Feature Representation for Cross-Lingual, Cross-Media Semantic Web Applications

Paul Buitelaar♦, Michael Sintek, Malte Kiesel♣

DFKI GmbH - ♦Language Technology Lab & ♣Knowledge Management Dept.
Saarbrücken/Kaiserslautern, Germany

{paulb,sintek,kiesel}@dfki.de

Abstract Currently, ontology development has been mostly directed at the representation of domain knowledge (i.e., classes, relations and instances) and much less at the representation of corresponding text and image features. To allow for cross-media knowledge markup, a richer representation of features is needed. At present, such information is mostly missing or represented only in a very impoverished way. In this paper we propose an RDF/S-based ontology for the integrated representation of domain knowledge and text/image features.

1 Introduction

Ontologies define the semantics for a *set of objects* in the world using a *set of classes*, each of which may be identified by a particular *symbol* (either linguistic, as image, or otherwise). In this way, ontologies cover all three sides of the "semiotic triangle" that includes *object, referent* and *symbol*, i.e., an *object* in the world is defined by its *referent* and represented by a *symbol* (Ogden and Richards, 1923 – based on Peirce, de Saussure and others).

Currently, ontology development and the Semantic Web effort in general have been mostly directed at the *referent* side of the triangle, and much less at the *symbol* side. To allow for cross-media knowledge markup, a richer representation is needed of these symbols, i.e. of the text and image features for the object classes that are defined by the ontology. At present, such information is mostly missing[1] or represented only in a very impoverished way, leaving the semantic information in an ontology without a grounding to the human cognitive and linguistic domain.

[1] According to the collection of ontologies available through OntoSelect (see Buitelaar et al., 2004) currently only about 9% of ontologies represent multilingual terms for classes and/or properties (http://views.dfki.de/ontologies/index.php?mode=stats)).

2 Cross-Lingual, Cross-Media Feature Extraction and Representation

An ontology describes a knowledge model of a particular domain of discourse at a particular point of time and is shared between two or more actors in the domain. As the ontology defines the agreed semantics of the domain, all relevant content will be marked-up with knowledge according to the ontology. The definition of the ontology in turn depends primarily[2] on the content that has already been interpreted. Accordingly, content production and interpretation will drive the adaptation of the ontology infrastructure, and ontology adaptation will drive content interpretation and production. In order to arrive at such a continuous 'hermeneutic cycle' of content and knowledge production and interpretation, a rich representation of domain knowledge and content features is needed. Here we propose an integrated approach that organizes content and knowledge in several layers, as displayed below:

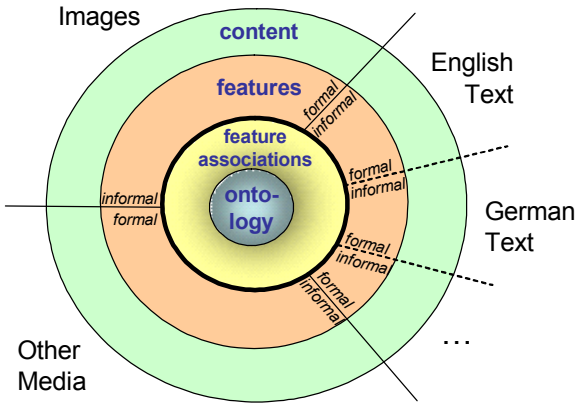

Figure 1: Interacting Layers in Feature Extraction and Representation

The *content layer* (outermost layer) consists of cross-media data (images, video and/or mixed image and text documents).

The *features layer* (1st inner layer) consists of extracted features for the data in the content layer. For multilingual data, this ranges from comparatively informal feature vectors gathered by use of statistical methods to formalized descriptions of the content of text documents, typically extracted by use of natural language processing and information extraction methods. For multimedia data, this will be mostly limited to informal features as used in colour histograms and similar.

[2] Aside from more generic knowledge of the physical world, time, space, etc. that will be inherited from an upper-level ontology.

The *feature association layer* (2nd inner layer) consists of feature representations occurring in the features layer. While in the *features layer* features are associated with cross-media data, in the *feature association layer* the features are associated with classes in the semantic model.

The *semantic model layer* (central layer) consists of classes, with which the data in the content layer is to be interpreted (i.e., annotated) by use of the extracted and represented features in the *features layer* and the *feature association layer*.

This integrated approach allows for cross-lingual, cross-media feature extraction and representation as follows:

image2text – For instance, if we know which terms express a class in English, we will be able to build a classifier for the classification of images that occur in the context of English terms for this class.

text2image – For instance, if we know which images represent instances for a specific class, we will be able to extract German terms for this class from surrounding German text.

text2text – For instance, if we know which terms express a class in English, and we know the context features (i.e. words) for these terms and possible translations for these words into German, we will be able to build a cross-lingual classifier for recognition of unseen German terms for this class.

image2class or text2class – For instance, if we know which terms express a class in English, and we know the context words for these terms, we will be able to detect a change in the semantic model for this class by monitoring any change in the context words, and similar with image feature models.

3 Towards Ontology-Based Feature Representation

The integrated ontology-based feature representation we propose is based on ongoing work in the context of the SmartWeb project on mobile Semantic Web access for intelligent information services in the football domain (http://www.smartweb-project.de/). To represent terminology for concepts in different languages we initiated an extension of RDF-based domain knowledge representation with the meta-class `ClassWithFeats`.

Although there is some overlap with the SKOS (Miles and Brickley, 2005) model for RDF-based thesauri, the proposed representation is richer as it will include not only multilingual terms for classes (and properties) but also context models for disambiguating these terms in knowledge markup (i.e., "world cup" as EVENT or ARTIFACT). More specifically, there is a technical and a conceptual reason why SKOS[5] does not fulfill the needs of our scenario: SKOS uses sub-properties of `rdfs:label` (`skos:prefLabel`, `skos:altLabel`) together with `xml:lang` to attach multilingual terms to concepts. Furthermore, the RDFS specification (Brickley and Guha, 2004; Hayes, 2004) defines the range of `rdfs:label` to be `rdfs:Literal`. From the definition of `rds:subPropertyOf` follows that the range of `skos:prefLabel` and `skos:altLabel` is also `rdfs:Literal` (or a specialization of `rdfs:Literal`). This is not sufficient in our scenario since we want to attach more information as linguistic information to classes than simple multilingual strings. This led to our decision to use the meta-class `Class-WithFeats`, which allows us to attach complex information to classes with the properties `lingFeat` and `imgFeat` (in the future, more properties will be defined for other media types like audio and video).

The conceptual problem we see with SKOS for the use in our scenario is that it mixes linguistic and semantic knowledge. SKOS uses `skos:broader` and `skos:narrower` to express "semantic" relations without clearly stating the semantics of these relations intentionally, and defines the sub-properties `skos:broaderGeneric` and `skos:narrowerGeneric` to have class subsumption semantics (i.e., they inherit the `rdfs:subClassOf` semantics from RDFS). We clearly keep the linguistic and semantic, ontology-based knowledge representations apart[6]: the ontology is represented using the semantic relations defined in RDFS or OWL (Full)[7] (McGuinnes and van Harmelen, 2004), and attach linguistic knowledge to the classes (and properties).

We further propose to integrate image-related features in this representation, which is beyond the scope of SKOS. Note that SKOS uses `foaf:depiction`, `skos:prefSymbol`, and `skos:altSymbol` to attach images to concepts, but not complex feature descriptions.

[5] Our argumentation applies to all approaches based on `rdfs:label` and `xml:lang` to attach multilingual labels to classes and relations.

[6] Note that our approach in effect integrates a domain-specific multilingual Wordnet into the ontology, although also the Wordnet model does not distinguish clearly between linguistic and semantic information (Miller et al., 1995). Alternative lexicon models that are more similar to our approach include (Bateman et al., 1995) and (Alexa et al., 2002), but these concentrate on the definition of a top ontology for lexicons instead of text/image features for domain ontology classes and properties as in our case.

[7] OWL Lite and OLW DL do not support meta-classes and meta-properties.

4 Application

The proposed feature representation is currently used in the SmartWeb ontology on sport events and related issues. Figure 2 shows the ontology with example classes and associated linguistic and image features: the ontology contains the class `o:FootballPlayer` with subclasses `o:Defender` and `o:Midfielder`. All these classes are instances of the meta-class `feat:ClassWithFeats` which allows them to use the feature-association properties `feat:lingFeat` and `feat:imgFeat`. The figure shows the linguistic features of German terms for the class `o:Defender` (*Abwehrspieler*) and `o:Midfielder` (*Mittelfeldspieler*). Note that the decomposition of *Abwehrspieler* contains *Spieler*, therefore implicitly relating the two classes in the ontology via linguistic features. Furthermore, the figure shows an image feature representation associated with the class `o:Midfielder`, stating that an instance of this class has a 'human' shape and certain color and texture features.

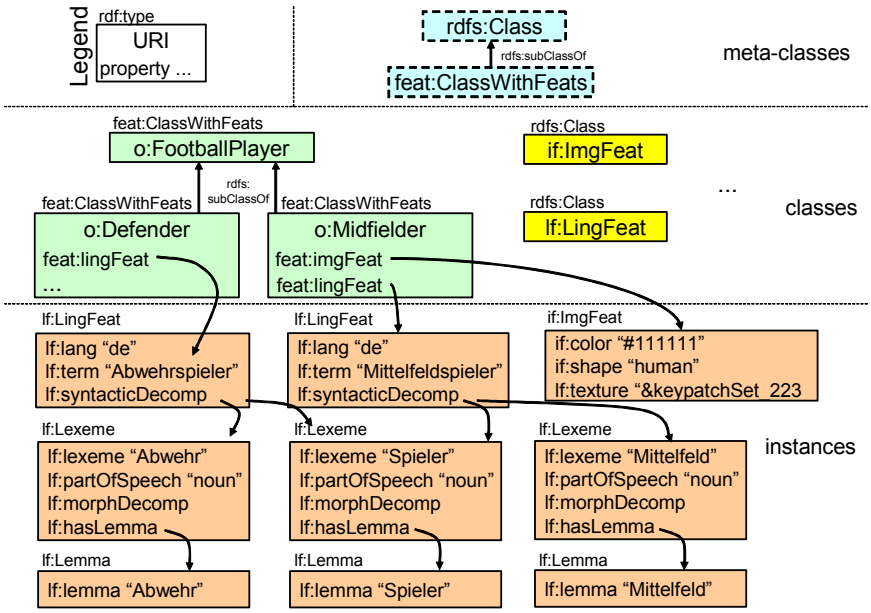

Figure 2: Ontology and Examples – *Defender*, *Midfielder* – of Domain Knowledge, Text (Linguistic) and Image Features (simplified)

Acknowledgements

This research has been supported in part by the SmartWeb project, which is funded by the German Ministry of Education and Research under grant 01 IMD01 A. We acknowledge input on the representation of image-related features by Yannis Avrithis, Eric Gaussier and Yiannis Kompatsiaris.

References

M. Alexa, B. Kreissig, M. Liepert, K. Reichenberger, L. Rostek, K. Rautmann, W. Scholze-Stubenrecht, S. Stoye *The Duden Ontology: an Integrated Representation of Lexical and Ontological Information* In: Proc. of the OntoLex Workshop at LREC, Spain, May 2002.

J. A. Bateman, R. Henschel and F. Rinaldi *Generalized Upper Model 2.0: documentation* Report of GMD/Institut für Integrierte Publikations- und Informationssysteme, Darmstadt, Germany, 1995.

D. Brickley, R.V. Guha, editors. *RDF Vocabulary Description Language 1.0: RDF Schema.* World Wide Web Consortium, 2004. `http://www.w3.org/TR/rdf-schema/`

P. Buitelaar, Th. Eigner, Th. Declerck *OntoSelect: A Dynamic Ontology Library with Support for Ontology Selection* In: Proc. of the Demo Session at the International Semantic Web Conference, Hiroshima, Japan, Nov. 2004.

P. Buitelaar and S. Ramaka *Unsupervised Ontology-based Semantic Tagging for Knowledge Markup* In: Proc. of the Workshop on Learning in Web Search at the International Conference on Machine Learning, Bonn, Germany, August 2005.

Th. Declerck, O. Vela, Z. Gantner and D. Manzano-Macho *Esperonto Deliverable 5.2: Multingualism and Ontologies* Dec. 2004

P. Hayes, editor. *RDF Semantics.* World Wide Web Consortium, 2004. `http://www.w3.org/TR/rdf-mt/`

D.L. McGuinness, F. van Harmelen, editors. *OWL Web Ontology Language Overview.* W3C Recommendation 10 February 2004. `http://www.w3.org/TR/owl-features/`

A. Miles, D. Brickley, editors. *SKOS Core Vocabulary Specification.* W3C Working Draft 10 May 2005. `http://www.w3.org/TR/swbp-skos-core-spec/`

A. Miles, D. Brickley, editors. *SKOS Core Guide.* W3C Working Draft 10 May 2005. `http://www.w3.org/TR/swbp-skos-core-guide/`

G. A. Miller *WORDNET: A Lexical Database for English*. Communications of ACM (11): 39-41, 1995.

Ch. K. Ogden and I. A. Richards *The meaning of meaning - A study of the influence of language upon thought and of the science of symbolism.* London: Kegan Paul, Trench, Trubner & Co., 1923.

K. Petridis, I. Kompatsiaris, M. G. Strintzis, S. Bloehdorn, S. Handschuh, S. Staab and N. Simou *Knowledge Representation for Semantic Multimedia Content Analysis and Reasoning* In: Proc. of the European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology, Royal Statistical Society, London, 25-26 Nov. 2004.