Combining and Standardizing Large-Scale, Practical Ontologies for Machine Translation and Other Uses

Eduard Hovy

Information Sciences Institute
of the University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292-6695
U.S.A.
tel: +1-310-822-1511
fax: +1-310-823-6714
email: hovy@isi.edu

Abstract

The increasing proliferation of large-scale knowledge bases, used in Computational Linguistics, Artificial Intelligence, Expert Systems, Databases, and related fields, creates the need for an integrated knowledge framework in which their contents can be related to one another. Over the past few years, the availability of several large-scale symbol taxonomies and axiomatizations (often called Ontologies) has generated considerable interest, mostly for how they can be used to assist with such integration. This paper outlines recent attempt at USC/ISI to create a single large Ontology for general free use over the Web, as performed under the aegis of the ANSI Ad Hoc Committee on Ontology Standardization. Eventually, this so-called Reference Ontology may combine the existing large-scale ontologies WordNet, MIKROKOSMOS, EDR, CYC, and SENSUS into a single framework. The paper provides semi-automated cross-ontology concept alignment heuristics and describes their use, operation, and results.

1. Introduction

In a world of increasing computerization, it is inevitable that different computer systems represent and treat the same basic entity or notion in different ways. The concept of *Person*, for example, connotes something different to a medical specialist (and hence to his or her computer system) than it does to a Census Bureau worker (and his or her system).

When the systems work in isolation from each other, there is no problem. Humans do the 'translating' between them. But when the systems are to be integrated, or just try to share data, the problem arises. Incommensurate views of the 'same' object mean incompatible systems, and incompatible systems means no data sharing, no knowledge transfer, and a necessary duplication of effort.

The only solution is to try to map terms to each other, either directly, using large bi-domain correspondence tables, or indirectly, using some neutral internal terminology that is structured to facilitate correspondence. The former approach has the disadvantage of requiring N^2 sets of mappings for N domains (one mapping between each two domains). The latter approach requires only N mappings (one mapping between each domain and the central one). When the domains are large, and there are many domains, the latter approach requires distinctly less work.

The advantages of a single, neutral ontology are fairly obvious:

1. They help standardize terminology. The name chosen for a term or a relation can be used by others in new domain modeling work, and can be used to effect communication among systems. They do not however enforce standards—something that is called *Income* in one domain can still be called *Income* there, even though the ontology name for the same concept may be *Salary*.

2. They assist knowledge transfer. When a portion of domain ontology, constructed by one expert in one area, is appropriate for another area, another expert working in the second area can easily find the portion via the neutral ontology and integrate it into the second domain ontology.

3. They facilitate interoperability. When a computer system is built to perform some actions using the neutral set of terms, it can more easily be ported and adapted to new domains that also use the neutral ontology.

2. The ANSI Ontology Standardization Effort

Practical experience makes clear that integrating different ontologies is fraught with difficulty.

However, so many ontologies, termsets, thesauri, and other terminology collections are being built and used today that it would seem wise to investigate automated methods of ontology integration. Even if they are only partially successful, these methods can significantly speed up new ontology efforts and system porting efforts. This point is illustrated in (Swartout et al., 1996), who describe practical experiences of the benefits of building domains models by starting with a large ontology.

In 1996, representatives from several Ontology projects and others interested in the issues formed an Ad Hoc Group on Ontology standards. Meeting twice a year, this group is a subcommittee of the ANSI Committee X3T2, currently called NCITS, and is headed by Mr. Robert Spillers, formerly of IBM. It contains representatives from various universities (including Stanford University and the NL Group at USC/ISI), research laboratories (including LADSEB and Lawrence Berkeley Laboratories), companies (including CYC and TextWise), U.S. Government officials, and private individuals (including John Sowa).

The group is creating a short document that will serve as a model ontology standard and has commissioned some work on the creation of an example of such a standard, called the Reference Ontology. An early version of the topmost portions of this Reference Ontology has been created out of the top regions of SENSUS (USC/ISI) and CYC (Cycorp).

3. SENSUS and the Reference Ontology

One way to create a standard ontology is to begin with a large, high-level, but rather content-neutral ontology, and then to systematically add to it other ontologies, termsets, data definitions, etc. To make the process more open to available data collections of any form, we define an Ontology rather loosely, as follows:

> An ontology is set of terms, associated with definitions in natural language (say, English) and, if possible, using formal relations and constraints, about some domain of interest, used in their work by human, data bases, and other computer programs.

Under this formulation, a termset, a data dictionary, and even a metadata collection are all types of ontologies. We specifically want it this way because we aim to incorporate information from any of these sources, if appropriate, into the Reference Ontology.

We first describe the starting point, and then discuss the problems of merging (automated assistance tools, overcoming and sidestepping incommensurabilites, etc.) below.

As initial central and high-level ontology we take SENSUS (Knight and Luk, 1994), built at USC/ISI to serve as the internal mapping structure (the Interlingua termbank) between lexicons of Japanese, Arabic, Spanish, and English, in several projects:

- GAZELLE: machine translation (Knight et al., 1995),
- SUMMARIST: multilingual text summarization (Hovy and Lin, 1998),
- C*ST*RD: multilingual text retrieval and management (Hovy, 1997).

The lexicons contain over 120,000 root words (Japanese), 60,000 (Arabic), 40,000 (Spanish), and 90,000 (English). SENSUS terms serve as connection points between equivalent language-based words.

SENSUS currently contains approx. 70,000 terms, linked together into a subsumption (isa) network, with additional links for part-of, pertains-to, and so on. SENSUS is a rearrangement and extension of WordNet (Miller, 1990) (built at Princeton University on general cognitive principles), retaxonomized under the Penman Upper Model (Bateman et al., 1989) (built at USC/ISI to support natural language processing).

SENSUS can be accessed publicly via the Web, at *http://mozart.isi.edu:8003/sensus/sensus_frame.html* via the Ontosaurus browser (Patil, 1996).

We consider SENSUS a good starting ontology because:

- it contains a large number of terms;
- the terms cover most of the general human areas of experience;
- it does not contain any particular domains already;
- it does not make deep ontological commitments to particular theories of existence, space, time, money, emotion, cognition, etc.;
- its notation is simple and easy to read.

4. Creating the Reference Ontology

When merging ontologies, three situations can arise (Hovy and Nirenburg, 1992). Given two most-closely related terms, one from each ontology, either:

1. The two terms are exactly equivalent—in which case they can be directly aligned;
2. One term is more general than the other—in which case the more specific term (and its subordinates, and possibly its siblings) can be integrated below the more general one;
3. The terms are incompatible (that is, identifying them would cause definitional and relational problems among other terms)—in which case either (1) one of the terms must be rejected and not incorporated, or (2) one of the terms and the other terms depending on it must be redefined, or (3) a separate 'microtheory' must be created in which the terms, and all the other terms depending on it, exist in parallel (Lenat and Guha, 1990), or (4) a weaker version of the offending term can be incorporated, without the definitions or relations that caused the inconsistency.

The most problematic case is the last one. In SENSUS, we do not support microtheories; in cases on incompatibility we work with the domain builders to either weaken or redefine the terms in either ontology, or to leave the term (and its dependents) out of the merged central ontology. In that case, the domain ontology will contain more information than the merged one, which simply means that some external user or system will not be able to exploit fully the domain ontology's content.

Integration proceeds in several stages:

1. Initially, direct term identification: a core set terms from the central and the domain ontologies that have equivalent meanings are identified and aligned;
2. Next, content merging: the domain's core terms' definitional constraints, inter-term relationships, and related axiomatizations are incorporated into their aligned equivalents in the central ontology, to the extent possible without overcommitting to a particular world view or causing internal inconsistencies;
3. Next, wider term alignment: when the domain core terms are integrated, the set of domain concepts with near-misses in the central ontology are addressed. Near-misses occur either because the domain terms are more specialized than or slightly different from the central ontology's terms (and hence that there are no direct equivalents for them, only ancestral ones), or because the domain terms' definitions are not clearly enough specified to allow them to be incorporated with certainty. By a combination of manual and automated work, and consultation with the domain ontology's builders, we identify the appropriate locations for such near-misses and incorporate them.
4. Next, inconsistency resolution: several types of inconsistency may occur. For each inconsistency, the type is identified and, as mentioned above, the appropriate actions are performed: either redefinition (in consultation with the domain experts), or weakening, or, if that fails, omission.
5. The cycle of steps 3 and 4 are repeated, with increasingly distant terms. In practise, given the use of automated alignment tools (described below) to identify close terms, and given the incorporation *en bloc* of whole domain subtrees that are more specialized than the central ontology, the number of distant terms for which incompatibilities exist may diminish rapidly.

5. Semi-Automated Ontology Alignment

5.1 Background

At first glance, it might seem impossible to align two ontologies automatically. Almost all ontologies, after all, depend to a large degree on non-machine-interpretable information such as concept names and English concept definitions to distinguish their concepts.

However, the technique described here, and recently tested on several different sources (both ontologies and dictionaries) offers some promise of success. This technique consists of the following:

1. a set of heuristics that make initial cross-ontology alignment suggestions

2. a function for integrating their suggestions

3. a set of alignment validation criteria and heuristics

4. a repeated integration cycle

5. an evaluation metric

This technique is not yet fully explored. Different portions of the technique have been used in different projects.

In order to enable Japanese- and Spanish-to-English machine translation in GAZELLE, Knight and Luk devised three heuristics (definition match, hierarchy match, bilingual match) to link English, Spanish, and Japanese words derived from various online dictionaries into SENSUS (Knight and Luk, 1994). They found 11,128 noun matches with 96% accuracy. To extend this linkage for Japanese verbs, (Okumura and Hovy, 1994) developed a fourth heuristic, matching on the case role structure of verbs.

In a subsequent application, a variant of this technique was tried with the linking of several ontologies. First, in 1996, the topmost regions of SENSUS and CYC (Lenat and Guha, 1990) were aligned (Hovy, 1996; Lehmann, 1997). The initial experiences served both ontologies well in terms of cleaning up hierarchy structure and sharpening definitions.

Later, in 1997, the 4790-concept ontology MIKROKOSMOS (Mahesh, 1996) was semi-automatically aligned and verified with the topmost 6768 concepts of SENSUS. During this procedure, several validation criteria were developed and tested. They are listed in Section 5.5.

In this paper we describe only the most recent version of this alignment and integration process, since it has evolved out of the previous work. We focus on the process of aligning the principal ontology, SENSUS (marked by "S@"), and the new ontology, MIKROKOSMOS (marked by "M@"). We only outline the subsequent validation steps that preceded the actual merging of the two ontologies, since that work is not complete.

5.2 Alignment Cycle

The cycle for creating alignment suggestions is as follows:

1. Load the initial ontologies (in original state) and load all patches (previous alignments and upgrades, as suggested by validation editing, etc.), to bring the ontologies into the most recent partially aligned state.

2. For all unaligned concepts, create a new set of cross-ontology match scores, running one or more of the heuristics NAME, DEF, TAX. Save the aligned pairs and scores.

3. Create a new set of alignment suggestions by combining the above match scores using the combination function. Sort them in descending (combined-score) order. Display the distribution of possible alignments and ask the user for a cutoff value. Extract and save the results.

4. Manual step: Check the suggestions by hand, deciding which are correct alignments, which are wrong, and which are near misses (at the moment, nothing special is done with near misses). Save the three sets of concepts separately.

5.3 Alignment Suggestion Heuristics

Three classes of alignment heuristics have been reported in the literature.

1. **Text matches** operate upon strings of letters, e.g.:

- Concept name matches (cognate matching): similar-enough names should provide some evidence that the developers consider the concepts similar, when the names are in the same language.
- Definition matches (word processing and overlap measures, such as first used in (Knight and Luk, 1994)): similar-enough natural language definitions should also provide some evidence of concept similarity.

2. **Hierarchy matches** exploit the taxonomization structure of ontologies. They include:

- Ambiguity filtering by shared superconcept (such as used in (Knight and Luk, 1994)): when a concept can be aligned to more than one alternative, consider only those whose superconcepts are somewhere aligned to the superconcepts of the target concept.
- Semantic distance (link distance) measures (such as the inverse-depth formula discussed in (Agirre et al., 1994)).

3. **Data item or form matches** operate upon the computer-related content or formal restrictions of the concepts themselves. They include:

- Internal cross-links among sets of concepts (such as the multilingual dictionary links employed by (Ageno et al., 1994; Rigau and Agirre, 1995)).
- Slot-filler restrictions (such as the verb case role restrictions used by (Okumura and Hovy, 1994)).

In the alignment of MIKROKOSMOS and the top of SENSUS, three alignment measures were used to create suggestions.

NAME (cognate) match: This match compares the names N1 and N2 of two concepts. It considers decreasing substrings of N1, chopping off the left. Composite-word names are split into separate words, and the maximum score is returned. Names shorter than 3 letters are ignored.

```
NAMESCORE := square of number of letters matched
             + 20 points if words are exactly equal
               or 10 points if end of match coincides
                  with end of hyphenated section of name.
```

Example name match suggestions and scores:

```
(alignval '|S@cuisine|
 '((NAME M@LIMOUSINE 26) (NAME M@VINE 19)
   (NAME M@MORPHINE 19)
   (NAME M@ENGINE-GOVERNOR 19)
   (NAME M@BUSINESS-COVERAGE-OF 16)
   (NAME M@AGRIBUSINESS-ACTIVITY 16)
   (NAME M@TABLE-LINEN 9) (NAME M@TRAINER 9)
   ... 120 more ...
 ))
(alignval '|S@Free World|
 '((NAME M@PERCENT-OF-WORLD-POPULATION 46)
   (NAME M@WORLD 35)
 ))
(alignval '|S@kept woman|
 '((NAME M@AMMAN 19)
   (NAME M@MAN-MADE-ATTRIBUTE 19)
   (NAME M@NAME-HUMAN 19) (NAME M@GERMAN 19)
   (NAME M@TRANSPORT-HUMAN 19)
   (NAME M@HUMAN-LIVING-EVENT 19)
   (NAME M@HUMAN-VOICE 19) (NAME M@HUMAN 19)
   ... 65 more ...
 ))
```

DEFINITION match: This match compares the English definitions D1 and D2 of two concepts. First, both definitions are separated into individual words (apostrophes, hyphens, etc., are removed) and all words are demorphed, using ISI's English demorpher (Lin 97). Function words and other stop words are removed, but duplicates are kept. The processed definition of M@FOOD, for example, is

```
("any" "substance" "that" "can" "be" "metabolized"
 "organism" "give" "energy" "build" "tissue")
```

With the remaining words, three values are computed:

- strength = ratio of number of words shared in both definitions to number of words in the shorter definition,
- reliability = number of shared words,
- defscore = strength * reliability.

DEFSCORE := (Shared(D1,D2) / min{D1,D2})*Shared(D1,D2)

Examples of Definition match suggestions and scores (strength, reliability, score):

```
(alignval '|S@cuisine|
 '((DEF M@KITCHEN (0.62  5   3.12))
   (DEF M@CHEESE (0.62  5   3.12))
   (DEF M@FOODSTUFF (0.62  5   3.12))
   (DEF M@PET-FOOD (0.62  5   3.12))
   (DEF M@CUTLERY (0.50  4   2.00))
   (DEF M@RACETRACK (  0.37 3   1.12))
   (DEF M@COOK (  0.40 2   0.80))
   ... 5 more ...
 ))
(alignval '|S@Free World|
 '())
(alignval '|S@kept woman|
```

```
'((DEF M@PATIENCE-ATTRIBUTE (0.40 2 0.80))
  (DEF M@ACRYLIC-PAINT (0.33 2 0.67))
  (DEF M@SHAREHOLDER (0.33 2 0.67))
  (DEF M@SCALAR-ATTRIBUTE (0.33 2 0.67))
  (DEF M@FLOUR-MILL ( 0.25 2 0.50))
  (DEF M@LIMITED-LIABILITY-PARTNER ( 0.25 2 0.50))
  (DEF M@LEVERAGE ( 0.25 2 0.50))
  ... 45 more ...
  ))
```

TAXONOMY match: for a given SENSUS concept, this match collects all the concepts in the MIKROKOSMOS taxonomy that are 'closer' than 10 links to it. The algorithm traverses the taxonomy in both superconcept and subconcepts directions, jumping across into the MIKROKOSMOS ontology as soon as possible. The inverse of the link distance is the match score.

```
TAXSCORE := 1 / number-of-links
```

```
(alignval '|S@end>come out|
'((TAX M@SOCIAL-EVENT 0.17)
  (TAX M@EMANATE 0.17)
  (TAX M@EMIT-LIGHT 0.17)
  (TAX M@EMIT-SOUND 0.17)
  (TAX M@REFLECT-LIGHT 0.17)
  (TAX M@EXTRACT 0.17) (TAX M@APPLY-FORCE 0.20)
  (TAX M@PACK 0.20)
  ... 22 more ...
  ))
(alignval '|S@scatterbrain|
'((TAX M@INTANGIBLE-OBJECT 0.17)
  (TAX M@MENTAL-OBJECT 0.17)
  (TAX M@SALAMANDER 0.17) (TAX M@FROG 0.17)
  (TAX M@CANARY 0.17) (TAX M@RAPTOR 0.17)
  (TAX M@CHICKEN 0.17) (TAX M@TUNA 0.17)
  (TAX M@SALMON 0.17) (TAX M@FISH 0.20)
  (TAX M@RESTAURANT-ROLE 0.17)
  (TAX M@SOCIAL-ROLE 0.25) (TAX M@EGO 0.20)
  ... 69 more ...
  ))
```

5.4 Combination Function

Several combination formulas were created and tried. Each formula had to have the following characteristics:

- it must increase with increasing values of NAME, DEF, and TAX
- it must 'normalize' the heuristics' scores
- it must mitigate the NAME scores' tendency to grow large quickly
- it must mitigate the TAX scores' tendency to diminish quickly
- it must return a nonzero score if any heuristic returned a nonzero score

The suggestion combination formula eventually used was:

```
SCORE := sqrt(NAMESCORE) * DEFSCORE
                          * (10 * TAXSCORE)
```

with the proviso that if NAMESCORE or DEFSCORE are zero, they are replaced by 1, and if TAXSCORE is 0, it is replaced by 0.01. Typically, the alignment scores ranged between zero and 16.

Since too little correctly aligned data was available, no automated training of combination formula coefficients was possible.

5.5 Validation Heuristics

As mentioned above, the process of validation and ontology merging is not yet complete. The heuristics mentioned here were developed by Tom Russ, Hans Chalupsky, and Eduard Hovy, all of USC/ISI. They are:

Hierarchy matching (Chalupsky and Hovy):

1. **New Superconcept test**: When a MIKROKOSMOS concept *Cm* is merged into SENSUS, it acquires all the SENSUS-superconcepts of its sister concept *Cs*. Some of these may be equivalent to MIKROKOSMOS concepts themselves. Did the alignment create any new superconcepts of *Cm* in MIKROKOSMOS (i.e., concepts that were not already superconcepts of *Cm*)? If so, *Cm* is wrongly aligned, for it may not acquire new MIKROKOSMOS superconcepts though a merge.

2. **Disjunction test**: Is mutual disjointness observed after alignment? That is, if *Cm* (or its parents, or its children) participate in mutual disjointness requirements, and they themselves have SENSUS equivalents, would the inclusion of *Cm* into SENSUS violate these requirements?

3. **Cycles/Bowties test**: A cycle is created when a superconcept chain leads back upon itself; a bowtie is created when concepts *Cm1* and *Cm2* in MIKROKOSMOS are respectively aligned to *Cs1* and *Cs2* in SENSUS, but while *Cm1* is superordinate to *Cm2*, *Cs1* is subordinate to *Cs2*. If merging a MIKROKOSMOS concept into SENSUS causes a cycle or bowtie to appear, its alignment is incorrect.

Content-based matching (Russ):

1. Several examples of inconsistent or missing names, or of type constraints violations, were encountered during the attempted merge. Automated procedures that checked each concept name and each filler of each slot were developed.

2. Both SENSUS and MIKROKOSMOS employ only one superconcept link, and hence overload the relationship (they should be converted to use both *Subclass-of* and *Element-of* instead). Since they misuse *isa* in different ways, inconsistencies appear.

3. SENSUS does not enforce property inheritance requirements; MIKROKOSMOS usually does. The resulting inconsistencies have to be reconciled.

6. Experiments and Results

6.1 Performance Figures

We report on the experiments aligning the MIKROKOSMOS ontology with the top region of SENSUS.

MIKRO: 4790 concepts
SENSUS: 6768 concepts (Pangloss origin: 367 concepts
WordNet origin: 6401 concepts)

The number of possible alignment arangements is enormous: 6768! / 4790! The number of concept pairs that each alignment heuristic has to consider is over 32.4 million (i.e., 4790 x 6768).

The alignment suggestion cycle was repeated 5 times. Alignment suggestions were computed once only by the NAME and DEFINITION match heuristics, at the beginning of the sequence, since these suggestions never change (the names and definitions remain the same). However, since every new alignment creates a new bridge across the ontologies, the TAXONOMY match heuristic is re-run every time. Its new suggestions, combined with the other heuristics' suggestions, produce the new suggestions in every run.

For each run, an appropriate matching value combination score cutoff level was chosen, by simple inspection of the score distribution (under the desire to avoid too much manual work checking low-scoring suggestions). The system's results of run 3, with cutoff value 7.8, are:

```
Now excluding from consideration all near-misses
Done.  6768 concepts have suggested alignments.

Alignment suggestion scoring statistics:
Numbers of possible alignment suggestions:
  NAME = 461628   DEF = 90701   TAX = 205143

NAME:
  score    number of suggestions
  >= 100 :  804
  >= 70  :  3036
  >= 50  :  2398
  >= 30  :  33889
  >= 25  :  67075
  >= 20  :  0
  >= 15  :  218879
  >= 10  :  1791
  >= 5   :  128785
  >= 0   :  4971
DEF:
  score    number of suggestions
  >= 3.00 :  448
  >= 2.00 :  2074
  >= 1.50 :  2997
  >= 1.00 :  16233
  >= 0.75 :  9461
  >= 0.50 :  34198
  >= 0.00 :  25290
TAX:
  score    number of suggestions
  >= 1.000 :  0
  >= 0.500 :  0
  >= 0.300 :  0
  >= 0.210 :  0
  >= 0.190 :  5810
  >= 0.150 :  0
  >= 0.140 :  35219
  >= 0.120 :  0
  >= 0.105 :  164114
  >= 0.090 :  0
  >= 0.000 :  0

Now saving suggestion definitions...
Now saving suggestion alignments...
Distribution of alignment scores:
  >= 30 :  0
  >= 20 :  0
  >= 15 :  1
  >= 12 :  2
  >= 10 :  6
  >= 8  :  160
  >= 6  :  739
  >= 5  :  493
  >= 0  :  155430
Please enter a cutoff value greater than 0: 7.8

Done: 170 suggestions collected.
170 suggestions to be saved...
```

The results of the five runs were as follows:

cutoff	1.4	10	7.8	12	15
new heur.	NAME,DEF,TAX	TAX	TAX	TAX	TAX
total	187	151	170	218	241
correct	73	11	18	36	106
near	51	92	51	60	2
wrong	63	48	101	122	39

After 5 runs, the heuristics provided the following results:

correct suggestions: 244 = 3.6% of all SENSUS concepts
incorrect suggestions:
- near miss: 256 = 3.8%
- far wrong: 383 = 5.6%

That is, the combined heuristics extracted 883 suggestions for validation (= 2.72% of the total number of pairs, or 13% of the portion of SENSUS under consideration). Of these, 244 (= 27.6%) were correct, 383 (= 43.4%) were incorrect, and 256 (= 30.0%) were nearly correct.

Recall is more difficult to measure; it would involve manually searching the entire MIKROKOSMOS to see how many concepts the heuristics should have picked out but did not. Informal estimates, based on browsing the MIKROKOSMOS files and seeing if plausible-looking concepts were picked up for suggested alignment, place the Recall level fairly high, at over 85%.

6.2 Alignment Examples

It is informative to inspect some of the suggestions returned by the heuristics. Correct alignments occurred when name, definition, and taxonomy heuristics combined:

|S@foodstuff<food|
= a substance that can be used or prepared for use as food
superconcepts: (|S@food|)

M@FOODSTUFF
(COMB = 13.35 NAME = 91 DEF = 10.00 TAX = 0.14)
= a substance that can be used or prepared for use as food
superconcepts: (M@FOOD M@MATERIAL)

|S@change of location,move|
= the act of changing your location from one place to another
superconcepts: (S@MOTION-PROCESS)

M@MOTION-EVENT
(COMB = 4.59 NAME = 26 DEF = 4.50 TAX = 0.20)
= a physical-event in which an agent changing location moves from one place to another
superconcepts: (M@CHANGE-LOCATION)

Often, however, it was difficult to decide whether an alignment suggestion was in fact correct or not:

|S@man<soul|
= the generic use of the word to refer to any human being "it was every man for himself"
superconcepts: (S@PERSON)

M@HUMAN
(COMB = 1.43 NAME = 19 DEF = 0.00 TAX = 0.33)
= homo sapien
superconcepts: (M@PRIMATE)

|S@library>bibliotheca|
= a collection of literary documents or records kept for reference
superconcepts: (|S@aggregation|)

M@LIBRARY
(COMB = 2.74 NAME = 59 DEF = 3.57 TAX = 0.00)
= a place in which literary and artistic materials such as books periodicals newspapers pamphlets and prints are kept for reading or reference an institution or foundation maintaining such a collection
superconcepts: (M@ACADEMIC-BUILDING)

|S@geisha|
= a Japanese woman trained to entertain men with conversation and singing and dancing
superconcepts: (|S@adult female| |S@Japanese<Asian|)

M@GEISHA
(COMB = 1.54 NAME = 46 DEF = 2.27 TAX = 0.00)
= a Japanese girl trained as an entertainer to serve as a hired entertainer to men
superconcepts: (M@ENTERTAINMENT-ROLE)

Both |S@library>bibliotheca| and |S@geisha| seem to align well to their partners, until one considers their superconcepts, which focus on different aspects (or qualia) of their nature. A library is both a collection of books and a building that houses such a collection; the word can be used to mean both in the same sentence quite naturally. Differentiating such senses is the focus of some recent work in both knowledge representation (Guarino, 1997) and computational lexicon work (Pustejovsky, 1996). This example illustrates why further ontological validation is required by machine; even with the human reading quite carefully, it is easy to miss distinctions that appear in ancestral nodes.

Finally, the alignment process can even be used to find errors...is an archipelago land or sea?

|S@archipelago|
= many scattered islands in a large body of water
superconcepts: (|S@dry land|)

M@ARCHIPELAGO
(COMB = 1.522 NAME = 131 DEF = 1.33 TAX = 0.00)
= a sea with many islands
superconcepts: (M@SEA)

7. Conclusion

This work is far from being complete. New alignment suggestion heuristics are easy to think up, especially when the ontologies contain many concept interrelationships. More sophisticated validation techniques are required, especially ones that automatically determine the extent of the effects of an alignment inconsistency. A simple browser-editor interface would facilitate manual checking.

Still, we found it greatly useful to be able to focus in on only about 13% of the SENSUS terms, and of these, to find just over half to be correct or nearly correct.

Acknowledgments

Many thanks to Bruce Jakeway for patiently sifting through ontologies and remaining cheerful.

References

Ageno, A., I. Castellon, F. Ribas, G. Rigau, H. Rodriguez, A. Samiotou. (1994). TGE: Tlink Generation Environment. In Proceedings of the 15th COLING Conference, Kyoto, Japan.

Agirre, E., X. Arregi, X. Artola, A. Diaz de Ilarazza, K. Sarasola. (1994). Conceptual Distance and Automatic Spelling Correction. Proceedings of the Workshop on Computational Linguistics for Speech and Handwriting Recognition. Leeds, England.

Bateman, J.A., Kasper, R.T., Moore, J.D., and Whitney, R.A. (1989). A General Organization of Knowledge for Natural Language Processing: The Penman Upper Model. Unpublished research report, USC/Information Sciences Institute, Marina del Rey, CA.

Guarino, N. (1997). Some Organizing Principles for a Unified Top-Level Ontology. New version of paper presented at AAAI Spring Symposium on Ontological Engineering, Stanford University, March 1997.

Hovy, E.H. and S. Nirenburg. (1992). Approximating an Interlingua in a Principled Way. Proceedings of the DARPA Speech and Natural Language Workshop. Arden House, NY.

Hovy, E.H. (1996). Semi-Automated Alignment of Top Regions of SENSUS and CYC. Presented to ANSI Ad Hoc Committee on Ontology Standardization. Stanford University, Palo Alto, September 1996.

Hovy, E.H. (1997). C*ST*RD: Information Clustering, Summarization, Translation, Reformatting, and Display. Unpublished draft, USC/Information Sciences Institute, Marina del Rey, CA.

Hovy, E.H. and C.Y. Lin. (1998). Automated Text Summarization in SUMMARIST. In M. Maybury and I. Mani (eds), Advances in Automatic Text Summarization. 1998. MIT Press, to appear.

Lehmann, F. (1997). Reworked Alignment of Top Regions of CYC and SENSUS. Presented to ANSI Ad Hoc Committee on Ontology Standardization. Stanford University, Palo Alto, March 1997.

Lenat, D.B. and R.V. Guha. (1990). Building Large Knowledge-Based Systems. Reading: Addison-Wesley.

Lin, C-Y. (1997). The ISI English Demorpher. Unpublished ms, USC/Information Sciences Institute, Marina del Rey, CA.

Knight, K. and S.K. Luk. (1994). Building a Large-Scale Knowledge Base for Machine Translation. In Proceedings of the AAAI Conference.

Knight, K., I. Chander, M. Haines, V. Hatzivassiloglou, E.H. Hovy, M. Iida, S.K. Luk, R.A. Whitney, and K. Yamada. (1995). Filling Knowledge Gaps in a Broad-Coverage MT System. In Proceedings of the 14th IJCAI Conference. Montreal, Canada.

Knight, K. (1997). SENSUS and SENSOR. Unpublished documentation, USC/Information Sciences Institute, Marina del Rey, CA.

Miller, G.A. (1990). WordNet: An Online Lexical database. International Journal of Lexicography 3(4) (special issue).

Mahesh, K. (1996). Ontology Development for Machine Translation: Ideology and Methodology. New Mexico State University CRL report MCCS-96-292.

Okumura, A. and E.H. Hovy. (1994). Ontology Concept Association using a Bilingual Dictionary. Proceedings of the 1st AMTA Conference. Columbia, MD.

Patil, R. (1996). The Ontosaurus Ontology Browser. Unpublished ms, USC/Information Sciences Institute, Marina del Rey, CA.

Swartout, W.R., P. Patil, K. Knight, and T. Russ. (1996). Toward Distributed Use of Large-Scale Ontologies. In Proceedings of the 10th Knowledge Acquisition for Knowledge-Based Systems Workshop. Banff, Canada.

Pustejovsky, J. (1996).

Rigau, G. and E. Agirre. (1995). Disambiguating Bilingual Nominal Entries against WordNet. Proceedings of the 7th ESSLI Symposium, Barcelona, Spain.