Integrating a large domain ontology of species into WordNet

Montse Cuadros[†], **Egoitz Laparra**[‡], **German Rigau**[‡], **Piek Vossen**[*], **Wauter Bosma** [*]

[†] TALP center, Universitat Politècnica de Catalunya, Barcelona, Catalonia
[‡] IXA NLP Group, University of the Basque Country, Donostia, Basque Country
[*] Dept. Letteren. Vrije Universiteit. Amsterdam. Netherlands
cuadros@lsi.upc.edu, {egoitz.laparra, german.rigau}@ehu.es, {p.vossen, w.bosma}@let.vu.nl

Abstract

With the proliferation of applications sharing information represented in multiple ontologies, the development of automatic methods for robust and accurate ontology matching will be crucial to their success. Connecting and merging already existing semantic networks is perhaps one of the most challenging task related to knowledge engineering. This paper presents a new approach for aligning automatically a very large domain ontology of Species to WordNet in the framework of the KYOTO project. The approach relies on the use of knowledge-based Word Sense Disambiguation algorithm which accurately assigns WordNet synsets to the concepts represented in Species 2000.

1 Introduction

Ontology alignment has been recognized as a major issue in the semantic web community (van Hage, 2008). On the Semantic Web (Maedche and Staab, 2001), data is structured by means of ontologies which describe the semantics of the data. In this scenario, data is represented by many different ontologies. However, information processing across ontologies is not possible without knowing corresponding mappings between them. Manually finding such mappings is tedious, not systematic, and clearly not possible with large-scale ontologies representing large collections of content data.

Due to the importance of the problem, many works have addressed ontology mapping using a variety of matching heuristics, e.g. (McGuinness et al., 2000), (Noy and Musen, 2001), (Rodriguez and Egenhofer, 2003). Recently, the Relaxation Labelling algorithm and structural constraints has been integrated successfully in a multi-strategy process for mapping ontologies (Daudé et al., 2000), (Doan et al., 2002).

There is also a meta-approach to ontology integration. The Linking Open Data Project (Bizer et al., 2008), launched by the W3C, aims to interlink existing ontologies. It encourages people to make RDFS/OWL data sets available online as Web services. On top of these Web services, it establishes links between equivalent concepts in different data sets.

Our work has been carried out in the framework of the KYOTO project[1] (Vossen et al., 2008). The goal of KYOTO is the construction of a system for facilitating the exchange of information across cultures, domains and languages. This system will allow people in communities to define the meaning of their words and terms in a shared Wiki platform. Domain terms will be anchored across languages and cultures to a common ontology that will allow a computer to use this knowledge to detect knowledge and facts in text. The system is being developed for the domain of environment. For example, the notion of environmental *migration* will become defined in the same way in all these languages. With these definitions it will be possible to find information on *migration* in documents, websites and reports so that users can directly ask the computer for actual information in their environment.

Thus, the KYOTO platform operates as a Wiki for establishing semantic interoperability across languages for a specific domain by creating domain wordnets that get interlinked through a shared knowledge base. The resulting semantic knowledge base is further used to apply automatic fact mining on document collections. The platform allows for continuous updating and modeling of the vocabulary by the people in the community, while their domain wordnets remain anchored to a generic wordnet, and to a common ontology. This architecture can be seen as a first attempt to implement the Global Wordnet Grid (GWG) on a practical scale for specific domains. In the GWG, all wordnets are anchored to a shared ontology (Fellbaum and Vossen, 2007), (Pease et al., 2008), (Fellbaum and Vossen, 2008).

In order to extend the coverage of the linguistic processors and knowledge tools of the KYOTO platform, we decided to extend the current vocabulary by integrating the Species2000 ontology as a domain extension of WordNet3.0. Species2000 is a very large ontology of around two million species.

The rest of the paper is organized as follows. After this short introduction, Section 2 presents the KYOTO system. In Section 3, we describe the KYOTO knowledge architecture, and in Section 4 the Species2000 ontology. Section 5 presents the automatic mapping of the Species2000 to WordNet3.0. A preliminary evaluation and error analysis is reported in sections 6 and 7. Finally, Section 8 draws some general conclusions and sketch future work directions.

2 KYOTO system

The KYOTO project pursues to help communities to model terms and concepts in their domain and to use this knowledge to apply text mining on documents. The knowledge cycle in the KYOTO system starts with a set of source documents of interest by the community, such as PDFs and

[1] http://www.kyoto-project.eu

websites. Linguistic processors apply tokenization, segmentation, morpho-syntactic analysis and some semantic processing to the text in different languages. The semantic processing involves detection of named-entities (persons, organizations, places, time-expressions) and determining the meaning of words in the text using a given wordnet in a language.

The output of this linguistic analysis is stored in an XML annotation format that is the same for all the languages, called the KYOTO Annotation Format (KAF, (Bosma et al., 2009)). This format incorporates standardized proposals for the linguistic annotation of text but represents them in an easy to use layered structure. In this format, the linguistic information of words, terms, constituents, syntactic dependencies is structured and stored in separate layers with references across the structures. This makes it easier to harmonize the output of different linguistic processors for different languages and to add new layers (mainly semantic) to the basic output, when needed (Bosma et al., 2009). All modules in KYOTO draw their input from these structures. For instance, the word-sense-disambiguation (WSD) (Agirre and Soroa, 2009) and named-entity recognition and classification (NERC) processes are carried out on the same KAF annotation in different languages and is therefore the same for all the languages (Agirre and Soroa, 2009). Both semantic processors use wordnet synsets to provide semantic interpretations to the terms occurring in the text. In the current system, there are processors for English, Dutch, Italian, Spanish, Basque, Chinese and Japanese.

The KYOTO system proceeds in two cycles (see Figure 1). In the first cycle, the **Tybot** (Term Yielding Robot) extracts the most relevant terms from the analysed documents. The Tybot is another generic program that can do this for all the different languages in much the same way. The terms are organized as a structured hierarchy and, wherever possible, related to generic semantic databases, i.e. wordnets for each language. In Figure 1, italic terms occur in the text, and underlined terms are not found in wordnet. Straight terms are hyperonyms in wordnet that do not necessarily occur in the text but are linked to ontological classes. The domain experts can view the terms in the term database and edit them using **Wikyoto** (Ronzano et al., 2010), i.e. adding or deleting terms, changing their meaning, adding definitions, changing relations, etc.

The result is a domain wordnet in a specific language. New terms can be also seen as possible candidates to extend the ontology if some fundamental semantic properties, like Rigidity (Guarino and Welty, 2004), apply. Through the ontology, the domain experts can establish the similarities and differences across the languages and hence cultures.

The second cycle of the system involves the actual extraction of factual knowledge from the annotated documents by the **Kybots** (Knowledge Yielding Robots). Kybots use a collection of profiles that represent patterns of information of interest. In the profile, conceptual patterns are modeled through the domain knowledge (wordnets and ontology) by means the so-called expression rules. Since the semantics is defined through the ontology, it is possible to detect similar information across documents, even if expressed differently, or expressed in different languages. In Figure 1, we give an example of a conceptual pattern that relates organisms that live in habitats. The Kybot can combine this pattern with words from the wordnet and morpho-syntactic structures. When a match is detected, the instantiation of the pattern is saved in a formal representation, either in KAF or in RDF. Since the wordnets in different languages are mapped to the same ontology and the text in these languages is represented in the same KAF, similar patterns can easily be applied to multiple languages.

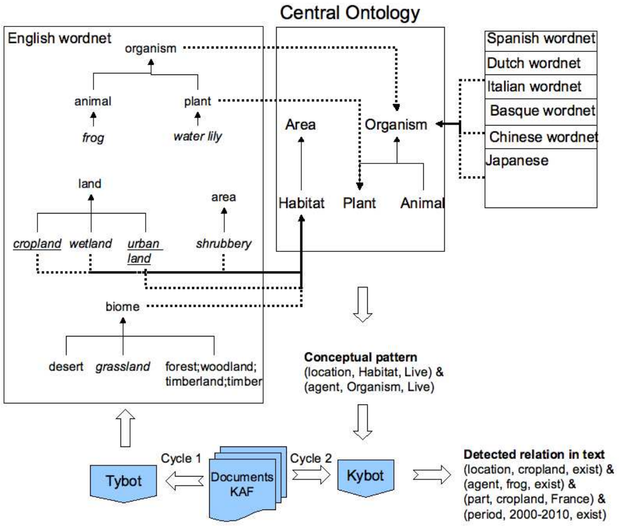

Figure 1: Two Cycles of processing in KYOTO

The main goal of the KYOTO project is to develop a knowledge sharing and transition platform that can be used by communities in the world. The KYOTO platform operates as a Wiki for establishing semantic interoperability across languages for a specific domain by creating domain wordnets that get interlinked through a shared knowledge base. The resulting semantic knowledge base is further used to apply automatic fact mining on document collections. The platform allows for continuous updating and modeling of the vocabulary by the people in the community, while their domain wordnets remain anchored to a generic wordnet, and to a common ontology. This architecture can be seen as a first attempt to implement the Global Wordnet Grid (GWG) on a practical scale for specific domains. In the GWG, all wordnets are anchored to a shared ontology (Fellbaum and Vossen, 2007), (Pease et al., 2008), (Fellbaum and Vossen, 2008).

Obviously, a large ontology as a language independent representation of meaning holds many promises for future research and usage provided that it is tightly connected to the wordnets used in the project. Universalia and idiosyncracies of lexicalizations in language can be expressed in a systematic way, allowing language-independent reasoning over linguistically expressed knowledge. If successful, the GWG can be built by the massive labour force of the Internet community and the results become available to the global community.

3 KYOTO knowledge architecture

When applying the principle of Global WordNet Grid (Fellbaum and Vossen, 2008) to a specific domain, numerous practical and fundamental problems to handle the domain data arise.

First of all, existing background knowledge, such as Species2000, should be integrated into the domain knowledge base since they are often maintained outside the wordnet community, without connecting their resources to the wordnet infrastructure.

Secondly, other new terms are automatically learned from the documents and web sites used in the community. Both background knowledge and domain terminology need to be aligned with existing generic wordnets to make the domain wordnet interoperable with general concepts. (Vossen and Rigau, 2010) describe the KYOTO approach for integrating all these resources in a useful and unique knowledge repository. The proposed solution has three different layers with different types of links between them that support different types of inferencing.

The amount and complexity of the KYOTO knowledge repository is enormous. The Global Wordnet Grid architecture suggests that the wordnets extended with the domain vocabulary are anchored through the domain extension of the ontology. In practice this means, that the ontology needs to be extended with millions of new concepts. For example, the KYOTO ontology needs to make a distinction between taxonomic groups and individual organisms. Instances of species are members of a taxonomic group and instances of an organism. Likewise, we can predict that if an instance of a frog ceases to exist, it is not implied that the taxonomic group Anura ceases to exist but only an instance of the organism Anura. The former is only the case when all members of Anura cease to exist. As a consequence, the ontology that represents all species in this domain should include all 2.1 million species twice (!), once as group and once as a type of organism.

Such a model leads to various practical problems. First of all, ontologies of that size cannot be loaded in any existing inferencing system. Inferences as the above can thus not be made because of the size of such an ontology. Another problem is that the vocabularies are linguistically too complex and diverse. Whereas the species can be considered as rigid concepts, as defined by (Guarino and Welty, 2002), this is not the case for most of the terms that are learned from the document collection. In the environment domain, the documents typically include terms for roles of species rather than the species as such, e.g. invasive species, migration species, threatened species. For mining facts from documents, these non-rigid role terms have more information value than the defining properties of the species.

For a knowledge sharing system as modeled by the Global Wordnet Grid, it is thus more important to precisely define what the roles and processes are in which species participate than to provide the defining properties of the species as such. Likewise, we propose a model of division of knowledge along the lines of the division of linguistic labor defined by (Putnam, 1975). Putnam argues that linguistic communities rely on the fact that experts know the defining properties of natural kind terms such as gold and can thus determine which instances of matter are gold and which are not. Most natural language users therefore have a shallow definition of what gold is and can still use this definition to communicate valuable information on gold, such as for trading gold or buying jewelry.

Along the same lines, we propose a digital version of this principle, where we state that a computer does not need to know the defining properties of each rigid concept but can rely on the capacity of the domain expert to determine what the instances are of, for example, a particular species. Vast amounts of words for rigid concepts can likewise remain in the vocabularies as long as we indicate their status as rigid concepts.

For instance, the KYOTO knowledge architecture distinguishes:

- instances, like ”Humber Estuary” represented by a wikipedia article[2] or DBpedia [3]
- concepts from wordnets, like <estuary_1> having the definition ”the wide part of a river where it nears the sea; fresh and salt water mix”
- ontological types (like estuary-eng-3.0-09274500-n).

The KYOTO knowledge model assumes that the terminology from the domain text corpus is merged with a generic wordnet in a language so that the domain terms are anchored to more general terms and concepts. This requires that the term hierarchy for the domain is somehow disambiguated to match specific word meaning from the generic wordnet. Once the term hierarchy is aligned with a generic wordnet, existing mappings from wordnet to ontologies can be used to apply the ontological distinctions to the domain terms. Named entities are more likely to be found in other resources such as Wikipedia, DBPedia and GeoNames. This requires another alignment operation, where the concepts in the external sources need to be matched to wordnet as well and through wordnet to the ontology. The situation becomes more complex when existing domain thesauri and taxonomies are added to the knowledge base. Modeling the vocabulary and concepts in a domain is a complex knowledge integration problem.

Furthermore, the following knowledge repositories are relevant as a background knowledge for the environment domain in KYOTO:

- Generic wordnets in each language ranging from 50,000 to 120,000 synsets.
- A term databases with about 500,000 terms extracted from about 1,000 documents in each language.
- Existing ontologies such as the EuroWordNet top-ontology (Vossen, 1998), SUMO (Pease et al., 2002) and DOLCE (Gangemi et al., 2003).
- GEMET (GEneral Multilingual Environmental Thesaurus): a core multilingual terminology for the environment[4]

[2] http://en.wikipedia.org/wiki/Humber

[3] http://dbpedia.org/page/Humber

[4] http://www.eionet.europa.eu/gemet

- Wikipedia: over 3 million articles in English and large volumes in other languages, by September 2009[5].
- DBPedia: 2.6 million things and 274 million pieces of information (RDF triples), by September 2009[6].
- GeoNames: 8 million geographical names and 6.5 million unique features whereof 2.2 million populated places and 1.8 million alternate names, by September 2009[7].
- The Species 2000 database with 2.1 million species, having taxonomic relations and labels in many different languages[8].

In Figure 2, we show an example of the three layers of the KYOTO model. We include in the vocabulary vast quantities of species obtained from Species 2000. The species hierarchy is partially linked to a generic wordnet (Toral et al., 2010). In addition, terms from the term database are mapped to the most specific synset as well. The wordnet synset hierarchy can be traversed to find the most specific Base Concept that is matched to the ontology. In this way, we can infer for all species in the vocabulary that they are both members of a taxonomic group and rigid subtypes of organism.

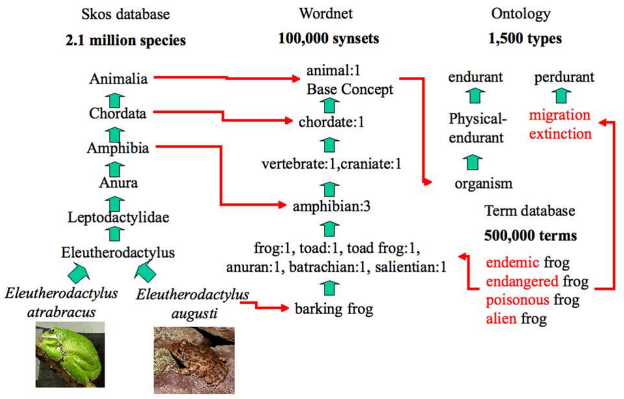

Figure 2: Division of knowledge over three layers

The wordnets for seven working languages of KYOTO have been represented in the Wordnet-LMF format (Soria et al., 2009) and stored in a DebVisDic server (Horák et al., 2006). The DebVisDic server also contains the SUMO ontology and a first version of the KYOTO ontology in OWL-DL. The SUMO ontology is fully mapped to WordNet3.0. The KYOTO ontology (version 1) consists of 786 classes divided over three layers. The top layer is based on DOLCE (DOLCE-Lite-Plus version 3.9.7, (Gangemi et al., 2003)) and OntoWordNet. This layer of the ontology has been modified for our purposes (Hicks and Herold, 2009). The second layer consists of concepts coming from the so-called Base Concepts in various wordnets (Vossen, 1998), (Izquierdo et al., 2007). Examples of base concepts are: *building, vehicle, animal, plant, change, move, size, weight*. The Base Concepts (BCs) are those synsets in WordNet3.0 that have the most relations with other synsets in the wordnet hierarchies and are selected in a way that ensures complete coverage of the nominal and verbal part of WordNet. This has been completed for the nouns (about 500 synsets) and is currently being carried out for verbs and adjectives in WordNet 3.0. Through the BCs, we will ensure that any synset in the wordnets is mapped to some concept in the ontology either directly or indirectly[9]. The most specific layer of the ontology contains concepts representing species and regions relevant to the KYOTO domain. These concepts were provided by the end users, and in certain cases, concepts have been added to link the domain specific terms to the ontology.

In the example shown in Figure 2, we see typical role concepts as terms. For these role concepts, we infer that they do not represent rigid subtypes but can be used to refer to instances of concepts that play a specific role. The role relation to the process needs to be defined more specifically through a mapping relation with the ontology. To properly define the semantics of this model, we need to define the precise relations between the concepts represented in the different repositories. This will be discussed in the next section.

4 The Species2000 thesaurus

After a review of available internet-based resources, the Species2000 project website[10] was selected as the source for the lists of animals, plant, fungi and microbes. Species2000 is a project which aims to create a comprehensive validated checklist of all the species in the world. The decision to choose this resource was based on factors/criteria such as: the consistency of the taxonomic system it utilizes; the ongoing expert validation of the Species2000 database; its currency in terms of being regularly updated; and the (comparatively) comprehensive nature of its coverage. To achieve these standards Species2000 brings together information from 52 databases from all around the world, which could be expanded and which together cover all of the major groups of organisms. These species are listed using a consistent taxonomic system which can be consulted through a web-interface at the above mentioned website address.

According to the Species 2000 website the databases currently used by the system account for approximately 60% of all known species. Because Species 2000 can be consulted through a web interface and is available as MySQL database. The MySQL database has been converted into Resource Description Framework (RDF) format. This domain specific thesaurus, provides an important vocabulary that can be used to model part of the knowledge in the environment domain. It contains around two million species structured according to a biological taxonomy. Each concept has at least a Latin name and often many alternative

[5] `http://www.wikipedia.org`

[6] `http://www.dbpedia.org`

[7] `http://www.geonames.org/`

[8] `http://www.sp2000.org`

[9] This set of BCs is more minimal than the BCs defined in EuroWordNet and BalkaNet. The original BC set contained too much redundancy and arbitrariness for our purposes.

[10] `http://www.sp2000.org/`

labels in different languages. An example of a Latin hierarchy is shown on Figure 3. Implicitly, each level of the hierarchy corresponds to a particular level of the biological classification.

To be able to exploit the data, we converted the Species2000 format to SKOS format and published it in Virtuoso. The taxonomic relations have been converted to skos:broader relations. To extend the language labels, we looked for the Latin name in DBPedia and collected all language labels for a matching record. The results are shown in Table 1.

Table 1: Language labels for Species 2000 concepts after alignment with DBPedia

Language	Species 2000	DBPedia extension
English	69,045	834,821
Spanish	1,731	358,499
Italian	17,552	215,511
Dutch	5,397	185,437
Chinese	58,774	83,756
Japanese	4,625	139,754
Total	157,124	1,817,778

The number of language labels increased from 157,124 to 1,817,778 labels. Note that a single concept can have many different synonymous labels. However, there are still many language gaps. That is, there are many Species 2000 concepts that only have a Latin name. Figure 4 shows an example of the SKOS entry corresponding to the subspecies ITS-207724, whose scientific Latin name is “Eleutherodactylus augusti”. This subspecies is also know as “Barking Frog” in English and “Rana-ladradora común” in Spanish. The rest of alternative labels for English, French, Dutch, Spanish and Italian (as well as for many other languages) have been acquired using the multilingual correspondences of DBpedia.

If sufficient nodes in the vocabulary are represented by labels in a language, the hierarchy can be used to create a mapping across the database and the wordnet in a language. For mapping the SKOS Species 2000 database to WordNet3.0, we thus can use the original Latin names occurring in the Species 2000 hierarchies and the corresponding 834,821 English labels. In fact, many species are named by its Latin name in WordNet3.0 as well.

5 Integrating Species2000 and WordNet3.0

In order to perform the integration, we designed a novel and more flexible approach to align Species2000 concepts to the WordNet3.0 synsets. First, we manually aligned to the WordNet3.0 synsets the Kingdoms appearing in the Species2000. Then, we perform the alignment automatically following a depth-breath order on each of the taxonomical branches occurring in the Species2000 ontology. Thus, we will align the Species2000 branches by using the original SKOS file which includes by order partial branches. For example, Figure 5 shows a partial view of Species2000 sequences of ordered taxonomic branches.

We also keep record of the alignment of a particular Species2000 concept occurring in a branch allowing to maintain an appropriate consistency of the aligment.

The alignment process have been carried out by using a robust and accurate knowledge-based Word Sense Disambiguation algorithm. We used a version of the Structural Semantic Interconnections algorithm (SSI) called SSI-Dijkstra (Cuadros and Rigau, 2008), (Laparra and Rigau, 2009). SSI is a knowledge-based iterative approach to Word Sense Disambiguation (Navigli and Velardi, 2005). Previously, the SSI-Dijkstra algorithm have been used for constructing KnowNets (Cuadros and Rigau, 2008) and for the integration of WordNet and FrameNet (Laparra and Rigau, 2009).

The original SSI algorithm is very simple and consists of an initialization step and a set of iterative steps. Given W, an ordered list of words to be disambiguated, the SSI algorithm performs as follows. During the initialization step, all monosemous words are included into the set I of already interpreted words, and the polysemous words are included in P (all of them pending to be disambiguated). At each step, the set I is used to disambiguate one word of P, selecting the word sense which is closer to the set I of already disambiguated words. Once a sense is disambiguated, the word sense is removed from P and included into I. The algorithm finishes when no more pending words remain in P.

SSI-Dijkstra uses the Dijkstra algorithm to obtain the shortest path distance between a node and some nodes of the whole graph. The Dijkstra algorithm is a greedy algorithm that computes the shortest path distance between one node an the rest of nodes of a graph. BoostGraph[11] library can be used to compute very efficiently the shortest distance between any two given nodes on very large graphs. We also use already available knowledge resources to build very large connected graphs. In fact, we perform the aligment by using two graphs. The first graph used only hyponym/hypernym relations with 97,666 edges and the second used the set of direct relations between synsets gathered from WordNet3.0 and the relations extracted from the sense annotated WordNet glosses, totalizing 595,339 edges. That is, the first one with only WordNet hyponymy/hypernymy relations and a second one with all WordNet and gloss relations.

Note that initially, the list I of interpreted words should include the senses of the monosemous words in W, or a fixed set of word senses. Remember that we already have the top Kingdom term of each taxonomic branch from Species2000 manually aligned to its appropriate WordNet synset.

Consider, the example in Figure 6. In this case, only “animalia” (aligned manually to animal#n#1) and “amphibia” appear in WordNet3.0. However, in English “eleutherodactylus” is also “barking_frog” which appears in WordNet3.0. Thus, the program stablishes the aligment shown in Figure 6.

The mapping also provides the proximity scores of the two graphs used and the synset WordNet Lexicographer file, in this case ANIMAL[12]. We use the two scores provided by the SSI-Dijkstra algorithm and the Lexicographer files to filter out inappropriate matchings. We only selected those

[11] `http://www.boost.org/doc/libs/1_35_0/libs/graph/doc/index.html`

[12] `http://wordnet.princeton.edu/man/lexnames.5WN.html`

```
Kingdom: Animalia ->
         Class: Chordata ->
                Order: Amphibia ->
                       Family: Anura ->
                               Genus: Leptodactylidae ->
                                      Species: Eleutherodactylus ->
                                               Infra species: Eleutherodactylus augusti
```

Figure 3: Example of the biological classification of an Species2000 concept

```
<skos:Concept
rdf:about="http://kyoto-project.eu/col2009ac/Animalia/Chordata/Amphibia/Anura/Leptodactylidae/Eleutherodac-tylus/ITS-207724">
        <skos:prefLabel xml:lang="la">Eleutherodactylus augusti</skos:prefLabel>
        <skos:prefLabel xml:lang="en">Barking Frog</skos:prefLabel>
        <skos:prefLabel xml:lang="es">Rana-ladradora común</skos:prefLabel>
        <skos:altLabel xml:lang="en">Eleutherodactylus</skos:altLabel>
        <skos:altLabel xml:lang="fr">Eleutherodactylus</skos:altLabel>
        <skos:altLabel xml:lang="nl">Eleutherodactylus</skos:altLabel>
        <skos:altLabel xml:lang="es">Eleutherodactylus</skos:altLabel>
        <skos:altLabel xml:lang="pt">Eleutherodactylus coqui</skos:altLabel>
        <skos:broader
rdf:resource="http://kyoto-project.eu/col2009ac/Animalia/Chordata/Amphibia/Anura/Leptodactylidae/Eleuthero-dactylus"/>
</skos:Concept>
```

Figure 4: Example of SKOS concept enriched with language labels from Dbpedia

```
Animalia : Chordata
Animalia : Chordata : Amphibia
Animalia : Chordata : Amphibia : Anura
Animalia : Chordata : Amphibia : Anura :  Leptodactylidae
Animalia : Chordata : Amphibia : Anura :  Leptodactylidae : Eleutherodactylus
...
```

Figure 5: Example of Species2000 sequences of ordered taxonomic branches

```
Animalia : Chordata : Amphibia : Anura :  Leptodactylidae : Eleutherodactylus

animal n 00015388-n "a living organism characterized by voluntary movement"
amphibia n 01625747-n "the class of vertebrates that live on land but breed in water;
                      frogs; toads; newts; salamanders; caecilians"
barking_frog n 01643507-n "of southwest United States and Mexico; call is like a dog’s bark"
```

Figure 6: Example of correct aligment

alignments appearing in the ANIMAL, PLANT lexicographer files and with the scores above average. Finally, a total number of 150,486 Species2000 concepts have been aligned to a WordNet3.0 synset, while filtering out 330,167 potential connections. The total number of concepts in Species2000 is 3,006,105. Thus, we are connecting to WordNet3.0 just a small amount of concepts. In fact, the mapping process just identifies in WordNet already occurring concepts from Species2000. The rest can be considered as new domain concepts not present in WordNet3.0. However, all Species2000 concepts will be now connected to a particular WordNet concept, either directly or indirectly because they are related through skos:broader relations to another concept that is mapped directly. Equivalent relations to WordNet3.0 concepts will be established for the 150,486 identified Species2000 concepts. The rest is aligned to more general WordNet3.0 concepts (the previous aligned concept in the Species2000 hierarchy) through the broader relation. Likewise, we have been able to combine the Species2000 database with the generic Wordnet with just a minimal manual effort to connect the top nodes of the 5 kingdoms in Species2000. For every concept in Species2000 we can thus access the richer wordnet relations and any ontology that is mapped to wordnet. In the case of KYOTO, this means that text mining patterns that are formulated with ontological labels at a generic level, such as organisms-live-in-habitats, can be applied to texts in different languages that contain specific names for species that are only found in Species2000.

6 Evaluation

In order to perform an initial evaluation of the aligment process, we selected randomly a small set of one-hundred filtered alignments. An independent evaluator (not an expert in the field) established the correctness of the mapping according to the following categories:

- C= correct
- B = matches the broader term

- BB = matches even higher up in the hierarchy

- X = incorrect

We ignored the infraspecies level. So if it was an infraspecies, the species level was also correct. The results show no incorrect cases (X). It seems that the filtering process performed correctly. For instance, the branch shown in Figure 7 was not included as a result of the mapping. However, almost all are B (48) or BB (52), and only one case is C.

Possibly, adjusting the filtering parameters we would obtain different coverage/accuracy figures.

7 Error analysis

We can partly explain this behavior looking at the example shown in Figure 8 trying to stablish the connection at the “genus” level of drosophila.

But, “genus Drosophila” also occurs in WordNet3.0 as synset eng-30-02197545-n. Thus, we are matching too high in the hierarchy. We are probably missing potential candidates since we are not taking into account the information of the level description of the Species2000 hierarchy. Thus, the general lookup strategy could be extended with domain specific heuristics to improve matching (e.g. use the genus, order, family clues). Such lookup modules need to be made for each domain and used optional in the software.

Furthermore, if the concept is not found in WordNet3.0, we use the previous aligned concept in Species2000 hierarchy. This is always a more abstract concept. In that case we should also change the SKOS mapping to skos:broaderMatch. That will make our results better.

8 Conclusions and future work

We have presented a robust approach to align a large domain ontology of Species to WordNet. The method relies on a knowledge-based Word Sense Disambiguation algorithm. The approach can be easily improved by taking account of ontology specific heuristics. For instance, by using clues from the hierarchy level since we always know if the term belongs to a genus, order, family, etc. We also plan to carry out a more complete evaluation on the filtering process. Through the mapping, we extended the wordnets for many languages with millions of domain concepts. The alignment of such domain ontologies can be performed on a regular basis to maintain an up-to-date integration of the work of the domain experts and the generic wordnets. Through the generic wordnets, the domain ontologies are mapped to a shared generic ontology.

Acknowledgments

This work has been supported by KNOW-2 (TIN2009-14715-C04-01 and TIN2009-14715-C04-04) and KYOTO (ICT-2007-211423). We want to thank the anonymous reviewers for their valuable comments.

9 References

E. Agirre and A. Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009)*, Athens, Greece, April. Eurpean Association for Computational Linguistics.

C. Bizer, T. Heath, K. Idehen, and T. Berners-Lee. 2008. Linked data on the web. In *WWW*, pages 1265–1266.

W. Bosma, P. Vossen, G. Rigau, A. Soroa, M. Tesconi, A. Marchetti, M. Monachini, and C. Aliprandi. 2009. Kaf: a generic semantic annotation format. In *Proceedings of the Generative Lexicon 2009*, pages 145–152.

M. Cuadros and G. Rigau. 2008. Knownet: Building a large net of knowledge from the web. In *Proceedings of COLING*.

J. Daudé, L. Padró, and G. Rigau. 2000. Mapping WordNets Using Structural Information. In *Proceedings of 38th annual meeting of the Association for Computational Linguistics (ACL'2000)*, Hong Kong.

A. Doan, J. Madhavan, P. Domingos, and A. Halevy. 2002. Learning to map between ontologies on the semantic web. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 662–673, New York, NY, USA. ACM.

C. Fellbaum and P. Vossen. 2007. Connecting the universal to the specific: Towards the global grid. In *Intercultural Collaboration I : Lecture Notes in Computer Science, Springer-Verlag*.

C. Fellbaum and P. Vossen. 2008. Challenges for a global wordnet. In *Proceedings of the First International Workshop on Global Interoperability for Language Resources(ICGL)*.

A. Gangemi, N. Guarino, C. Masolo, and A. Oltramari. 2003. Sweetening wordnet with dolce. *AI Mag.*, 24(3):13–24.

N. Guarino and C. Welty. 2002. Evaluating ontological decisions with ontoclean. *Commun. ACM*, 45(2):61–65.

N. Guarino and C. Welty. 2004. An Overview of OntoClean. In Steffen Staab and Rudi Studer, editors, *Handbook on Ontologies*, International Handbooks on Information Systems, pages 151–172. Springer, Berlin.

A. Hicks and A. Herold. 2009. Evaluating ontologies with rudify. In Jan L. G. Dietz, editor, *Proceedings of the 2nd International Conference on Knowledge Engineering and Ontology Development (KEOD'10)*, pages 5–12. INSTICC Press.

A. Horák, K. Pala, A. Rambousek, and M. Povolný. 2006. Debvisdic - first version of new client-server wordnet browsing and editing tool. In *In Proceedings of the Third International WordNet Conference - GWC 2006*, pages 325–328.

R. Izquierdo, A. Suárez, and G. Rigau. 2007. Exploring the automatic selection of basic level concepts. In *Proceedings of the International Conference, Recent Advances on Natural Language Processing RANLP'07*, Borovets, Bulgaria.

E. Laparra and G. Rigau. 2009. Integrating wordnet and framenet using a knowledge-based word sense disambiguation algorithm. In *Proceedings of the International Conference, Recent Advances on Natural Language Processing RANLP'09*, Borovets, Bulgaria.

```
Animalia : Mollusca : Gastropoda : Baso mmatophora : Planorbidae : Armiger
score WN hierarchy=0.272727272727273
score WN+gloss=0.0769230769230769
synset=eng-30-09808591-n
lexicographer file=PERSON
```

Figure 7: Example of filtered aligment

```
Animalia : Arthropoda : Insecta : Diptera : Drosophilidae : Drosophila
score WN hierarchy=0.5
score WN+gloss=0.19047619047619
synset=eng-30-02197413-n
lexicographer file=ANIMAL
```

Figure 8: Example of a too high aligment

A. Maedche and S. Staab. 2001. Ontology learning for the semantic web. *IEEE Intelligent Systems*, 16(2):72–79.

D. McGuinness, R. Fikes, J. Rice, and S. Wilder. 2000. The chimaera ontology environment. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 1123–1124. AAAI Press / The MIT Press.

R. Navigli and P. Velardi. 2005. Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(7):1063–1074.

N. Noy and M. Musen. 2001. Anchor-prompt: Using non-local context for semantic matching. In *In Proc. IJCAI 2001 workshop on ontology and information sharing*, pages 63–70.

A. Pease, I. Niles, and J. Li. 2002. The suggested upper merged ontology: A large ontology for the semantic web and its applications. In *In Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web*.

A. Pease, C. Fellbaum, and P. Vossen. 2008. Building the global wordnet grid. In *Proceedings of the 18th International Congress of Linguists (CIL18)*.

H. Putnam. 1975. The Meaning of ’Meaning’. *Minnesota Studies in the Philosophy of Science*, 7:131–193.

M. A. Rodriguez and M. J. Egenhofer. 2003. Determining semantic similarity among entity classes from different ontologies. *Knowledge and Data Engineering, IEEE Transactions on*, 15(2):442–456.

F. Ronzano, M. Tesconi, S. Minutoli, and A. Marchetti. 2010. Collaborative management of kyoto multilingual knowledge base: The wikyoto knowledge editor. In *Proceedings of the 5th Global WordNet Conference (GWC’10)*.

C. Soria, M. Monachini, and P. Vossen. 2009. Wordnet-lmf: fleshing out a standardized format for wordnet interoperability. In *IWIC ’09: Proceeding of the 2009 international workshop on Intercultural collaboration*, pages 139–146, New York, NY, USA. ACM.

A. Toral, M. Monachini, C. Soria, M. Cuadros, G. Rigau, W. Bosma, and P. Vossen. 2010. Linking a domain thesaurus to wordnet and conversion to wordnet-lmf. In *Proceedings of Second International Conference on Global Interoperability for Language Resources (ICGL’10)*.

W. R. van Hage. 2008. *Evaluating Ontology-Alignment Techniques*. Ph.D. thesis, Vrije Universiteit Amsterdam.

P. Vossen and G. Rigau. 2010. Division of semantic labour in the global wordnet grid. In *Proceedings of the 5th Global WordNet Conference (GWC’10)*.

P. Vossen, E. Agirre, N. Calzolari, C. Fellbaum, S. Hsieh, C. Huang, H. Isahara, K. Kanzaki, A. Marchetti, M. Monachini, F. Neri, R. Raffaelli, G. Rigau, M. Tesconi, and J. VanGent. 2008. Kyoto: a system for mining, structuring and distributing knowledge across languages and cultures. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC’08)*, Marrakech, Morocco, may.

P. Vossen, editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks* . Kluwer Academic Publishers .