigate the possibility to generate interoperable and multilingual semantic annotation of Cultural Heritage or eHumanities texts. One of the semantic resources we consider for this task is the Thompson Motif-index of folk-literature (TMI)[1] [5], which contains around 36,000 terms, cataloguing typical narrative content of folk tales and myths from around the world. The terms, or 'labels', of the classification system are available only in English.

Our general hypothesis is that converting resources such as TMI into a LOD compliant combination of multi-layered linguistic annotation and their taxonomic classes can support the automatic detection and semantic annotation of motifs in literary work, across genres and languages.

[1] An electronic version of TMI is available at: http://www.ruthenia.ru/folklore/thompson/

A motif is an element conveying an idea or theme e.g. in film or music, but also in folklore or scientific texts[2]. Motifs are cognitively complex notions expressed in lexically and syntactically highly variable but compact structures. Linguistic features of motifs have so far not been systematically investigated, but these have been exposed and aim to be worked out by the authors of this paper, in collaboration with the international AMICUS network[3], with a clear motivation for enhanced indexing and modelling of cultural heritage data (cf. [1], [3] and [4]).

The TMI catalog focuses on motifs that emphasize ideas or themes. For example, "*K3. Substitute in contest*" is one motif in TMI (its parent node being "*K0-K99. Contests won by deception*", subsumed under *"K. Deceptions*"). Dozens of subtypes are assigned to this single motif; these catalogue descriptions, or labels, are short phrases such as "*Supernatural substitute in tournament for pious warrior*", "*Wise man disguised as monk beats learned heretic in debate*". The TMI lists 23 main categories[4] and provides a deep hierarchical structure of motifs.

To semantically annotate texts in German and Hungarian with this resource, we aim to enrich TMI with German and Hungarian labels. Our strategy consists in providing first for the linguistic annotation of the phrasal heads detected in the English labels[5], and to try to find equivalent lexical entries in German and Hungarian retrieved from online multilingual lexical resources.

2 Access to Online Lexical Resources in the LOD

The scarcity of freely available professional on-line multilingual lexical data made us turn to the lexical resources offered by the collaborative dictionary project Wiktionary, and the access provided to within the Lexvo service[6], which has been deployed within the Linked (Open) Data (LOD) framework[7]. Some observations we could make on this combination of resources are described in this section.

We noted first that in Wiktionary, variants of an entry (e.g. singular or plural form), often do not feature identical sense or translation information.[8] It is necessary to link those entries into a consistent unit, and to use an appropriate model for this. Two candidates can be considered for this modeling: ISO-LMF[9] and *lemon*

---

[2] Some random examples for motifs in folk tales are e.g. the cruel stepmother, the poor girl who was chosen as wife in preference to a rich one, or a supernatural who substitutes the hero in a tournament.

[3] http://amicus.uvt.nl

[4] E.g. Animal Motifs, Magic, the Dead, Marvels, Tests, the Wise and the Foolish, Deceptions, Reversals of Fortune

[5] The details of this linguistic analysis are described in a submission currently under review.

[6] http://www.lexvo.org/

[7] http://linkeddata.org/

[8] One example is the English Wiktionary entry "creator" (http://en.wiktionary.org/wiki/creator), which lists the basic morpho-syntactic information, associated senses and translations whereas the entry "creators" (http://en.wiktionary.org/wiki/creators#English) only states that it is the plural of "creator".

[9] http://en.wikipedia.org/wiki/Lexical_Markup_Framework

(developed in the Monnet project and related to the W3C community)[10]. An advantage of the *lemon* approach would be that one could represent the Wiktionary data in the RDF format, making Wiktionary data available in the Linked Data framework. Nevertheless, as a first step we ported the XML dump of Wiktionary into a TEI[11] and MAF[12] compliant format (see Section 3).

Lexvo is a service that "brings information about languages, words, characters, and other human language-related entities to the Linked Data Web and Semantic Web"[13]. Lexvo points to Wiktionary entries, displaying for each word that can be queried (in a variety of languages) a link to senses that are encoded either in the LOD version of WordNet[14] or/and of OpenCyc[15], but in those versions the senses are available only for English entries. Since the the Wiktionary data is not yet available in a machine-readable format, Lexvo cannot display the senses available in the resource. This is an additional argument for porting Wiktionary to RDF. Due to the same reason, linguistic information associated to each word in WIktionary cannot be made available in Lexvo. A Lexvo specific shortcoming is the fact that it refers only to the English version of Wiktionary, regardless of entries that are in fact written in other languages, ignoring as a consequence several pieces of language-specific information.

3 Porting Wiktionary to a Standardised Representation

Our starting point is the XML dump[16] of Wiktionary. Nevertheless, the data do not really deliver what one might expect from xml data, namely well-formed structured information. The content is formatted making use of a lightweight markup system which is used in different Wiki applications, and is neither standardized (various applications use considerably divergent forms of the wikitext language) nor truly structure-oriented. It is designed in a format-oriented manner to be transformed into HTML.

Our initial goal was to transfer these data into an XML format suitable for further processing. Although, as mentioned above, we consider ISO-LMF and *lemon* as the final candidates, for pragmatic reasons, we eventually opted for TEI p5[17] as our

[10] http://greententacle.techfak.uni-bielefeld.de/drupal/sites/default/files/lemon-cookbook.pdf and [8].
[11] http://www.tei-c.org/index.xml
[12] http://lirics.loria.fr/doc_pub/maf.pdf
[13] http://www.lexvo.org
[14] http://semanticweb.cs.vu.nl/lod/wn30
[15] http://sw.opencyc.org
[16] http://dumps.wikimedia.org
[17] As the TEI p5 dictionary module was conceptualized as the digital representation of printed dictionaries, it appears not to be the most natural candidate for the task at hand. However, the main motive behind adopting the dictionary module of this “de facto” text encoding standard was that ongoing lexicographic projects of the ICLTT had yielded tools to process this kind of data. Besides an online dictionary editor geared towards the particular needs of TEI, there are also a number of thoroughly tested XSLT stylesheets to visualize the particular kind of data. A second reason, equally important, is the fact that the ICLTT’s dictionary working group has been working recently on a TEI dictionary schema suitable for use in NLP applications.

starting point. While several attempts at preparing Wiktionary for use in NLP applications have been made before [2, 5, 7], the tool we present here is – to our knowledge – the first such application targeting TEI p5, and the first such tool provided with a graphical user interface.

The actual conversion process is carried out in three main steps. Each of these steps can be performed separately, which allows the interested user to pursue the transformation process in detail.

First, the comparatively large database dump (287 MB) was split into manageable smaller chunks. This process resulted in a collection of roughly 85000 entries.

In the second phase of the conversion, the top-level constituents of these entries were identified and transformed into XML elements. This task turned out to be pretty straightforward as the *entries* (we stick to traditional lexicographic nomenclature here) display a rather flat hierarchical structure. The resulting chunks each contain a particular type of data, the main constituents of the dictionary entries. The number of constituent parts varies with the size of the individual entries (from 3KB up to 338KB). In the result sets, there are chunks containing grammatical data such as for instance part of speech. There are chunks containing etymological information and/or usage information. Many entries contain morphological data, in numerous cases complete inflectional paradigms. The files also hold data concerning hyphenations of word forms and their pronunciation. However, the central concern of our work here has been semantic data. This kind of information is stored in sections describing the various meanings of words. These, in turn, are linked to translations, synonyms, antonyms, hyperonyms, hyponyms, and often to examples.

The last step in the transformation process has been the conversion of the above described constituents into TEI p5. Iterating through all the untyped chunks, the program attempts to identify the right category and subsequently to translate it into TEI p5. At this point, the main challenge for the programmer was the merging of data on the same hierarchical level (e.g. meanings and translations) into neatly nested XML structures. Successful data conversion depends largely on the quality of the underlying markup. While many errors can be compensated by some trickery in the program, inconsistencies remain.

The actual tag set applied in our project can also be seen as a contribution aiming at developing the TEI guidelines towards an encoding system suitable to be used in NLP applications.[18] We will not go into the gory details of modeling TEI documents here, just one small digression: one particularly useful module of the TEI p5 guidelines was the chapter on *feature structures*. This mechanism allowed us to model the representation of the morpho-syntactic data in accordance with the MAF standard (Morpho-syntactic Annotation Framework, ISO TC 37). Canonical TEI for inflected word forms such as *gingst* "(you) went" usually look like this:

[18] An initiative towards this end was the workshop *Tightening the representation of lexical data, a TEI perspective* at the TEI's members meeting this year in Würzburg (Germany).

```
<form>
    <orth>gingst</orth>
    <gramGrp>
      <gram type="pos">verb</gram>
      <gram type="number">plural</gram>
      <gram type="person">2</gram>
      <gram type="tense">preterite</gram>
      <gram type="mood">indicative</gram>
    </gramGrp>
</form>
```

We tried to encode such structures in a more MAF-like manner, which is still TEI conformant:

```
<form ana="#v_pret_ind_pl_p2"><orth>gingst</orth></form>
```

In this encoding scheme, the morpho-syntactic identifiers used in the *ana* atribute of the form element is defined as a set of TEI conformant feature structures. The values used here refer to a feature value library, which is also linked to the ISO data categories.

Although the conversion tool already works quite nicely, a number of issues registered in its requirement specification remain to be solved. It goes without saying that the first thing that comes to mind, is the issue of other languages, which is on top of our agenda. First candidates for this are English and French.

The second issue is moving on to LMF which is a project reaching far beyond our Wiktionary tool. Creating LMF data from TEI is something apparently non-trivial.

One other important task to be achieved in the near future is setting up a service delivering the data. First steps towards implementing a restful server have been taken. We hope that by the time this paper is presented, our TEI version of the German-language Wiktionary will be up and running.

4 Further work on porting Wiktionary to the Semantic Web

Although our work represents a step in making the full Wiktionary information available for NLP applications, it is not sufficient to represent links between entries (for example, one entry being the plural of the other, etc), or to make this information available in the Web or in the LOD and so to establish links between entries and senses in Wiktionary, WordNet or OpenCyc, on the one, but also between TMI and LOD data sets on the other hand. Just to name an example: In TMI the concept "A0: Creator" is the upper class of a large number of (hierarchically ordered) terms. We collected all the head nouns of those terms, and can build so a kind of domain specific "WordNet". This list of nouns is for sure very different and more complex than what we find in WordNet or OpenCyc. We need a way to relate the semantic organization

of TMI and WordNet/OpenCyc (or other data sets), also on the base of linguistic information we can find in the (Semantic) Web. There is therefore a need to port both Wiktionary and the analyzed labels of TMI onto a LOD compliant RDF. For this we are getting also advices from the Monnet project[19].

Acknowledgments

Part of the described in this paper is supported by the R&D project "Monnet", which is co-funded by the European Union under Grant No. 248458, and by the AMICUS network, which is sponsored by a grant from the Netherlands Organization for Scientific Research, NWO Humanities, as part of the Internationalization in the Humanities programme

References

1. Declerck, T., K. Eckart, Lendvai, P., L. Romary, T. Zastrow (2010a). Towards a Standardised Linguistic Annotation of Fairy Tales. In: Proc. of the LRT standards workshop at LREC-2010.
2. Krizhanovsky, A. (2010). The comparison of Wiktionary thesauri transformed into the machine-readable format. (http://arxiv.org/abs/1006.5040)
3. Lendvai, P., Declerck, T., S. Darányi, P. Gervás, R. Hervás, S. Malec, F. Peinado (2010a). Integration of Linguistic Markup into Semantic Models of Folk Narratives: The Fairy Tale Use Case. In: Proceedings of the Seventh International conference on Language Resources and Evaluation, Pages 1996-2001, Valetta, Malta, European Language Resources Association (ELRA).
4. Lendvai, P. (2010). Granularity Perspectives on Modeling Humanities Concepts. In: S. Darányi, P. Lendvai, (eds.). First International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts, Vienna, Austria. University of Szeged, Hungary.
5. Navarro, E., Sajous, F., Gaume, B., Prévot, L., Hsieh, S.-K., Kuo, T.-Y., Magistry, P., Huang, C.-R. (2009). Wiktionary and NLP: Improving synonymy networks. In: Proceedings of the 2009 Workshop on Peoples's Web Meets NLP, ACL-IJCNLP. Singapore: pp. 19-27.
6. Thompson, S. (1955-58). Motif-index of folk-literature: A classification of narrative elements in folktales, ballads, myths, fables, medieval romances, exempla, fabliaux, jest-books, and local legends. Revised and enlarged edition. Bloomington, Indiana University Press.
7. Zesch T., Mueller C., Gurevych I. (2008). Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In: Proceedings of the Conference on Language Resources and Evaluation. LREC 2008.
8. McCrae, J, Aguado-de-Cea G, Buitelaar P, Cimiano P, Declerck T, Gomez-Perez A, Gracia J, Hollink L, Montiel-Ponsoda E, Spohr D et al.. In Press. Interchanging lexical resources on the Semantic Web. Language Resources and Evaluation 2011.

[19] http://www.monnet-project.eu

XSLT Conversion between XLIFF and RDF

Dimitra Anastasiou
SFB/TR8
Computer Science/Languages Science
University of Bremen
Bremen, Germany
anastasiou@uni-bremen.de

Abstract. This paper focuses on the conversion between the open standard XML Localisation Interchange File Format (XLIFF) and the Resource Description Framework (RDF). XLIFF is a localisation standard supported by proprietary and free and open source software (FOSS) localisation tools, while the latter is a standard model, basic ingredient in Semantic Web. We developed a converter based on Saxon XSLT Processor which translates XLIFF to RDF.

Keywords: Conversion, Localisation, Semantic Web, Standards.

1 Introduction

Generally speaking, standards incorporate a solid body of knowledge and provide a unified framework. In addition, when metadata is standardised, resources can be identified, catalogued, and processed faster and more efficiently. Although standards as such are a benefit for information management, in the last years we have seen too many standards evolving in information science. In our opinion, the existence of too many standards in tandem with their inflexible structure (of some standards) adds complexity and leads to lack of interoperability; interoperability between Web resources is crucial for communication between application components.

This paper focuses on XLIFF[1] and RDF[2] and the conversion based on Saxon from the former to the latter. Our work is motivated by the insight that Web resources should be multilingual and XLIFF as a localisation standard is capable to help localise ontologies and thus create multilingual linked data. A wider target range of users and applications will then be reached. The automatic conversion from XLIFF into RDF can be used as an API both by localisation tools and Semantic Web applications.

In section 2 we describe some related work about combining multilinguality with Semantic Web. In sections 3 and 4 some examples of XLIFF and RDF are provided. Section 5 discusses the XLIFF-RDF interoperability and then we conclude the paper.

[1] http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xliff, 12/09/11
[2] http://www.w3.org/RDF/, 12/09/11

2 Related Work

In 2004 [1] stated that Human Language Technology faces new multilingual and multicultural challenges for the Semantic Web and presented relevant ongoing initiatives. One year later, [2] pointed out the usefulness of a multilingual Semantic Web, particularly to help translate websites through the use of ontologies, manage group knowledge in multilingual form, and create international communication base for industry and commerce. [3] used the Universal Networking Language (UNL) as a step between the process of acquiring knowledge from textual sources and translating it into one of the state-of-the-art knowledge representation formalisms for building multilingual ontologies.

The Multilingual Semantic Web workshop started in 2010 and continues with annual workshops; the same holds for the XLIFF International Symposium. Some research projects: the Multilingual Web[3], Flarenet[4], META-NET[5], and Monnet[6] see the symbiotic relationship between multilingual resources and Semantic Web.

As far as the conversion between XLIFF and other standards is concerned, the Okapi Framework provides XLIFF conversion utilities, e.g. to Translation Memory eXchange (TMX). [4] describes how to convert documents to XLIFF and back to the original format through text extraction, pre-translation, translation, reverse conversion, and translation memory improvement. A framework which combines many localisation standards is the MultiLingual Information Framework (MLIF) [5]; an overview about localisation standards can be found in [6]. A model that has been proposed to associate linguistic data to ontologies is the 'Linguistic Information Repository' (LIR) [7], designed to account for cultural and linguistic differences among languages. Lemon[7] is another model sharing lexical information on the Semantic Web; noteworthy is the converter between lemon and the Lexical Markup Framework (LMF).

Our main motivation for XLIFF2RDF conversion is the concept of 'ontology localization', a term coined by [8]: "*Ontology Localization is the adaptation of an ontology to a particular language and culture*". [9] state that ontology localisation is an activity with both pragmatic and economic goals. The former can be seen in the fostering reuse of ontologies already available for the domain in question instead of building them from scratch, and the latter, a result of the former, is seen in the stage of cost reduction compared to building a completely new ontology.

3 XLIFF

XLIFF is an open localisation standard supported by proprietary and FOSS localisation tools. It is under the auspices of OASIS and is understood by many

[3] http://www.multilingualweb.eu/en, 12/09/11
[4] http://www.flarenet.eu/, 12/09/11
[5] http://www.meta-net.eu/, 12/09/11
[6] http://www.monnet-project.eu/Monnet/Monnet/English?init=true, 12/09/11
[7] http://lexinfo.net/, 12/09/11

actors: software providers, localisation service providers, and localisation tools providers. Semantic localisation metadata is very important in a localisation workflow to distinguish between the responsibilities of each stakeholder (project manager, engineer, translator, proofreader), between translatable and non-translatable content, annotate (in the case of translatable content) the status of the strings and so on. Particularly in software localisation, coordinates of menus dialogue boxes, version control, count of screenshots belong to the most important metadata. The following example contains an XLIFF file with three translation units (TUs). TU elements include a <source>, <target> and associated elements.

```
1.  <?xml version="1.0" encoding="UTF-8" ?>
2.  <xliff version="1.2" xmlns="urn:oasis:names:tc:xliff:document:1.2">
3.  <file original="minimal_XLIFF.html" source-language="en-us" target-language="de-de"
    datatype="html">
4.  <body>
5.  <trans-unit id="#1">
6.  <source>book</source>
7.  <target>Buch</target>
8.  </trans-unit>
9.  <trans-unit id="#2">
10. <source>book publisher</source>
11. <target>Buchverlag</target>
12. </trans-unit>
13. <trans-unit id="#3">
14. <source>This book is good!</source>
15. <target>Dieses Buch ist gut!</target>
16. </trans-unit>
17. </body>
18. </file>
19. </xliff>
```

Example 1. XLIFF file with three translation units. *Line 1*: XML declaration, *Line 2*: XML schema, *Line 3*: file metadata, *Lines 5-16*: file data (three TUs).

4 RDF

RDF is family of W3C specifications which describe Web resources. Here is a brief explanation of *Resource*, *Property*, and *Property value* by means of the XLIFF Ex.1:

- A *Resource* is anything that can have a URI, e.g. `minimal_XLIFF.html`;
- A *Property* is a Resource that has a name, such as `trans-unit`, `source`;
- A *Property value* is the value of a Property, such as `This book is good!`

The example 1 can be represented in an RDF graph as follows:

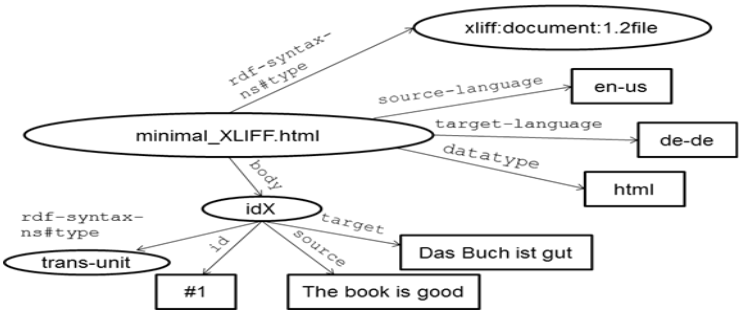


Diagram 1. RDF graph of Example 1

Accordingly, every XLIFF file can be represented in an RDF graph. The circles are the *resources,* the labels on the arrows are the *properties*, and the content of the rectangles are the *property values*. `idX` is a placeholder for a resource representing the `body`.

Building a bridge for interoperability between RDF and other standards is something common: WSDL-RDF, RDF-Topic Maps, OWL-RDF, and others. However, these standards, which RDF can be converted from and into, also come from the Semantic Web world and not from the localisation scene.

As far as the representation of multilingual information in RDF is concerned, RDF used the RFC 3066 standard (published in 2001) for language tags for literals in natural languages. The revision RFC3066bis included productive use of language, country and script codes. [10] suggested a small change to the RDF model theory to permit access to the language tag in the formal semantics, giving this ontology a precise formal meaning; their approach defined a new property called `rdflg:lang`.

5 Interoperability

The greatest contribution of XLIFF is the nature of its content, i.e. the capture of translation pairs, rather than the formalisation vehicle of the knowledge, be it XML or RDF. We do not intend to reify XLIFF, but to make XLIFF portable to RDF. The reasons why an XLIFF2RDF mapping and conversion are useful follow:

i. Any file format which can be converted into XLIFF can be then converted to RDF;
ii. RDF ontology labels can be translated using XLIFF;
iii. Web resources can be described by XLIFF metadata.

A practical implementation of standards' interoperability between XLIFF and RDF(S) is distinguished between two parts: mapping XLIFF elements and attributes to RDF and automatically converting from XLIFF into RDF. The mapping of three XLIFF files has been described in [11]. In order to cover more than three use cases, automatic conversion is needed. We created different types/use cases of XLIFF files and accordingly incremental EXtensible Stylesheet Language Transformations (XSLTs) to translate various XLIFF files: a file with 3 translation units, with file processing metadata, with alternative translations, a document containing two files, and a modularised file containing a lot of metadata and inline markup.

A sample of an XSLT follows:

```
1.  <xsl:stylesheet version="1.0"
    xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
    xmlns:a="urn:oasis:names:tc:xliff:document:1.2"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:xliff="http://docs.oasis-open.org/xliff/xliff-core/xliff-core.html#">
2.  <xsl:template match="/">
3.  <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
4.  <xliff:file>
5.  <xsl:attribute name="rdf:about">
6.  <xsl:value-of select="a:xliff/a:file/@original"/>
7.  </xsl:attribute>
8.  <xsl:attribute name="source-language">
9.  <xsl:value-of select="a:xliff/a:file/@source-language"/>
```

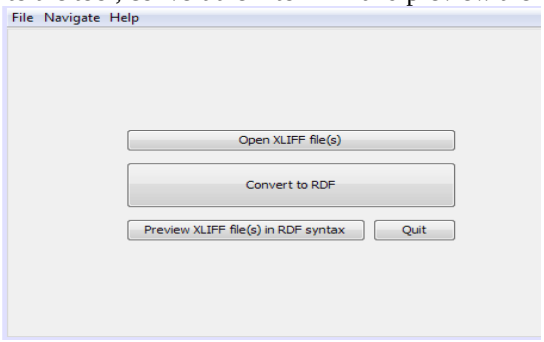
Example 2. Sample of the XSLT

It should be mentioned that there is discrepancy between interoperability between data based on standards and interoperability between standards. Conversion between

standards plays a small part within the wider scope of interoperability which includes, among others, supporting relevant standards and conforming with specifications.

5.1 Converter

The development of a conversion tool to translate from XLIFF into RDF automates and thus accelerates the process. We used NetBeans IDE to create a GUI of the conversion tool (see Screenshot 1). For our conversion utilities we used the Saxon home edition 9.3 version[8]. The home edition is an open source product available under the Mozilla Public License. It provides implementations of XSLT 2.0, XQuery 1.0, and XPath 2.0 and is available for both Java and .NET. The user can input one or more XLIFF file(s) to the tool, convert them to RDF and preview them.



Screenshot 1. XLIFF2RDF conversion tool

The converter is under Google code hosting[9] website. There users can freely get a local copy of the tool or create their own clone.

6 Discussion and Conclusion

In this paper we discussed the interoperability between the localisation standard XLIFF and RDF. We showed ongoing initiatives, projects, and tools combining multilinguality with Semantic Web. We developed a converter from XLIFF to RDF by using and adapting the Java API of the XSLT processor Saxon. We wrote some sample XLIFF files and adopted a modular transitional file provided in the XLIFF latest specifications in order to create corresponding XSLTs.

In our opinion, localisation is often regarded only as a business strategy to increase return on investment and not as a research field which can both enrich and gain from the Semantic Web and Linked Data. Localisation standards and particularly XLIFF has received little attention although it covers many actors' needs.

In Semantic Web context, it is an arbitrary decision in which natural language the ontology labels are provided, and thus many researchers see the need for multilingual ontologies; challenges, like cross-lingual mapping and translation follow the existence

[8] http://saxon.sourceforge.net/, 12/09/11
[9] http://code.google.com/p/xliff-rdf/, 28/03/11

of multilingual ontologies. Our conversion tool is a contribution to build a bridge between localisation and Semantic Web resources, so that localisation tools can localise ontologies and Semantic Web resources are populated with localisation-related metadata. After the XLIFF2RDF conversion, metadata can be reused in the Semantic Web to represent multilingual ontologies. The XLIFF2RDF conversion tool is hosted on Google code hosting website. There other users can freely get a local copy of the tool; thus replication of the tool is allowed. The conversion tool fulfills its basic requirements, i.e. XLIFF files are represented in RDF. Not only minimal XLIFF examples with one TU, but with more TUs and also with file processing metadata, alternative translations, etc. can be successfully converted. Five use cases have been successfully tested, however more quantitative and qualitative examples are planned to be converted. We plan to extend the conversion API for other standards. At first place, we plan to translate from XLIFF into OWL. Also interoperability between other localisation and internationalisation standards is also among future prospects. In terms of quality assurance, existing validation tools will be part of our tool.

Acknowledgment. We gratefully acknowledge the support of the Deutsche Forschungsgemeinschaft (DFG) through the Collaborative Research Center SFB/TR 8 Spatial Cognition - Subproject I5-DiaSpace.

References

1. Declerck, T., Buitelaar, P., Calzolari, N. & Lenci, A. Towards A Language Infrastructure for the Semantic Web. *Proceedings of LREC* (2004)
2. Hahn, W. and Vertan, C. Challenges for the Multilingual Semantic Web. *Proceedings of the International MT Summit X* (2005)
3. Cardeñosa, J., Gallardo, C., Iraola, L., & De la Villa, M. A New Knowledge Representation Model to Support Multilingual Ontologies. A Case study. *Proceedings of the International Conference on Information and Knowledge Engineering*, 313-319 (2008)
4. Raya, R. XML Localisation Interchange File Format as an intermediate file format. *IBM developerWorks* (2004) http://www.maxprograms.com/articles/xliff.html
5. Cruz-Lara, S., Bellalem, N. Ducret, J. & Kramer, I. Standardizing the management and the representation of multilingual data: the MultiLingual Information Framework. *International Workshop on Language Resources for Translation work, Research and Training*, 35--38 (2006)
6. Anastasiou, D., Morado Vázquez, L. Localisation Standards and Metadata. *Proceedings of the 4th Metadata and Semantics Research Conference (MTSR 2010)*, Communications in Computer and Information Science, Springer, 255--276 (2010)
7. Peters, W., Montiel-Ponsoda, E. & Aguado de Cea, G. Localizing Ontologies in OWL. *Proceedings of the ISWC07 OntoLex workshop* (2007)
8. Suarez-Figueroa, C., M. and Gomez-Perez, A. First attempt towards a standard glossary of ontology engineering terminology. *Proceedings of the 8th International Conference on Terminology and Knowledge Engineering* (2008)
9. Cimiano, P., Montiel-Ponsoda, E., Buitelaar, P., Espinoza, M., & Gomez-Perez, A. A note on ontology localization. *Journal of Applied Ontology* (JAO), 5(2), 127--137 (2010)
10. Carroll, J.J., & Phillips, A. Multilingual RDF and OWL. *The Semantic Web: Research and Applications, Lecture Notes in Computer Science*, Vol. 3532/2005, 15--19 (2005). doi: 10.1007/11431053_8
11. Anastasiou, D. XLIFF Mapping to RDF. *JIAL (The Journal of Internationalisation and Localisation)*, to appear (2011)