Mapping WordNet to the Kyoto ontology

Egoitz Laparra*, German Rigau*, Piek Vossen◇

* IXA group UPV/EHU, Donostia Spain, Donostia, Spain {egoitz.laparra, german.rigau}@ehu.com
◇ Dept. Letteren. Vrije Universiteit. Amsterdam, Netherlands p.vossen@let.vu.nl

Abstract

This paper describes the connection of WordNet to a generic ontology based on DOLCE. We developed a complete set of heuristics for mapping all WordNet nouns, verbs and adjectives to the ontology. Moreover, the mapping also allows to represent predicates in a uniform and interoperable way, regardless of the way they are expressed in the text and in which language. Together with the ontology, the WordNet mappings provide a extremely rich and powerful basis for semantic processing of text in any domain. In particular, the mapping has been used in a knowledge-rich event-mining system developed for the Asian-European project KYOTO.

Keywords: Lexicon, Lexical Database, Ontologies, Semantics, WordNet

1. Open Domain Semantic Processing

Traditionally, Information Extraction (IE) is the task of filling template information from previously unseen text which belongs to a predefined domain (Peshkin and Pfeffer, 2003). Standard IE systems are based on language-specific pattern matching, consisting of language-specific regular expressions and associated mappings from syntactic to logical forms. The major disadvantage of traditional IE systems is that they focus on satisfying precise, narrow, pre-specified requests e.g. facts about terrorist events.

Alternatively, the KYOTO system[1] can be applied to different languages and domains in a uniform way thanks to a very rich knowledge model which maps text items to ontological statements. To achieve this, it uses a very rich knowledge model in the form of wordnets and a generic ontology (Vossen and Rigau, 2010). After basic morpho-syntactic processing, the system integrates the semantic information of named-entities and open-class words by applying a graph-based Word Sense Disambiguation to words in context, scoring all possible synsets from a local wordnet. Each of these synsets is mapped to a shared ontology. From this mapping, all possible ontological implications are derived and integrated into a very rich text representation. The output of the linguistic processors 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)). Currently, the semantic information is included for all possible senses according to a given wordnet together with all its associated ontological statements[2].

[1] Available at `www.kyoto-project.eu/`

[2] A demo for English, Spanish and Basque is provided at `http://ixa2.si.ehu.es/demokaf/demokaf.pl`

In KYOTO, the knowledge extraction is done by so-called Kybots (Knowledge Yielding Robots). Kybots are defined by a set of profiles representing information patterns. In a profile, conceptual relations are expressed using ontological and morpho-syntactic patterns. Since the semantics is defined through the ontology, it is possible to detect similar data even if expressed differently in the same or (with minor modifications) in a different language.

In order to combine the Kybots with a rich and flexible knowledge model, we defined a three layered architecture. Each layer contains a different source of knowledge: a local lexicon, a local wordnet and a common ontology. To guarantee maximum coverage, it is necessary to connect every synset of wordnet to its corresponding ontological features, i.e. to achieve semantic closure. Connecting all synsets manually is very expensive, too time consuming and not free of errors and inconsistencies. We therefore describe a method to obtain a full mapping between the English WordNet and the ontology. In the next sections, we first describe the knowledge model and next the heuristics for mapping all WordNet nouns, verbs and adjectives to the ontology.

2. Knowledge Model

As a semantic model, the system exploits a three-layered knowledge architecture which integrates using formal semantic relations three different types of resources: a central ontology, local wordnets and large background vocabularies linked to the wordnets (Vossen and Rigau, 2010). This model follows the principle of the division of labour (Putnam, 1975). Following this principle, we can state that a computer does not need to distinguish between instances of rigid concepts (as defined by (Guarino and Welty, 2002)),

such as a *wigeon* and a *mallard*. We assume that domain experts have the necessary knowledge to keep them apart and that these properties are not necessary for text mining applications. Rigid concepts thus do not need to be defined formally in the ontology but can be kept in the available background resources, such as databases with millions of species (Cuadros et al., 2010). We assume that hyponymy relations to rigid synsets in WordNet declare those subconcepts as rigid subtypes too unless we provide information to the contrary. Instead, the ontology should help describing non-rigid concepts, e.g. *endangered birds*, *migratory birds* which refer to processes and states for which we find valuable information in textual sources.

In this model, the ontology does not need to be the central hub for all terms in a domain in all languages and the three knowledge layers can be developed separately by combining the efforts from three different communities:

- Domain experts in social communities that continuously build background vocabularies;
- Wordnet specialists that define the general concepts for a language
- Semantic Web specialists that define top-level and domain-specific ontologies that capture formal definitions of concepts

The semantic model provides complete mappings to the ontology for all nominal, verbal and adjectival WordNet3.0 synsets (Fellbaum, 1998)[3]. The mappings also harmonize information across different part-of-speech (POS). For instance, migratory events represented by different synsets for verbs *migrate*, nouns *migration* and adjectives *migratory* inherit the same ontological information corresponding to the *ChangeOfResidence* class. Consequently, event mentions in text exhibiting a large variety of syntactic structures will be modeled semantically in the same way if we are able to establish the appropriate connection to the same ontological types. This knowledge model provides an extremely powerful basis for semantic processing in any domain. Furthermore, through the equivalence relations of wordnets in other languages to the English WordNet, this semantic framework can also be applied to the other languages.

2.1. KYOTO ontology

As a central ontology, we used the KYOTO ontology (Hicks and Herold, 2009)[4]. This ontology currently consists of 1,964 classes divided over three layers. The top layer is based on DOLCE [5] (Gangemi et al., 2003) and OntoWordNet. The second layer are the Base Concepts[6] (BCs) which cover all nominal and verbal WordNet synsets (Izquierdo et al., 2007) at an intermediate level of abstraction. BCs are synsets in WordNet 3.0 that have many relations with other synsets and are selected in a way that ensures complete coverage of the nominal and verbal part of WordNet, i.e. resulting in semantic closure of the full vocabulary by the ontology. Examples of BCs are: *building, vehicle, animal, plant, change, move, size, weight*. In total, there are only 309 nominal and 578 verbal BCs, which have been incorporated as new types into the Kyoto ontology as a second layer. The nominal BCs have been manually assigned to various existing ontology types. For the verbal BCs, we first included new types for the 15 verbal WordNet lexicographer's files as subclasses of the type *perdurant* (events, processes, phenomena, activities and states). Next, the verbal BCs were integrated as subclasses under the type corresponding with their WordNet lexicographer's file. A third layer consists of domain classes introduced for detecting events and qualities in a particular domain (i.e. environment). Special attention has been paid to represent the processes (perdurants) in which objects (endurants) of the domain are involved and the qualities they may have.

2.2. Wordnet to ontology relations

In addition to the regular synset to synset relations in WordNet, we defined a specific set of relations for mapping the synsets to the ontology, which are all prefixed with sc_ standing for synset-to-concept. We differentiate between rigid and non-rigid concepts in WordNet through the mapping relations:

- sc_equivalenceOf [synset, type]: the synset is fully equivalent to the ontology Type and inherits all its properties; the synset is Rigid
- sc_subclassOf [synset, type]: the synset is a proper subclass of the ontology Type and inherits all its properties; the synset is Rigid
- sc_domainOf [synset, type]: the synset is not a proper subclass of the ontology Type and is not disjoint (therefore orthogonal) with other synsets that are mapped to the same Type either through sc_subclassOf or sc_domainOf; the synset is non-Rigid but still inherits all properties of the target

[3]This knowledge model is freely available through the KYOTO website as open-source data.

[4]Available at http://www.kyoto-project.eu/

[5]DOLCE-Lite-Plus version 3.9.7

[6]http://adimen.si.ehu.es/web/BLC

ontology Type; the synset is also related to a Role with a sc_playRole relation

- sc_playRole [synset, role]: the synset denotes instances for which the context of the Role applies for some period of time but this is not essential for the existence of the instances, i.e. if the context ceases to exist then the instances may still exist (Mizoguchi et al., 2007) [7]
- sc_participantOf [synset (endurant), perdurant]: instances of the concept (denoted by the synset) participate in some perdurant, where the specific role relation is indicated by the playRole mapping.
- sc_hasState [synset, quality], : instances of the concept are in a particular state denoted by the ontological quality value. This state is is not essential and can change.

This model extends existing WordNet to ontology mappings. For instance, in the SUMO to Wordnet mapping (Niles and Pease, 2003), only the sc_equivalenceOf and sc_subclassOf relations are used, represented by the symbols = and + respectively. The SUMO-Wordnet mapping likewise does not systematically distinguish rigid from non-rigid synsets. In our model, we separate the linguistically and culturally specific vocabularies from the shared ontology while using the ontology to interface the concepts used by the various communities.
Through the mapping relations, we can keep the ontology relatively small and compact whereas we can still define the richness of the vocabularies of languages in a precise way. The classes in the ontology can be defined using rich axioms that model precise implications for inferencing.

3. Mapping methodology

3.1. Connecting Nouns and Verbs to the Ontology

We followed a semi-automatic approach to create mappings from all WordNet synsets to the KYOTO ontology. Firstly, we derived a complete mapping for nouns and verbs by exploiting the BCs. All nominal and verbal synsets in WordNet are indirectly related to at least one BC through the hypernym relations and likewise to the ontology. All nominal and verbal WordNet concepts have been aligned to its corresponding ontological types. All concepts corresponding to a Base Concept have been aligned by using the "sc_equivalentOf" relation. The rest of nominal and verbal concepts have been mapped as "sc_subClassOf" to the class in the ontology corresponding to the its Base Concept.
We also created mappings for those nominal concepts connected by morpho-semantic links to events. We used the WordNet morpho-semantic database[8] to create a set of rule-based heuristics to derive additional connections among nominal concepts from WordNet and event types in the ontology. WordNet 3.0 contains derivational links connecting noun and verb senses, e.g. between *cannibal*^1_n and *cannibalise*^2_v. WordNet morpho-semantic database also includes the semantic type of the relationship (e.g. agent). The database uses 14 semantic relations between verbs and nouns. Table 1 shows an example and the relative frequency for the three most frequent relations.
For each type of semantic relation we created an heuristic to establish a connection for each synset to its corresponding ontological type. Additionaly, in order to filter out incorrect assigments, the heuristics perform a validation test regarding semantic properties for both nominal and verbal synsets of the relation. To distinguish among potential candidate connections, these rules use as a background knowledge the EuroWordNet Top Ontology[9] (Álvez et al., 2008). We use the EuroWordNet Top Ontology to test for their *static* of *dynamic* attributes. We also use the hierarchical WordNet structure to test for endurant or perdurant properties[10].
Following (Nervo and Laure, 2011) we used the following equivalences:

- endurant = physical_entity^1_n \ process^1_n
- perdurant = process^1_n ∪ event^1_n ∪ state^1_n

We also add the following equivalences:

- static = state^1_n
- dynamic = process^1_n ∪ event^1_n
- instrument = instrumentation^1_n
- object = object^1_n
- substance = substance^1_n
- person = person^1_n

[7] Some terms involve more than one role, e.g. gas-powered-vehicle. Secondary participants are related through sc_hasCoParticipant and sc_playCoRole mappings.

[8] http://wordnetcode.princeton.edu/standoff-files/morphosemantic-links.xls

[9] http://adimen.si.ehu.es/web/WordNet2TO

[10] In some cases we could also use the KYOTO ontology itself.

event (8,158 links)	
intensify^2_v	make more intense, stronger, or more ...
intensification^1_n	action that makes something stronger or ...
agent (3,043 links)	
cannibalise^2_v	eat human flesh
cannibal^1_n	a person who eats human flesh
result (1,439 links)	
acquit^2_v	pronounce not guilty of criminal charges ...
acquittal^1_n	a judgment of not guilty
by-means-of (1,273 links)	
approbate^2_v	approve or sanction officially
approval^1_n	a message expressing a favorable opinion ...
undergoer (878 links)	
remit^2_v	send (money) in payment
remittal^1_n	a payment of money sent to a person in a ...
instrument (813 links)	
accelerate^2_v	cause to move faster
accelerator^1_n	a pedal that controls the throttle valve...
uses (740 links)	
signalize^2_v	provide with traffic signals
signal^1_n	any nonverbal action or gesture that ...
state (528 links)	
survive^2_v	continue in existence after (an ...
survival^1_a	a state of surviving; remaining alive
property (318 links)	
beautify^2_v	be beautiful to look at
beauty^1_a	the qualities that give pleasure to the ...
location (288 links)	
hospitalize^2_v	admit into a hospital
hospital^1_a	a medical institution where sick or ...

Table 1: Examples for the most frequent morpho-semantic links

3.1.1. Heuristic for event relations

As an example, consider the heuristic for event relations[11]. Using the heuristic for the event relations, all 8,158 event relations have been connected to the ontology since no test was required. Consider the example presented in table 2.

00227165-v intensify –event–> 00374224-n intensification
00227165-v intensify –BCverb–> 00156601-v increase
00374224-n intensification –BCnoun–> 00363260-n change

Table 2: Example of event relation

Thus, we include the connections to their corresponding Base Concepts as shown in table 3, and being nominalisations of the same events, we also establish additional relations to its related Base Concept types.

[11]The KYOTO deliverable D8.3 ”Domain extension of central ontology” presents the whole set of heuristics applied for all type of relations

eng-30-00227165-v sc_subClassOf Kyoto#increase-eng-3.0-00156601-v
eng-30-00374224-n sc_subClassOf Kyoto#change-eng-3.0-00191142-n
eng-30-00227165-v sc_subClassOf Kyoto#change-eng-3.0-00191142-n
eng-30-00374224-n sc_subClassOf Kyoto#increase-eng-3.0-00156601-v

Table 3: Direct and inferred relations

3.1.2. Connecting Adjectives

In order to establish appropriate connections to the Kyoto ontology for adjectives, we also use the Princeton semantic relations to connect adjectives to its more appropriate nominal or verbal concepts. Through their hypernymy chain we also map each adjective as ”sc_qualityOf” to the type corresponding to the Base Concept of the noun or verb. We also map each adjective as ”sc_subclassOf” to the ”Kyoto#quality-eng-3.0-04723816-n” type. The total number of adjectival synsets is 18,156. There are 10,693 synsets which

are satellites of 7,464 kernel synsets. However, not all kernels have antonym relations to other kernels. There are 3,618 kernels without antonymy relation. They could have satellite synsets but do not constitute a complete cluster. We thus distinguish between two types of clusters: complete clusters and semi-clusters. If a kernel has an antonymy relationship with another kernel then it forms a complete cluster. If a kernel has no antonymy relation it forms a semi-cluster. Examples of complete clusters with one or more antonyms are shown in Table 4.

00031974-a active <–antonym–> 00033574-a inactive
01105620-a nonspecific <–antonym–> 01103021-a specific
01105620-a nonspecific <–antonym–> 01105233-a specific

Table 4: Some complete adjectival clusters

An example of a semi-cluster without antonyms is shown in Table 5.

01380267-a aerial

Table 5: An adjectival semi-cluster

There are 1,897 complete clusters with two or more semi-clusters. There are 1,849 complete clusters with two semi-clusters, 44 clusters with three semi-clusters and 4 clusters with four semi-clusters. We process each cluster and semi-cluster to establish appropriate connections to nouns and verbs depending to the available semantic relations in WordNet 3.0 relating the adjectives to nouns or verbs. We processed the adjectives depending on the structure and the adjective relations contained in the clusters to nouns and verbs. Thus, we have nine different sets. We developed several heuristics considering the following WN relations between adjectives and nouns or verbs (no priority is given to none of them):

- Related form
- Participle of verb
- Pertainym (pertains to noun)
- Attribute

As an exemple we will ilustrate the mapping process for a particular example corresponding to one of the 334 complete clusters[12] . Recall that this type of cluster have the antonym kernels with its corresponding antonym noun or verb. For instance, consider the complete mapping shown in Table 6.

[12]The KYOTO deliverable D8.3 "Domain extension of central ontology" presents the whole set of heuristics applied for all types of adjectival clusters

00031974-a active <–antonym–> 00033574-a inactive
04635104-n activeness <–antonym–> 04635631-n inactiveness
00031974-a active <–related–> 04635104-n activeness
00033574-a inactive <–related–> 04635631-n inactiveness
04635104-n activeness –BCnoun–> 04616059-n trait
04635631-n inactiveness –BCnoun–> 04616059-n trait

Table 6: An example mapping for a complete adjectival cluster

We use these relations to derive the Base Concepts of the adjectives (kernel synsets and its satellites) by using the Base Concepts of the nouns and verbs. Table 7 shows an example of the new Base Concepts derived for adjectives.

00031974-a active –BCadj–> 04616059-n trait
00033574-a inactive –BCadj–> 04616059-n trait

Table 7: An example mapping for a complete adjectival cluster

4. Evaluation

We have not carried out a direct manual evaluation of samples of the knowledge structure. It is not trivial to carry out such an evaluation since the semantics applies to hundreds of thousands of words and concepts. We did however use the model within an event-mining task that is described in more detail in (Vossen et al., 2012) and (Vossen et al., fc). Within this task, the sequential representation of the words in a text is first expanded to WordNet synsets and next to all the ontological implications that apply to these synsets on the basis of our model. The implications are collected by determining for each synset the set of synsets to which it is related through hyponymy that have a mapping to the Kyoto ontology. Next these ontological types are inserted in the text representation and then expanded to other implications that follow from the ontological structure. For example, the term *alis shad* is a hyponym or *migratory fish* in WordNet. The latter has a mapping to the perdurant ontology type *Migration* as a process in which it is involved. The ontology states that *Migration* is a subclass of *Change-of-Location*. The event *Change-of-Location* has axioms that imply a has-path, has-source and has-destination role. Likewise, we insert mappings to all these classes and the roles into the textual representation of *alis shad* in the text. We showed that these representation can be used by the Kybots to successfully extract events, participants and their roles regardless of the many ways in which they are expressed. Furthermore, the evaluation showed that we can use word-sense-disambiguation scores to prefer certain seman-

tic interpretations above others but also that suboptimal word-sense-disambiguation gives the highest F-measure in terms of precision and recall. The richness of the model expands the textual representation 20 times using this methods. Nevertheless, we applied the system to thousands of documents for which processing time and efforts remained linear.

5. Conclusions

We described the connection of WordNet to a generic ontology based on DOLCE. In total, we created for all nominal, verbal and adjectival synsets, 114.016 mappings to the BCs, 185.666 mappings to the central ontology together with 30.000 mappings from ontology labels to implications in the ontology. This provides an extremely powerful basis for semantic processing of full text in any domain. Through the equivalence relations of wordnets in other languages to the English WordNet, this semantic framework can also be applied to the other languages.

Together with the ontology, the WordNet mappings provide a extremely rich and powerful basis for semantic processing of text in any domain. Through the equivalence relations of wordnets in other languages to the English WordNet, this semantic framework can also be applied to the other languages. This provides a common framework for semantic processing of text for all the languages.

6. Acknowledgments

Partial support provided by KYOTO ICT-2007-211423 and KNOW2 TIN2009-14715-C04-04.

7. References

J. Álvez, J. Atserias, J. Carreras, S. Climent, E. Laparra, A. Oliver, and G. Rigau. 2008. Complete and consistent annotation of wordnet using the top concept ontology. In *Proc. of LREC08*.

W. E. Bosma, Piek Vossen, Aitor Soroa, German Rigau, Maurizio Tesconi, Andrea Marchetti, Monica Monachini, and Carlo Aliprandi. 2009. Kaf: a generic semantic annotation format. In *Proceedings of the GL2009 Workshop on Semantic Annotation*, Pisa, Italy.

Montse Cuadros, Egoitz Laparra, German Rigau, and Piek Vossen. 2010. Integrating a large domain ontology of species into wordnet. In *Proceedings of 7th international conference on Language Resources and Evaluation (LREC'10)*, La Valleta, Malta.

C. Fellbaum. 1998. *WordNet: An Electronical Lexical Database*. The MIT Press, Cambridge, MA.

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

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

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'09)*, pages 5–12. INSTICC Press.

R. Izquierdo, A. Suarez, and G. Rigau. 2007. Exploring the automatic selection of basic level concepts. In Galia Angelova et al., editor, *International Conference Recent Advances in Natural Language Processing*, pages 298–302, Borovets, Bulgaria.

Riichiro Mizoguchi, Eiichi Sunagawa, Kouji Kozaki, and Yoshinobu Kitamura. 2007. The model of roles within an ontology development tool: Hozo. *Applied Ontolology*, 2:159–179, April.

Verdezoto Nervo and Vieu Laure. 2011. Towards semi-automatic methods for improving WordNet. In *International Workshop on Computational Semantics (IWCS), Oxford, UK*.

Ian Niles and Adam Pease. 2003. Linking lexicons and ontologies: Mapping wordnet to the suggested upper merged ontology. In *In Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE 03), Las Vegas*, pages 23–26.

L. Peshkin and A. Pfeffer. 2003. Bayesian information extraction network. In *In Proc. of the 18th International Joint Conference on Artificial Intelligence*.

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

P. Vossen and G. Rigau. 2010. Division of semantic labor in the global wordnet grid. In *Proc. of Global WordNet Conference (GWC'2010). Mumbay, India*.

P. Vossen, A. Soroa, B. Zapirain, and G. Rigau. 2012. Cross-lingual event-mining using wordnet as a shared knowledge interface. In *Proceedings of the 6th Global Wordnet Conference, Matsue, Japan*, pages 362–369.

P. Vossen, E. Agirre, G. Rigau, and A. Soroa. fc. 10. kyoto: a knowledge-rich approach to the interoperable mining of events from text. In Alessandro Oltramari, Lu Qin, Piek Vossen, and Ed Hovy, editors, *New trends of Research In Ontologies and Lexical Resources*. Springer.