Language as a Foundation of the Semantic Web

Gerard de Melo
Max Planck Institute for Informatics
Saarbrücken, Germany
demelo@mpi-inf.mpg.de

Gerhard Weikum
Max Planck Institute for Informatics
Saarbrücken, Germany
weikum@mpi-inf.mpg.de

ABSTRACT

This paper aims to show how language-related knowledge may serve as a fundamental building block for the Semantic Web. We present a system of URIs for terms, languages, scripts, and characters, which are not only highly interconnected but also linked to a great variety of resources on the Web. Additional mapping heuristics may then be used to derive new links.

1. INTRODUCTION

Language is the basis of human communication and the key to the tremendous body of written knowledge available on the Web. To date, however, the language domain remains strongly underrepresented on the Semantic Web. In what follows, our efforts to address this highly significant issue shall be described. We define URIs for language-related entities and link them to a multitude of resources on the Web.

2. LANGUAGE INFRASTRUCTURE

The first step consists of establishing the basic infrastructure for referring to language-related entities.

2.1 Languages, Scripts, and Characters

The ubiquitous two-letter ISO 639-1 codes for languages ('*en*', '*fr*', etc.) are defined for no more than around 180 languages. While the slightly more recent ISO 639-2 standard provides around 500 three-letter codes and hence covers the major languages of the world, it cannot by any means be considered complete, lacking codes for Ancient Greek, American Sign Language, and of course thousands of rare minority languages spoken around the world. The same holds for URIs derived from the English Wikipedia, which merely describes a few hundred languages.

To address this situation, we have created URIs of the form `http://www.lexvo.org/id/iso639-3/eng` for all of the 7 000 languages covered by the ISO 639-3 standard. For each URI, background information about the language from several sources is provided, for instance language names in many languages, geographical regions (using URIs based on ISO 3166 / UN M.49), identification codes, relationships between languages, etc.

The language URIs are linked to URIs that have been set up for the scripts defined by the ISO 15924 standard. Examples include Cyrillic, Indian Devanagari, and the Korean Hangul system. By extracting Unicode Property Values from the Unicode specification, these script URIs have also been connected with the specific characters that are part of the respective scripts.

URIs of the form `http://www.lexvo.org/id/char/5A34` are provided for each of the several thousand characters defined by the Unicode standard. A large number of Unicode code points represent Han characters used in East Asian languages. We have extracted additional data from the Unihan database and other sources to provide semantic information about such characters.

2.2 URIs for Terms

String literals cannot serve as subjects of an RDF triple. For expressing lexical knowledge, several ontologies have defined OWL classes that represent words or other terms in a language. However, the URIs for individual terms are often created on an ad hoc basis. For instance, the W3C draft RDF/OWL Representation of WordNet [3] has defined URIs for the words covered by the WordNet lexical database [2].

We propose a standard, uniform scheme for referring to terms in a specific language[1]. Given a term t in a language L, the URI is constructed as follows:

- The term t is encoded using Unicode, and the NFC normalization procedure [1] is applied to ensure a unique representation. Conventional unnormalized Unicode allows encoding a character such as '*à*' in either a composed or in a decomposed form.
- The resulting Unicode code point string is encoded in UTF-8 to obtain a sequence of octets.
- These octet values are converted to an ASCII path segment by applying percent-encoding as per RFC 3986. Unacceptable characters as well as the '*%*' character are encoded as character triplets of the form '*%4D*' with the respective octet value stored as two uppercase hexadecimal digits.
- The base address `http://www.lexvo.org/id/term/` as well as the ISO 639-3 code for the language L followed by the '/' character are prepended to this path segment to obtain a complete URI.

[1]Different levels of abstraction exist. When considering terms in a specific language, we do not distinguish the meanings of polysemous words in a language, e.g. the verb and noun meanings of the English term '*call*'. In contrast, we do consider the Spanish term '*con*', which means '*with*', distinct from the French term '*con*', which means '*idiot*'.

Capturing links to terms is particularly significant in light of the important role of natural language for the Semantic Web. In general, a non-information resource URI string itself does not convey reliable information about its intended meaning, because an URI (including class or property names) can be chosen quite arbitrarily. Oftentimes the meaning is specified using natural language definitions or characteristic labels. Formally, however, RDFS `label` is merely an annotation property that provides human-readable display labels, which can be identifier strings such as '*minCardinality*'.

In order to make the meaning of URIs more formal, we propose explicitly linking to term URIs of one or more natural languages using a lexicalization property, whenever appropriate. Such a property formally captures the semantic relationship between a concept and its natural language lexicalizations or between an arbitrary entity and natural language terms that refer to it.

2.3 RDF Service

Our language infrastructure is backed by an RDF Web service that makes our URIs dereferenceable. The service relies on HTTP content negotiation with 303 redirects to provide RDF or HTML representations of basic knowledge about the URIs. For instance, any term URI is linked to the respective language URI.

3. INTEGRATING ADDITIONAL DATA

The value of the infrastructure is greatly increased by importing from and linking to other resources in the spirit of the ongoing Linked Data endeavours proposed by Tim Berners-Lee.

- **Princeton WordNet** [2] is likely to be the most commonly used lexical resource for natural language processing. We consider the current WordNet 3.0 and link from each term to the respective WordNet synsets.
- **Wiktionary** is an effort to collaboratively create dictionaries. The individual language-specific sites each contain a wealth of multilingual lexical information, but do not share common formatting conventions and unfortunately are catered for human use rather than machine processing. We have implemented information extractors for several Wiktionaries.
- **Wikipedia** is a very well-known collaboratively authored encyclopedia. We use article names, redirects, etc. to connect millions of terms in many languages to the respective Wikipedia pages.
- **DBpedia and YAGO:** DBpedia is an effort to make information from Wikipedia available in a more machine-processable form. YAGO is an ontology derived from Wikipedia and WordNet. We establish links from term entities to the respective DBpedia and YAGO entities.
- **Upper Ontologies:** The Suggested Upper Model Ontology (SUMO) is a formal ontology that, unlike most OWL ontologies, provides an extensive set of axioms for the entities it defines. OpenCyc is a large taxonomy related to the Cyc knowledge base but lacking rules and axioms. We imported links from English terms to entities in SUMO and OpenCyc.
- **Thesauri:** A thesaurus captures terminological information about a domain using taxonomical and other relations. We link from term entities to the respective concepts in the GEneral Multilingual Environmental Thesaurus (GEMET), the United Nations FAO AGROVOC thesaurus, the US National Agricultural Library Thesaurus, and several others.

4. AUTOMATIC MAPPING LINKS

Additional links between data sources can be created automatically using appropriate mapping heuristics. We currently apply the following scoring model: Given two entities $x \in S_1$, $y \in S_2$ from two different data sources S_1, S_2, we consider the respective sets of linked terms $T(x)$, $T(y)$, and compute

$$
\begin{aligned}
m_1(x,y) &= \frac{1}{2} \sum_{t_1 \in T(x)} \sum_{t_2 \in T(y)} \mathrm{sim}(t_1, t_2) \\
m_2(x,y) &= \frac{1}{\sum_{t_1 \in T(x)} \sum_{y' \in S_2} \max_{t_2 \in T(y')} \mathrm{sim}(t_1, t_2)} \\
&\quad \text{or 0 if denominator} < 1 \\
m_3(x,y) &= \frac{1}{\sum_{t_2 \in T(y)} \sum_{x' \in S_1} \max_{t_1 \in T(x')} \mathrm{sim}(t_1, t_2)} \\
&\quad \text{or 0 if denominator} < 1 \\
m(x,y) &= \alpha_1 m_1(x,y) + \alpha_2 m_2(x,y) m_3(x,y)
\end{aligned}
$$

Here, α_1, α_2 are weighting parameters both set to 0.5, and $\mathrm{sim}(t_1, t_2)$ is a string similarity metric. In our case, it is simply 1 if the two strings are equal after establishing converting to lower case and establishing Unicode NFCK [1], and 0 otherwise. The score $m(x, y)$ then constitutes a conservative estimate of how much evidence we have for a match between x and y. For instance, we obtained around 12,000 mappings between AGROVOC and the NAL Thesaurus with a sampled precision of 0.969 ± 0.027 (95% Wilson interval).

5. CONCLUSIONS

We have presented a new infrastructure that brings language knowledge to the Semantic Web, thereby addressing the need for unique references to linguistic entities such as languages, scripts, characters, and terms. We have argued why we believe such knowledge will constitute an important foundation of the Semantic Web. The data is strongly linked to existing resources on the Web, including wordnets, thesauri, ontologies, and versions of Wikipedia and Wiktionary, and can be used to derive additional equivalence links between entities. Further details about the data may be obtained at `http://www.lexvo.org`.

6. REFERENCES

[1] M. Davis and M. Dürst. Unicode normalization forms, rev. 29. Technical report, Unicode, 2008.

[2] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May 1998.

[3] M. van Assem, A. Gangemi, and G. Schreiber. RDF/OWL Representation of WordNet. W3C Working Draft, World Wide Web Consortium, June 2006. http://www.w3.org/TR/wordnet-rdf/.