Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology

Ian Niles and Adam Pease (presenter)
Teknowledge
1800 Embarcadero Rd
Palo Alto CA 94303
650 424 0500
650 493 2645
[iniles | apease]@teknowledge.com

Abstract

Abstract: *Ontologies are becoming extremely useful tools for sophisticated software engineering. Designing applications, databases, and knowledge bases with reference to a common ontology can mean shorter development cycles, easier and faster integration with other software and content, and a more scalable product.*

Although ontologies are a very promising solution to some of the most pressing problems that confront software engineering, they also raise some issues and difficulties of their own. Consider, for example, the questions below:

- *How can a formal ontology be used effectively by those who lack extensive training in logic and mathematics?*
- *How can an ontology be used automatically by applications (e.g. Information Retrieval and Natural Language Processing applications) that process free text?*
- *How can we know when an ontology is complete?*

In this paper we will begin by describing the upper-level ontology SUMO (Suggested Upper Merged Ontology), which has been proposed as the initial version of an eventual Standard Upper Ontology (SUO). We will then describe the popular, free, and structured WordNet lexical database. After this preliminary discussion, we will describe the methodology that we are using to align WordNet with the SUMO. We close this paper by discussing how this alignment of WordNet with SUMO will provide answers to the questions posed above.

keywords: natural language, ontology

1. SUMO

The SUMO (Suggested Upper Merged Ontology) is an ontology that was created at Teknowledge Corporation with extensive input from the SUO mailing list, and it has been proposed as a starter document for the IEEE-sanctioned SUO Working Group [1]. The SUMO was created by merging publicly available ontological content into a single, comprehensive, and cohesive structure [2,3]. As of February 2003, the ontology contains 1000 terms and 4000 assertions. The ontology can be browsed online (http://ontology.teknowledge.com), and source files for all of the versions of the ontology can be freely downloaded (http://ontology.teknowledge.com/cgi-bin/cvsweb.cgi/SUO/).

2. WordNet

WordNet [4,5,6] is an extremely large and freely available online database. The database is divided by part of speech into nouns, verbs, adjectives, and adverbs. The nouns are organized as a hierarchy of nodes, where each node is a word meaning or, as it is termed in WordNet, a synset. A synset is simply a set of English words that express the same meaning in at least one context. For example, {accession, addition} is a synset which expresses the meaning of adding to something. In version 1.6 of WordNet, there are 66, 054 noun synsets, 17,944 adjective synsets, 3,604 adverb synsets, and 12,156 verb synsets

As an example of a record in the WordNet database, consider the following.

```
00047131 04 n 02 accession 0 addition 0 001 @
09536731 n 0000 | something added to what you
have already; "the librarian shelved the new
accessions"; "he was a new addition to the staff"
```

The first part of the record states that the number 00047131 is the unique identifier of the noun synset {accession, addition}. The part of the record between the "@" symbol and the "|" symbol indicates that this synset is directly subsumed by the synset whose identifier is 09536731. This latter synset corresponds to the meaning of acquisition. The final element of the example record above (the text after the "|" symbol) consists of a gloss of the synset and some usage examples.

WordNet is of interest not only because it is a vast repository of lexical data, but also because it is so widely used. It has been leveraged for automated sense-disambiguation, term expansion in IR systems, and the construction of structured representations of document content. In fact, WordNet is so popular that it is almost considered a de facto standard in the NLP community. The many uses to which WordNet has been put are described in a recent book (Fellbaum, 1998).

WordNet is continually updated, and several versions of the database are currently used in Information Retrieval and Natural Language Processing applications. The latest version at this writing is 1.7. However, we decided to use version 1.6 of WordNet for the mapping project described in this paper, because when the project began version 1.7 has not yet been ported to Windows. This should not pose any compatibility problems, because a mapping from the synsets of WordNet 1.6 to the synsets of WordNet 1.7 is due to be released.

3. Mapping Methodology

The first problem decision we had to make in the mapping project was to settle on the relations to be used to map WordNet synsets to SUMO concepts. There are three possible relations of interest: synonymy, hypernymy, and instantiation.

Some examples should make clear these three relations and their use in mapping SUMO concepts to WordNet synsets. Consider the following entry in the WordNet noun database.

```
00008864 03 n 03 plant 0 flora 0 plant_life 0 027
@ . . . | a living organism lacking the power of
locomotion
```

Since this synset is synonymous with the SUMO concept of 'Plant', the WordNet entry is augmented as follows:

```
00008864 03 n 03 plant 0 flora 0 plant_life 0 027
@ . . . | a living organism lacking the power of
locomotion  &%Plant=
```

The '&%' prefix indicates that the term is taken from the SUMO ontology, and the '=' suffix indicates that the mapping relation is synonymy.

Let us now consider a case where a WordNet synset is mapped to a SUMO concept which is broader in meaning than the synset. Consider, for example, the following entry in the WordNet noun file.

```
04719796 09 n 01 Christian_Science 0 001 @
04718274 n 0000 | religious system based on
teachings of Mary Baker Eddy emphasizing
spiritual healing
```

As one might expect, there is no term in the SUMO that is equivalent in meaning to 'Christian_Science'. However, the ontology does contain the more general concept of 'Religious Organization'. Accordingly, we add the annotation '&%ReligiousOrganization+' to the end of the WordNet entry for 'Christian_Science'. Note that the suffix '+' indicates that the concept is a hypernym of the associated synset.

The final sort of mapping relation used in this project is instantiation. This relation indicates that the thing denoted by the WordNet synset is a member of the class denoted by the SUMO concept. Consider, for example, the following entry in the WordNet noun database.

```
00034393 04 n 02 Underground_Railroad 0
Underground_Railway 0 001 @ 00032687 n 0000 |
abolitionists secret aid to escaping slaves; pre-
Civil War in US
```

In this case, the most closely related SUMO concept is 'Organization'. However, this relationship is not one of equivalence in meaning, nor one of subsumption of meaning. The Underground Railway is a particular organization. We indicate this fact by adding the annotation "&%Organization@" to the end of the entry for 'Underground_Railway'.

The WordNet database augmented with SUMO mappings can be exploited in a variety of ways. For example, an application that already uses WordNet can be reconfigured to take advantage of the new SUMO mappings field. If, for some reason, one does not want to make use of all of the augmented files, it will be a simple matter to write a script that extracts the synset/SUMO concept associations from the augmented file and writes them to a new, dedicated file. Finally, if one wants to make use of the mappings within a knowledge-based system, it is a simple matter to write a script that uses the mappings to populate a knowledge base with reverse pointers to

WordNet. In the SUMO, these reverse pointers would be formulas of the following forms:

```
(subsumingExternalConcept <SUMO concept> <WordNet
synset ID> WordNet1.6)

(synonymousExternalConcept <SUMO concept>
<WordNet synset ID> WordNet1.6)

(instance <WordNet synset ID> <SUMO concept>)
```

4. Mapping Examples

In many cases, the mappings from WordNet to the SUMO posed no practical or theoretical problems. In fact, most of the high-level notions in the WordNet database found a ready equivalent in the SUMO. Consider, for example, the following augmented noun entries:

```
00008019 03 n 06 animal 0 animate_being 0 beast 0
brute 0 creature 0 fauna 0 . . . | a living
organism characterized by voluntary movement
&%Animal=

00008864 03 n 03 plant 0 flora 0 plant_life 0 . .
. | a living organism lacking the power of
locomotion &%Plant=

00009457 03 n 02 object 0 physical_object 0 . . .
| a physical (tangible and visible) entity; "it
was full of rackets, balls and other objects"
&%Object=
```

All of these cases are unproblematic, and the many others like them gave us much encouragement during early stages of the mapping project.

Nevertheless there were some challenging cases that are worth examining closely. Consider, for example, one of the WordNet synsets for 'Space'.

```
00015975 03 n 01 space 0 003 @ 00013018 n 0000 %p
00014887 n 0000 %p 06271859 n 0000 | the
unlimited 3-dimensional expanse in which
everything is located; "they tested his ability
to locate objects in space"
```

It was problematic to relate this notion to the SUMO, because the SUMO does not have a concept of "Space" and it is not immediately apparent whether and how such a concept would be useful for knowledge engineering and data modeling tasks. This problem became much more tractable when we considered the parallel notion of "Time", which is represented with the concept of 'TimeMeasure' (and its subsumed classes) in the SUMO. This fact led us to wonder if perhaps there was some concept of measure that could be similarly used to represent "Space". Finally, we decided that the SUMO concept of 'LengthMeasure' captures the quantitative aspect of the common sense notion of "Space" in much the same way that 'TimeMeasure' embodies the quantitative aspect of "Time". Accordingly, we augmented the synset entry above with the annotation "&%LengthMeasure".

Another interesting mapping case concerns WordNet synsets that have an irreducible subjective component. Consider, for example, the following synsets:

```
00082055 04 n 01 best 0 002 @ 00503611 n 0000 !
00082178 n 0101 | the supreme effort one can
make: "they did their best"

00125560 04 n 01 stunt 0 002 @ 00021392 n 0000 ~
00277241 n 0000 | a difficult or unusual feat;
usually done to gain attention

00025630 04 n 02 going 0 sledding 1 001 @
00020977 n 0000 | advancing toward a goal;
"persuading him was easy going" or "the proposal
faces tough sledding"
```

The attribution of these terms involves a criterion which varies from subject to subject and even with respect to the same subject over time, for all of us have different ideas at different times about what is "best", "difficult", etc. Since the SUMO is supposed to to be a repository of precisely defined concepts, these concepts have to be objective ones. Nevertheless, we decided that it might be useful to have a general SUMO concept for the many WordNet synsets like the ones above. Accordingly, we defined the concept of 'SubjectiveAssessmentAttribute' and made it an immediate subclass of 'NormativeAttribute' in the SUMO.

Another interesting sort of case arises when a single concept from the SUMO maps to more than one synset in WordNet, or vice versa. In some cases, WordNet posits a linguistic distinction which does not correspond to a logical difference. Consider, for example, the two following synsets:

```
00002086 03 n 04 life_form 0 organism 0 being 0
living_thing . . . | any living entity

00002880 03 n 01 life 0 002 @ 00002086 n 0000 ~
05988126 n 0000 | living things collectively;
"the oceans are teeming with life"
```

These two synsets mean essentially the same thing, but the first emphasizes being an instance of the general class of living things while the second denotes this class directly. Although this distinction may have linguistic importance, it does not have any bearing on knowledge engineering needs. For this reason, both synsets have been assigned the annotation "&%Organism=". For an example of one synset mapping to more than one SUMO concept, consider the following entry in WordNet:

00128951 04 n 02 substitution 0 exchange 1 004 @ 00125689 n 0000 ~ 00129213 n 0000 ~ 00129804 n 0000 ~ 00129915 n 0000 | the act of putting one one thing or person in the place of another: "he sent Smith in for Jones but the substitution came too late to help

It's clear that this notion of substitution involves removing something from a particular place and then putting something else into that same place. However, it is very difficult to formulate precise temporal and spatial constraints for this substitution. For this reason, we simply augmented the entry above with the annotation "&%Removing+ &%Putting+".

5. Motivation for the Mapping

In the introduction to this paper, it was noted that there are three important issues that arise in connection with ontologies.

· How can a formal ontology be used effectively by those who lack extensive training in logic and mathematics?

· How can an ontology be used automatically by applications (e.g. Information Retrieval and Natural Language Processing applications) that process free text?

· How can we know when an ontology is complete?

The WordNet/SUMO mappings will help resolve each of these issues. In particular, these mappings can function as a natural language index to the concepts in the ontology, as a bridge between these structured concepts and the free text that is processed by an ever increasing number of applications, and as a "completeness check" on the content of the ontology.

Let us discuss each of the three issues in turn. First of all, the mappings between WordNet and the SUMO can be regarded as a natural language index to the SUMO. Thus, we have developed a tool which permits the user to enter English terms and which returns SUMO concepts that are associated with the input terms via WordNet synsets. By interacting with this tool, the user is able to see all SUMO concepts that are related to natural language terms of interest, and this makes it much easier for him/her to do knowledge engineering and data modeling tasks with the ontology. This tool has been integrated with the SUMO browser, which is available online at: http://ontology.teknowledge.com/.

Aside from facilitating the creation of SUMO-compliant knowledge and data elements, the mappings may also be an important knowledge source for Natural Language Processing applications. The mappings can be used by these applications to assign the structured meanings of the SUMO to free text. The simplest way of doing this would be simply to assign every SUMO concept to every word which is related to it via a WordNet synset. More sophisticated approaches could use some sort of sense-disambiguation algorithm to pinpoint the precise SUMO concept which is intended in a given context. In either case, the document representations consisting of SUMO concepts could then be used to create automatically generated summaries or they could be used to facilitate semantic searching.

Aside from resulting in a knowledge source that should prove very useful to Natural Language Processing applications and human users of the ontology, the mapping process has functioned as a completeness check on the SUMO. As we assigned SUMO concepts to WordNet synsets, we came across some cases where the most specific subsumer in the SUMO for a given synset was too broad in meaning. In other words, from time to time we found content in WordNet that was not part of the SUMO but that should be, we judged, part of an upper-level ontology. Consider, for example, the following entry in the WordNet database.

00038917 04 n 01 failing 0 002 @ 00038702 n 0000 ! 00037826 n 0101 | failure to reach a minimum required performance; "his failing the course led to his disqualification"

This entry and the many others in WordNet relating to the level of performance in a competitive situation convinced us that we needed a subclass of 'Attributes' for qualities related to competition, e.g. passing, failing, winning, losing, etc. Note that this class is not a subclass of the concept 'SubjectiveAssessmentAttribute' discussed in the previous section, because in some cases there is an objective fact of the matter about a participant's standing in a particular competition, e.g. when one player checks another in chess. Other additions to the SUMO that have been motivated by WordNet include the concept 'EmotionalState', many concepts in the 'Process' branch of the SUMO, and 'SoundAttribute' (a subclass of 'Attribute'). To the extent that WordNet synsets have suggested concepts that are appropriate for an upper-level ontology, definitions and axioms corresponding to these synsets have been crafted and added to the SUMO. In this way, we believe that we have refined the SUMO into an ontology that can be used to express anything that anyone would every want to say in a formal context.

The SUMO/WordNet mapping project was completed in December 2002. All of the synsets in WordNet 1.6 have been mapped to at least one SUMO concept. The files containing all of the SUMO/WordNet mappings can be freely downloaded from the SUMO ontology web page: http://ontology.teknowledge.com/.

6. References

[1] SUO, (2003), The IEEE Standard Upper Ontology web site, http://suo.ieee.org

[2] Niles, I & Pease A. (2001) "Towards A Standard Upper Ontology." In Proceedings of FOIS 2001, October 17-19, Ogunquit, Maine, USA.

[3] Niles, I & Pease A. (2001) "Origins of the IEEE Standard Upper Ontology." In Working Notes of the IJCAI-2001 Workshop on the IEEE Standard Upper Ontology, August 4-10, Seattle, Washington, USA.

[4] Fellbaum, Christiane. (1998) "WordNet: An Electronic Lexical Database." MIT Press.

[5] Miller, G. A. (1993) "Nouns in WordNet: A Lexical Inheritance System."

[6] Miller, G. A., Beckwith, R., Fellbaum, Christiane, Gross, Derek, and Miller, K. (1993) "Introduction to WordNet: An On-line Lexical Database."