2 Formal ontology as interlingua: the SUMO and WordNet linking project and global WordNet

Adam Pease and Christiane Fellbaum

2.1 WordNet

WordNet[1] is a large lexical database for English. With its broad coverage and a design that is useful for a range of natural-language processing applications, this resource has found wide general acceptance. We offer only a brief description here and refer the reader to Miller, 1990 and Fellbaum, 1998 for further details. WordNet's creation in the mid-1980s was motivated by current theories of human semantic organization (Collins and Quillian, 1969). People have knowledge about tens of thousands of concepts, and the words expressing these concepts must be stored and retrieved in an efficient and economic fashion. A semantic network such as WordNet is an attempt to model one way in which concepts and words could be organized.

The basic unit of WordNet is a set of cognitively equivalent synonyms, or synset. Examples of a noun, verb, and adjective synset are { *vacation, holiday* }, { *close, shut* }, and { *soiled, dirty* }, respectively. Each synset represents a concept, and each member of a synset encodes the same concept. Differently put, synset members are interchangeable in many contexts without changing the truth value of the context. Each synset also includes a definition, or 'gloss', and an illustrative sentence.

The current version of WordNet (3.0) contains over 117,000 synsets that are organized into a huge semantic network. The synsets are interlinked by means of bidirectional semantic relations such as hyponymy, meronymy, and a number of entailment relations. For example, the relation between *oak* and *tree* is such that *oak* is encoded as a hyponym (subordinate) of *tree* and *tree* is encoded as a hypernym (superordinate) of *oak*. *Leaf* and *trunk* are meronyms (parts) of *tree*, their holonym. Meronyms are transitive, so linking *leaf* and *trunk* to *tree* means that *oak* (and *beech* and *maple* etc.) inherits *leaf* and *trunk* as parts by virtue of its relation to *tree* (Miller, 1990, 1998). Concepts expressed by other parts of speech (verbs, adjectives) are interlinked by means of additional relations (Fellbaum, 1998).

[1] Available freely on line at http://wordnet.princeton.edu

2.1.1 Types and instances

In earlier versions of WordNet, all subordinates (hyponyms) of a given synset were encoded as **kinds-of**, or Types. For example, *mountain peak* is a Type of *topographic_point* and *town* is a Type of *municipality*. The current version further differentiates hyponyms by means of the Instance relation (Miller and Hristea, 2006). Instances are always Named Entities, and they are the leaves of trees, i.e., they never have any subordinates. For example, *Aconcagua* is not a Type but an Instance of a *mountain peak* and *Princeton* is an Instance of a *town*. While the distinction among Types and Instances is a valuable addition to WordNet, other distinctions are lacking. For example, WordNet designates nouns like *brother* and *architect* as Types of *persons*, in the same way that *dwarf, midget* is a kind of *person.* Yet the relation between *dwarf* and *person* on the one hand, and *architect* and *person* on the other hand is not the same. This can be seen by the fact that somebody can refer to the same person as both a *dwarf* and an *architect*, violating the rule that nouns with the same superordinates cannot refer to the same entity (e.g., *ash* and *oak* are both subordinates of *tree* and thus an *oak* can never be an *ash*). Rather, *architect* is a Role that a person assumes. Roles often refer to professions or functions associated with a person or temporary states (stage-level predicates) (Carlson, 1980) like *patient* and *customer*.

2.1.2 Formal vs linguistic relations

One could include in WordNet not just its current roughly one dozen semantic links, but all the hundreds of relations that are found in a formally specified logical theory like SUMO, such as **part-of**, **beforeOrEqual**, **authors** etc. Having done that, the question would be how to relate informal linguistic notions with more formal ontological relations.

Some WordNet relations, like **part-of**, would appear to be similar. However, **part-of** relations concern not only classes, but also instances. For instance, the synset { *lock* } part of the synset { *door* }, means that locks can be or often are parts of doors (a modal relation between classes or types), and the synset { *Pillar of Hercules* } part of { *Gibraltar* } means that a particular rock is part of a particular island in this case. In contrast, SUMO's **part** relation is logically specified with axioms that limit it to a relation between instances, and a relation between classes would be specified with a more complex axiom, likely involving SUMO modal relations or normative attributes.

By keeping ontological relations in the formal ontology, and linguistic relations in the lexicon, one can avoid merging two different levels of analysis and yet still capture the information that is needed about both formal concepts and linguistic tokens. A formal ontology such as SUMO also contains formal rules that specify complex relations that cannot be captured explicitly as simple links in a graph.

What is needed is an interlingual ontology that accommodates not only English WordNet but all wordnets (Elkateb *et al.*, 2006; Vossen, 1998).

2.1.3 Lexical vs conceptual ontologies

WordNet is often called an ontology, although its creators did not have in mind a philosophical construct. WordNet merely represents an attempt to map the English lexicon into a network by means of a few semantic relations. Many of these relations are implicit in standard lexicographic definitions.

The lexicon can be defined as the mappings of concepts onto words. A structured lexicon like WordNet can reveal whether the mapping is arbitrary or follows certain patterns and principles according to which concepts get labeled with a word. WordNet shows that the lexicon obeys clear patterns of semantic organization (Fellbaum, 1998) (compare Levin's 1993) syntactically based patterning of the verb lexicon).

But there are structural gaps where the geometry of the relations would require a word, yet where the language does not have one. People intuitively distinguish the class of wheeled vehicles (like cars and motorbikes) from vehicles that run on rails (trains, trams), yet English does not have a simple word for these concepts. Fellbaum (1998) argues for the existence of specific gaps on the basis of syntactic evidence that distinguishes verb classes.

A look at the lexicons of other languages quickly reveals crosslinguistic differences in lexicalization patterns; a well-known case is kinship relations (e.g. Kroeber, 1917). Languages do not seem to label concepts arbitrarily. Instead, the labeled concepts (words) follow patterns that are revealed by the fact that words can be related to one another via a few relations like hyponymy (Fellbaum, 1998). Nevertheless, the concept-word mappings of any given language are to some extent accidental; existing words do not fully reflect the inventory of concepts that is available. That inventory can be represented in a non-lexical ontology such as SUMO.

2.1.4 SUMO

The Suggested Upper Merged Ontology (SUMO)[2] (Niles and Pease, 2001) is a formal ontology stated in a first-order logical language called SUO-KIF (Genesereth, 1991; Pease, 2003). It contains some 1,000 terms and 4,000 axioms

[2] Ontologies, tools and mappings to WordNet are available freely on line at www.ontology portal.org.

(which are any logical formula) using those terms in SUO-KIF statements. The axioms include some 750 rules. SUMO is an upper ontology, covering very general notions in common-sense reality, such as time, spatial relations, physical objects, events and processes. SUMO is capped at roughly 1,000 terms in order to keep it manageable and easily learned. There is no objective test for a concept being considered 'upper level' or 'domain level', so this cutoff is purely arbitrary.

To explain the metrics above we describe a 'term' as a named concept in the ontology, which has an associated definition in logic. An axiom is any statement in logic. A rule is a particular kind of axiom that has two parts: an antecedent and a consequent. If the conditions of the antecedent are true, then the consequent must also be true. 'If an entity is a man, then he is mortal' is an example of a rule, albeit one stated in English, rather than logic.

SUMO has been extended by lower-level ontologies. A Mid-Level Ontology (MILO) has several thousand more terms with associated definitions for concepts that are more specific than the ones in SUMO, and yet are general enough not to be considered part of a topic-specific domain ontology. Domain ontologies cover over a dozen specific areas including world government, finance and economics, and biological viruses. Together with SUMO and MILO they comprise roughly 20,000 terms and 70,000 axioms. Note that in this chapter we will refer broadly to the entire collection of SUMO and its extensions as 'SUMO'.

In addition to the mappings to WordNet and linguistic paraphrases in multiple languages discussed in more detail below, the SUMO family of products includes an open-source ontology development and inference system called Sigma (Pease, 2003). SUMO is also the basis for some current work in language understanding (Pease, 2003; Pease and Fellbaum, 2004).

A formal ontology is distinguished from other ontological efforts in that it contains first-order logical rules which describe each of the terms. For example:

```
(<=>
    (earlier ?INTERVAL1 ?INTERVAL2)
    (before
        (EndFn ?INTERVAL1)
        (BeginFn ?INTERVAL2)))
```

is part of the description of the relation **earlier**. It states that 'interval 1 precedes interval 2' means that the ending time of interval 1 is before the starting time of interval 2. The axiom is bidirectional, as indicated by '<=>', which means that the axiom also says that if the end of interval 1 is before the start of interval 2 then interval 1 is earlier than interval 2. The ontology also has axioms that define the meaning of **before**, **EndFn** etc.

Where in a semantic network or a frame-based ontology one would largely have to use natural-language definitions to express the meaning of a word or concept, in a formal ontology it is solely the axioms as mathematical statements that give the terms their meaning. One could replace all the term names with arbitrary unique symbols and they would still have the same meaning. This entails that the meaning of the terms can be tested for consistency automatically with an automated theorem prover, rather than the ontologist having to rely completely on human inspection and judgment.

WordNet includes the word *earlier*, but it does not include formal axioms such as the one shown above that explains precisely to a computer what *earlier* means. Nothing in WordNet would allow a computer to assert that the end of one event is before the start of another if one event is earlier than the other.

Because the names of terms in SUMO are just convenient labels, nothing guarantees that the name of a term is going to parallel conventional usage in English. SUMO does not contain semantic relations among words of the kind found in WordNet, such as synonymy. Having two symbols that are logically equivalent is a redundancy in any mathematical theory. SUMO does contain links between formally axiomatized concepts and various labels, which include lexicalized items in different human languages, as well as locally evocative terms or phrases appropriate in a more restrict application context. For example, there is no need in SUMO to create formal terms **Above** and **HigherThan** with the same axiom:

```
(=>
    (orientation ?OBJ1 ?OBJ2 Above)
    (not
        (connected ?OBJ1 ?OBJ2)))

(=>
    (orientation ?OBJ1 ?OBJ2 HigherThan)
    (not
        (connected ?OBJ1 ?OBJ2)))
```

They would be logically redundant terms. However, SUMO can and does relate formal terms to linguistic elements appropriate in different contexts, including relating the formal term **Above** to the English WordNet synset containing the words *above*, *higher_up*, *in_a_higher_place*, *to_a_higher_place*, the Tagalog word *itaas*, the German *oberhalb* etc.

2.2 Principles of construction of formal ontologies and lexicons

Because a lexicon must accurately reflect the inventory and use of words in a given language, lexicographers do not have licence to judge whether a word has a rightful place in the lexicon. A word deemed redundant cannot be eliminated (rather, it might be treated as a synonym). And words cannot be considered to be 'missing' from the lexicon of a particular language, either, as argued by Zaefferer and Schalley, 2003. A structured lexicon like WordNet includes strings to fill 'lexical gaps' justified on structural or syntactic grounds (Fellbaum, 1998).

In contrast to a lexicon, an ontology is an engineered product. There are many ways to name, categorize, and define concepts, especially general notions in the realm of metaphysics (Loux, 2002). The absence of a word in a particular language does not prohibit creation of a term to cover a useful concept in an ontology. Similarly, the presence of a word does not entail inclusion of a term with the same name in an ontology. Every lexicalized concept should be covered by a term, but duplicate lexicalization for the same concept (synonymy) belongs in the lexicon and is not needed in a formal ontology.

In a formal ontology, the meaning of the terms only consists of the formal mathematics used to define those terms. The names of the terms could be replaced by arbitrary unique character strings and their meaning would still be the same. This independence from language gives some confidence in SUMO as a starting point for a true interlingua. While language serves as a starting point for many formalizations, it is only just that. In a lexicon, the meanings of words are determined by their use, while in a formal ontology meaning is determined only by the formal axioms. A word is coined to label a concept. The word, whether written or spoken, forms an index into the meanings. Many taxonomies, frame systems, and other informal ontologies combine the linguistic and the formal aspects, with some properties (most commonly type–instance and class–subclass relations) expressed formally, and more complex information, such as the example axiom for *earlier* given in the section above, left implicit in the name of the term or in natural-language definitions.

There have been attempts to state principles for organizing ontology (Guarino and Welty, 2000b). While there is universal agreement that such principles exist, specific proposals differ. One principle is parsimony, or simplicity, which argues for inclusion of only those terms that are needed to cover the topics or domains of interest. Two terms in an ontology should not have the same formal definitions. There is no proper notion of a synonym in a formal ontology, because the names of concepts are not important.

2.3 Mappings

We have mapped SUMO to WordNet in two phases. The first phase (Niles and Pease, 2003) consisted of mapping just SUMO itself, consisting of approximately 1,000 formally defined terms. Each synset in WordNet 1.6 was examined manually, one at a time, and a particular SUMO term was chosen as the closest equivalent. Three types of mappings were employed: rough equivalence, subsuming, and instance.

In a second phase, we looked at mapping all the word senses that occurred three or more times in SemCor (Miller *et al.*, 1993), a version of the Brown Corpus which was manually annotated with WordNet synsets. For each synset we also created a new concept in the MILO if one did not already exist in SUMO, and linked the synset to the new, more specific term.

Since a fundamental aspect of WordNet is the grouping of words in synsets, there are many cases in which a WordNet synset has several synonymous words that map to a single SUMO term. For example, the synset { *artificial_satellite, orbiter, satellite* } *man-made equipment that orbits around the earth or the moon* maps to the formally defined term of ARTIFICIALSATELLITE. The mapping is an 'equivalence' mapping since there is nothing that appears to differentiate the linguistic notion from the formal term in this case.

A more common case of mapping is a 'subsuming' mapping. For example, { *elk* } *large northern deer with enormous flattened antlers in the male; called elk in Europe and moose in North America* maps to the SUMO term HOOFEDMAMMAL.

WordNet is considerably larger than SUMO and so many synsets map to the same more general formal term. As an example of an 'instance' link, the synset { *george_washington, president_washington, washington* } *first President of the United States; commander-in-chief of the Continental Army during the American Revolution (1732–1799)* is linked to the SUMO term HUMAN. Because WordNet discriminates among different senses of the same linguistic token (polysemy), the synset { *evergreen_state, wa, washington* } *a state in northwestern United States on the Pacific* is linked via an 'instance' relation to the different term STATEORPROVINCE.

In current work, we are updating the links to point to more specific terms in the domain ontologies, when available. We have also added many links to synsets that are new to WordNet 2.1 (and later, 3.0) from 1.6. We further created three new types of links that are the negations of the original links. However, only the addition of the 'negated subsuming' link appears to be needed at this time. As an example, the synset { *concealing, concealment, hiding* } *the activity of keeping something secret* has a 'negated subsuming' link to DISSEMINATING.

One of our recent tasks was to explore automatic additions of links from SUMO to synsets that are new to WordNet 2.1 (and later, 3.0). Evidence to date is that the hypernym links in WordNet, although very different from SUMO's subclass links at the higher levels of the ontology and lexicon respectively, are far more reliably similar near the leaf levels of each structure.

Consider the synset { *morphogenesis* } *differentiation and growth of the structure of an organism...* This synset entered WordNet in version 1.7 and was not initially mapped to SUMO. It has a hypernym link to { *growth* } *the process of an individual organism growing organically...* That synset has a 'subsuming' link to the SUMO term GROWTH, which is a formally defined subclass of a PROCESS. Adding a subsuming link from GROWTH to { *morphogenesis* } was reasonable in this case and was done automatically. Further experiments will be needed to determine the reliability of this method, but initial results are promising.

A limitation of the current linking approach is that a single mapping from a lexical entity to a formal term does not fully capture the meaning of some lexical items, even if there is the option of linking to or creating a very specific formal term. The verb synset { *continue* } *exist over a prolonged period of time; 'The bad weather continued for two more weeks'* cannot be expressed as a single term, because it can refer to many unrelated types of **Process**(es). It expresses a temporal relation to an earlier point in time, referenced in the context of previous sentences. One would need a more complex relation structure to express the semantics of this lexical item. Note, however, that the existence of such problematic cases does not negate the utility and, we would contend, the necessity of expressing the relations in simpler cases.

We therefore conclude that a full semantic inventory of language is beyond what has currently been attempted here. The authors are aware that a much richer corpus is needed, such as the PhraseBank proposal as described in Pease and Fellbaum, 2004, which would capture a logical semantics for complex template linguistic expressions, rather than individual lexical items. At least some of this need may be addressed by integrating with FrameNet (Fillmore *et al.*, 2003). Because SUMO and WordNet are fully related, and significant parts of WordNet and FrameNet have been related, this may be possible in the future.

2.4 Interpreting language

Relating language and ontology is a necessity if we wish to create a deep semantic interpretation of language as in the Controlled English to Logic Translation (CELT) system (Pease and Murray, 2003). To take an example from Parsons, 1990, let us say we wish to interpret the sentence 'Brutus stabbed Caesar with a knife on Tuesday.' There are many issues with the interpretation, especially the possibility of a Davidsonian (Davidson, 1967a) semantic interpretation as shown in Parsons and as performed in CELT. The logical form below is output from CELT interpreting Parson's example sentence.

```
(exists (?S ?K ?T)
  (and
    (instance ?S Poking)
```

```
        (instance ?K Knife)
        (instance ?T Tuesday)
        (agent ?S Brutus)
        (patient ?S Caesar)
        (time ?S ?T)
        (instrument ?S ?K)))
```

The SUMO-WordNet mappings provide the relation between the English root word *stab* and the formal SUMO term of POKING. Note that this is a ‘subsuming’ mapping in the current version, since there is no direct SUMO equivalent to *stabbing*. They provide a mapping to KNIFE and axioms that state that a KNIFE has the capability of being used as an INSTRUMENT in a CUTTING event. Note that the SUMO terms being referred to here are not words. They are formal terms with definitions in first-order logic. Note also that the relationship between KNIFE and CUTTING is not just a link, but a logical axiom suitable for use in theorem proving. Specifically,

```
(=>
    (instance ?X Knife)
    (capability Cutting ?X instrument))
```

Contrast the form above with what we would have to generate if there were no formal ontology with a mapping to a lexicon, for example:

```
(exists (?S ?K ?T)
  (and
    (instance ?S stabs)
    (instance ?K knife)
    (instance ?T Tuesday)
    (agent ?S Brutus)
    (object ?S Caesar)
    (on ?S ?T)
    (with ?S ?K)))
```

There would be no logical definition of *stabs* or any of the other predicates or terms to explain the meanings to a machine. A human would have to interpret the meaning of the terms in the logical form, leaving us about where we started in having an English sentence that has to be interpreted by a human, rather than understood by a machine.

2.5 Global WordNet

As the English WordNet gained wide acceptance in the natural-language processing community, researchers in other countries began to construct wordnets in their languages. Vossen (1998) coordinated the effort to create eight

European wordnets that follow a common design and are interlinked via an Interlingual Index (ILI). At the time of writing, wordnets exist in over forty languages spoken around the world (Singh, 2002; Sojka *et al.*, 2004, 2006). Besides the obvious advantage for NLP applications in a given language, interconnected wordnets hold great potential for crosslinguistic applications. Furthermore, the construction of wordnets in typologically and genetically unrelated languages sheds light both on the commonalities and the differences in the ways languages map concepts onto words. To facilitate the construction of international WordNets and to enable their mapping, Vossen (1998) conceived of the Interlingual Index, or ILI.

2.5.1 The Interlingual Index

When EuroWordNet – the first international set of wordnets – was begun, it soon became clear that words and synsets could not just be translated from the English wordnet into the European languages; this became even more obvious for the more recent wordnets in Indian and Asian languages. Not only are there language-specific ‘lexical gaps’, seemingly accidental holes where a word in one language has no correspondence in another language, but there are differences in the ways languages structure their words and concepts. Vossen (1998) discusses the case of *spoon* in English and Dutch. Dutch has no exact equivalent for English *artefact*, which serves as a superordinate to a large class of synsets. As a result, the hierarchy where Dutch *lepel* (‘spoon’) is embedded is flatter than that of *spoon*.

EuroWordNet comprises three modules to which the individual languages refer. These are the Top Concept Ontology, the Domain Ontology, and the Interlingual Index (ILI). The ILI initially consisted of all English WordNet (1.5) synsets. Each international wordnet either links its synsets to the matching synsets in the ILI or adds a synset that is not yet in the ILI. The ILI thus becomes the superset of all concepts in all wordnets.

Equivalence relations between the synsets in different languages and Princeton WordNet are made explicit in the ILI. Each synset in the language-specific wordnet has at least one equivalence relation to an entry in the ILI. Thus, synsets linked to the same entry in the ILI can directly map the corresponding synsets and words, allowing for a variety of crosslinguistic applications. ILI entries are also linked to the Top Concept Ontology and the Domain Ontology. For further details and discussion see Vossen, 1998.

As pointed out earlier, lexical ontologies are limited in how they can express word meanings through relations in a graph, plus definitions in natural language. Similarly, much of the meaning of each term in the ILI is given by its name and the English definition, taken from the Princeton WordNet or newly created for a language-specific synset. The model is therefore limited to users who are reasonably fluent in English in addition to the target language. The lack of mathematical axioms defining terms in the ILI means that it cannot be shown mathematically to be consistent or correct. Finally, the meaning of the terms has to rely on a human interpretation of linguistic definition, rather than on a precise mathematical specification. While two people may disagree on aspects of a definition expressed in natural language, there can be no disagreement between two mathematically competent people about the meaning of a mathematical formula, other than whether it faithfully reflects some view of reality.

Many wordnets have been linked to English WordNet and thus also to SUMO. The formal and language-independent nature of SUMO holds some promise in enabling creators of new wordnets to verify these cross-language links by testing them against a formal, logical definition, rather than WordNet's definitions and semantic relations. Given the utility that has been gained from just having a lightweight interlingua of just over 100 terms for EuroWordNet (Vossen, 1998), a more extensive, precise, and language-independent formal ontology holds considerable promise.

2.6 SUMO translation templates

In an effort to make SUMO more understandable to a wider community, we have created a system that performs rough natural-language paraphrasing of the formal axioms that are stated in first-order logic. While it is awkward and does not present an advance in the study of language generation, this system nevertheless allows SUMO users who do not understand formal logic to have a better idea of the axiomatic semantics of the terms. Because the SUMO terms are stated as English words or phrases, having translations of the terms is also required for non-English speakers whether or not they are conversant with logic. There are currently translation templates for English, German, Czech, Italian, Hindi, Chinese (traditional characters and pinyin), and Romanian. Partial translation sets have been created for Tagalog and Cebuano. Korean, Estonian, and Hungarian are under development.

Acknowledgment

This work has been sponsored by the United States Central Intelligence Agency and US Army CECOM. Ian Niles conceived of the SUMO-WordNet mapping project and did almost all of the original mappings.