Measuring and Improving the Quality of World Knowledge Extracted From WordNet

Aaron N. Kaplan and Lenhart K. Schubert
University of Rochester

The University of Rochester
Computer Science Department
Rochester, New York 14627

Technical Report 751

May 2001

Abstract

WordNet is a lexical database that, among other things, arranges English nouns into a hierarchy ranked by specificity, providing links between a more general word and words that are specializations of it. For example, the word "mammal" is linked (transitively via some intervening words) to "dog" and to "cat." This hierarchy bears some resemblance to the hierarchies of types (or properties, or predicates) often used in artificial intelligence systems. However, WordNet was not designed for such uses, and is organized in a way that makes it far from ideal for them. This report describes our attempts to arrive at a quantitative measure of the quality of the information that can be extracted from WordNet by interpreting it as a formal taxonomy, and to design automatic techniques for improving the quality by filtering out dubious assertions.

This work was funded by NSF research grant IIS-0082928. The authors would like to thank David Ahn, Greg Carlson, and Henry Kyburg for regular discussions during the course of this work, and David Ahn for help labeling data.

WordNet [4] was designed as a dictionary and thesaurus for human use. It differs from classical dictionaries and thesauri in that rather than being simply an alphabetized list of self-contained entries, it is richly cross-referenced with relationships that psycholinguistic research indicates are involved in the organization of the human mental lexicon. Of these relationships, the one with the most instances in WordNet is the hyponymy relationship, which links a more specific word (hyponym) like "cat" to a more general one (hypernym) like "mammal."

For a variety of reasons, researchers in artificial intelligence and natural language processing have found taxonomic concept hierarchies to be useful components of their computational systems. The various taxonomies used in AI, while they differ in content, in general have a standard form: they are trees or lattices, in which the nodes represent predicates or properties, and the links indicate subsumption relationships. It is tempting to interpret WordNet's noun hierarchy in this way, since it is a lattice structure whose nodes are sets of synonymous nouns (nouns in natural language are often taken to express predicates or properties), and whose links indicate a relationship expressed in English as "is a" or "is a kind of," which seems similar in flavor to the subsumption relationship. Because it has this familiar-looking structure, because it has such broad coverage (some 94,000 nouns expressing some 66,000 unique concepts), and because it is hand-constructed and therefore presumably of high quality compared to taxonomies collected by automatic clustering techniques (*e.g.* [9, 3]), it is tempting to try to use WordNet as a source of taxonomic knowledge for a logical reasoning system.

The creators of WordNet did not have in mind a precise interpretation of the kind typically used in the knowledge representation field, and consequently interpreting it in such a precise way sometimes yields incorrect information. In Section 2, we will list a number of factors that make a formal interpretation of WordNet problematic, but in order to introduce the approach to be taken in this paper, let us consider one example. In WordNet, the words 'gold' and 'noble metal' are linked by the same relationship that links 'noble metal' with 'metallic element,' namely the hyponymy relationship. According to the consensus in semantics, in the sentence "Gold is a noble metal," the word 'gold' names an individual, the phrase 'noble metal' names a set (or a property), and the sentence is an assertion that the individual is a member of the set (or that the individual instantiates the property). In the sentence "A noble metal is a metallic element," 'noble metal' and 'metallic element' both name sets (or properties), and the sentence is an assertion that the first set is a subset of the second (or that instances of the first property are also instances of the second). A knowledge representation useful for logical inference must differentiate between the subset and membership relations, because they have different entailments.

Nevertheless, since WordNet contains such a wealth of information, it may be useful to use it *as if* it conformed to some more precise interpretation, as long as the performance of the system using the information degrades gracefully when given occasional incorrect information. A primary goal of the work described in this paper was to measure the quality of the information that can be obtained from WordNet in this way. The intention was to define a precise interpretation of the sort usually used in computational knowledge representations, and to measure by statistical sampling the proportion of assertions in WordNet that are false under that strict interpretation. This proportion could then be used as the degree of confidence that a system places in information extracted from WordNet. As a secondary goal, we hoped to find ways of improving this proportion by automatically identifying assertions that are likely to be false under the imposed interpretation. Work towards the first goal could feed into the second: in the process of sampling and evaluating the truth of assertions from WordNet, we would begin to understand the ways in which WordNet tends to deviate from the interpretation we had imposed, and we would also be creating a corpus of labeled instances that might be useful as training data for a classifier that could automatically identify assertions likely to be false under the imposed interpretation.

The first step in this proposed course of research, therefore, would be to fix a precise interpretation that could be imposed on WordNet. We originally assumed that this would be trivial, but it turned out to be one of the primary problems, and one which we still have not solved.

Imposing a precise interpretation on WordNet includes two subproblems. The first, which is easily accomplished, is to define a semantics for the representation. This means specifying what sort of object a node in the hierarchy represents, and defining what it means for two nodes to be connected by a (directed) link. Our solution is a rather standard one: we stipulate that each node represents a predicate, and that a link means that the predicate P represented by the upper node subsumes the predicate Q represented by the lower node, *i.e.* that $\forall x Q(x) \rightarrow P(x)$. The second part of defining an interpretation for the hierarchy is to identify the particular set that each node represents. For some nodes, this is more difficult than we anticipated, and whenever we are unable to identify the set represented by a node, we are accordingly unable to evaluate the truth or falsity of a link connecting that node to another. For example, while it is clear that 'metallic element' and 'noble metal' name properties such that the first subsumes the second, and that 'noble metal' and 'gold' do not name properties that stand in this relation, it is less clear whether this relationship holds between 'feeling' and 'calmness,' or between 'abstraction' and 'attribute.' We have gradually built a set of rules for identifying the denotation of a WordNet node, but at present we still frequently come across nodes whose denotation is not sufficiently constrained by those rules. Consequently, when sampling WordNet assertions to measure the proportion that are true under the interpretation we have imposed, we frequently have difficulty deciding whether a given assertion is true or not. This difficulty is reflected in low inter-annotator agreement scores.

Therefore, we have made only limited progress towards the primary goal—we can make only a rather rough estimate of the proportion of assertions that are true. Nevertheless, it still seems that some progress towards the second goal may be possible. While there are many assertions whose truth or falsity is difficult to determine, there are also many that are clearly true and many that are clearly false. We have, as hoped, identified some regularities among the false assertions, and have begun to demonstrate that some instances of systematic problems can be picked out automatically.

So on the one hand, we have arrived at an interpretation of WordNet under which many of the assertions are clearly true, and some are clearly false, and we believe we can make progress in automatically detecting the false ones; but on the other hand, so many of the assertions are not clearly true or false that automatic detection of clearly false assertions might not make a noticeable difference in the overall quality of the knowledge source. It seems that a change of focus might be in order.

This paper will proceed as follows. In Section 1, we will place this work in perspective by describing a sample of related work in two areas: the design of computational systems in which WordNet is used as a knowledge source, and proposals for redesign or modification of WordNet to make it more useful for such purposes. In Section 2, we characterize the ways in which we feel WordNet differs from an ideal taxonomy for a symbolic inference system. In Section 3, we will discuss our experiments in fixing a precise interpretation for WordNet and evaluating the quality of the information extracted thereby. In Section 4, we cover our preliminary work in detecting instances of systematic errors. Finally, in Section 5 we draw some conclusions from the observations we have made, and indicate where this work may lead.

1 Related Work

Despite its failings, WordNet is used very widely. To borrow a phrase from Churchill, it is the worst ontology devised by the wit of man, except for all the others. Whatever faults it may have, its coverage exceeds that of most similar resources by two orders of magnitude, and it is freely available without licensing restrictions. Consequently, it has been used in many computational applications. A bibliography of WordNet-related work on the web[1] currently lists some 200 papers, and it is undoubtedly far from complete.

WordNet is useful for statistical language processing because it allows abstraction. When statistics about individual words are unreliable because of the sparseness of data in the training corpus, these statistics can be aggregated by abstracting up the taxonomic hierarchy, under the often-valid assumption that semantically similar words occur in similar contexts. Resnik [14] has done some foundational work in this area.

Such statistical techniques are generally used for tasks such as syntactic disambiguation. Harabagiu and Moldovan have described applications for WordNet, or more precisely for an enriched version of WordNet

[1] http://www.cis.upenn.edu/~josephr/wn-biblio.html

they have built by automatically creating additional links based on information found in nodes' glosses, that involve more high-level reasoning. For example, in [8], they describe marker-propagation inference methods for establishing discourse coherence by discovering implicit connections between succeeding sentences.

The success of such applications indicates that WordNet does contain useful taxonomic information. Nevertheless, many authors have discussed what they view as flaws in its organization. Pustejovsky's "generative lexicon" research program [12] is based on the recognition that each word can express a variety, perhaps an infinity, of different meanings in different contexts, and that it is impossible to enumerate all of these possible senses independently of context, as WordNet attempts to do. Pustejovsky argues that a lexical entry should provide, rather than a single sense of a word, the information necessary to derive the sense that a word will take on in any given context. In [13], he proposes some principles for structuring lexical hierarchies; no broad-coverage lexicon has yet been built according to these principles, but one has the sense that if it were, it would be more regular and consistent, and therefore more appropriate for use in an inference system, than WordNet.

The work of Guarino and his various co-authors [7, 5], like that of Pustejovsky in [13], is concerned with defining principles and methodology that could be used to construct an ideal (or at least a better) taxonomy. In [5], they examine the WordNet noun taxonomy in light of the principles they have developed, thereby producing a rather comprehensive analysis of the range of systematic ambiguity and imprecision, or at least deviation from the usual semantics of taxonomies, that WordNet contains. In [10], one of us has presented a critique of the axioms by which Guarino *et al.* define the meta-properties (properties of properties) involved in their proposed constraints on taxonomies. While there are a number of technical flaws in their mathematical development, the underlying intuitions may be sound, and many of the observations about WordNet made in [5] (a preprint of which has just been made available as we write this) coincide with our own and with those of Pustejovsky [13]. Our discussion in Section 2 of the semantics of WordNet's hyponymy relationship reiterates some of the main points of Gangemi *et al.*'s recent paper, but we add a quantitative analysis of the prevalence of the problems discussed, and we take issue with some parts of their analysis.

In [5], six top-level ontological categories are defined in terms of Guarino *et al.*'s primitive meta-properties. Then, various subtrees of the WordNet hierarchy rooted at nodes in the first, second, and third levels are attached by subsumption links to each of the six categories, resulting in a new taxonomy with the same coverage as the original WordNet, but with a top-level structure that conforms to Guarino *et al.*'s principles. A similar approach was taken in building the SENSUS ontology, which was created by attaching subtrees of WordNet to nodes in the PENMAN upper model [1], a relatively small taxonomy designed according to principles of systemic functional linguistics.[2] At first, we found rearranging the top levels of WordNet to be an attractive idea, because WordNet's high-level nodes tend to represent vague, abstract concepts whose precise meanings are hard to pin down from just the synonyms and brief glosses provided, while concepts lower in the hierarchy tend to be more concrete ones whose intended meaning is clear. Perhaps the hope is that since the concepts in the lower levels of the hierarchy are relatively concrete and easily understood, the builders of WordNet might have ended up with a structure that conforms to one's favorite formal ontological theory, even if they didn't have such a theory explicitly in mind, so that by merely rearranging the very highest levels of WordNet, one can create a taxonomy with WordNet's broad coverage that conforms to one's ontological theory. Unfortunately, having examined these modified WordNets, our conclusion is that the rearrangement was too coarse-grained to effect any significant improvement in the taxonomy. Instances of the various problems we will discuss in Section 2 still occur frequently. Gangemi *et al.* state that their top-level reorganization of WordNet is just the first step, the ultimate goal being to rearrange the entire WordNet according to their ontological principles. This is certainly a worthwhile endeavor, but one that will take almost as much effort as the construction of the original WordNet.

---

[2]The SENSUS ontology was developed by Kevin Knight, Eduard Hovy, and Richard Whitney. To our knowledge, there is no published paper describing this rearrangement of WordNet to merge it with the PENMAN upper model. For some time the ontology could be browsed on the web at http://mozart.isi.edu:8003/sensus/sensus_frame.html, but at last check that page was not working. [11] describes how WordNet was merged with the Longman Dictionary of Contemporary English and various other lexical resources, but the results of that work have never been publicly available because of licensing restrictions.

2 Problems with WordNet's Semantics

In order to explain why we find elements of WordNet's organization objectionable, let us sketch how we envision a taxonomic hierarchy being used in a language processing system, and then look at the ways in which WordNet is incompatible with such a use.

The sort of application we have in mind is one in which sentences of natural language are translated into sentences of a logic, and then an inference engine uses those logical sentences, plus a base of previously obtained commonsense world knowledge expressed in the same logic, to perform some task such as executing a user's request, answering a question, or discovering implicit relationships between entities in a text. Taxonomic information would be one important part of the commonsense knowledge base, in part because it makes possible more efficient storage of other sorts of information. For example, if the commonsense knowledge base contains the information that mammals are warm-blooded, and that dogs, cats *etc.* are mammals, then it can infer that dogs are warm-blooded, and that cats are warm-blooded, *etc.* as needed, without storing each of these facts explicitly.

Taxonomic information, in this context, consists of assertions of the form $\forall x P(x) \rightarrow Q(x)$. A set of such assertions can also be expressed as a directed graph, where the nodes correspond to predicates of the logic, and in which the universal implication $\forall x P(x) \rightarrow Q(x)$ is expressed as a link from P to Q. This graph is useful because it lends itself to efficient inference techniques, but it is merely another way of expressing information that can also be expressed in the logic.

The predicates of the logic correspond to open-class words (nouns, verbs, adjectives, adverbs) in the natural language. More specifically, there is a predicate that corresponds to each lexical entry, so that for a polysemous word like 'bank,' which can refer (at least) to financial institutions and to the sides of rivers, there would be (at least) two predicates, say `bank1` and `bank2`. Part of the system's task when translating from natural language to logical form is resolving lexical ambiguity by associating each word with a predicate. The system's vocabulary of predicates will inevitably be incomplete; we will not specify how the system handles such cases, but simply concentrate on what happens when it is able to translate each word token with a known predicate, and therefore associate each with a node in the taxonomy.

We first examine how the semantics of WordNet's hyponymy relationship compares to the subsumption relationship of our ideal taxonomy, and then consider some problems in identifying the denotation of the nodes themselves.

2.1 Semantics of Hyponymy

A link in the sort of hierarchy we have just described means that every member of one class is a member of the other. The intended meaning of a hyponymy link in WordNet is not so well-defined. The creators of WordNet state that x is a hyponym of y "if native speakers of English accept sentences constructed from such frames as *An x is a (kind of) y* [4]. While this principle sounds similar to the formal relationship we have described, we will now see that the hyponymy relation by which WordNet is structured differs in important ways from the subsumption relationship between predicates.

Subsumption vs. Instantiation

As Gangemi *et al.* have noted [5], one readily apparent way in which WordNet differs from a subsumption hierarchy is that some of WordNet's nodes represent individuals, rather than predicates or properties, as illustrated in the introduction with the example of 'gold' as used in "Gold is a noble metal." Using WordNet in the most straightforward way as a source of taxonomic knowledge would entail using its set of nodes as a vocabulary of predicates, but with nodes like 'gold' as used here, this would be incorrect. The English word 'gold' *can* express a predicate, as in the sentence "I have some gold in my hand," which, ignoring a number of subtleties irrelevant to the current discussion, could be translated into logical form as

$\exists x[\mathtt{gold1}(x) \wedge \mathtt{in}-\mathtt{my}-\mathtt{hand}(x)]$; but in “Gold is a noble metal,” it is a different sense of ‘gold’ being used, namely a sense that represents the element, gold, an individual of which the predicate ‘noble metal’ is being asserted: $\mathtt{noble}-\mathtt{metal}(\mathtt{gold2})$. (To stay within first-order logic and WordNet’s sense-enumerative approach, this would have to be represented by using different symbols `gold1` and `gold2`, as we have just done. If one is willing to go beyond first-order logic, one can use a reification operator, such as the operator `K` (for “kind”) in [16]. Then the predicative sense of the word would translate in logical form as a predicate constant `gold`, and the individual sense would translate as the term `K(gold)`.)

WordNet contains many proper nouns, which are taken to represent individuals in some semantic theories and predicates in others. If one takes them to represent individuals, then every occurrence of a proper name in WordNet is a case of the problem we are discussing. If one takes then to represent predicates, then they don’t violate our proposed semantics for the taxonomic hierarchy, but nevertheless they are a special kind of predicate that it would be useful to be able to differentiate from the others.

In a random sample of 200 hypernym/hyponym pairs, one of us judged that 37 of them (about 20%) were related by instantiation rather than subsumption. Objectively speaking, these are not necessarily errors, but merely indicate a difference between the semantics of WordNet and the semantics of a subsumption hierarchy. However, it is a long-understood and uncontroversial principle of knowledge representation that for the purposes of inference, the relationships of subsumption and instantiation need to be kept distinct ([2] is a classic paper on the subject). Therefore, for an inference system to make use of the information in WordNet, we need to augment each link with a label indicating whether it represents subsumption or instantiation. In Section 4 we discuss an effort to build a classifier that does this automatically.

Inappropriate Use of Multiple Inheritance

WordNet allows a word to have multiple hypernyms; for example, ‘eaglet’ has the two hypernyms ‘eagle’ and ‘young bird.’ This example is compatible with the semantics of a subsumption hierarchy: every eaglet is both an eagle and a young bird. However, as Guarino [6] and Gangemi *et al.* [5] have pointed out, there are many cases in WordNet where multiple inheritance is used to indicate something other than the conjunction of two properties. For example, ‘fibrous tissue’ is a hyponym of both ‘animal tissue’ and ‘plant tissue,’ and ‘hoodoo’ (“a practitioner of voodoo”) is a hyponym of both ‘priest’ and ‘voodoo’ (the latter being a hyponym of ‘religion’).

In a random sample of fifty nodes with multiple hypernyms, one of us judged that seven of them (about 15%) were attempts to express something other than subsumption by both hypernyms.

2.2 Determining the Denotation of Individual Nodes

Knowing that WordNet’s hyponymy relation was not as well defined as the predicate subsumption relationship, our initial intention in this project was to sample hypernym/hyponym pairs to determine what fraction of such links would be false if interpreted as subsumption relationships, and to develop heuristics for identifying such links. However, such a judgment can only be made if one can first identify the intended meaning of the hypernym and hyponym nodes, and we found that to be difficult more frequently than we expected. In fact, the number of links whose truth we couldn’t confidently determine is at least as large as the number of links we judged clearly incorrect.

There are three sources of information that contribute to determining the meaning of a node. First, there are the words themselves. Each node consists of a “synset,” a set of one or more words or multi-word compounds which are asserted to be synonymous, or more precisely, to have a sense in common. Second, each node has a brief gloss that can further constrain the meaning, or can indicate the difference between various senses of the same word. Finally, there is the node’s position in WordNet as a whole—its direct and transitive connections to other nodes. This third source is clearly important, but for our purposes we can only use it in a limited way, since our goal is to measure the reliability of those very connections. We will now describe some of the situations in which these three sources of information fail to determine a unique denotation with enough precision to make subsumption judgments.

Confounding of Senses

Even though WordNet does incorporate the understanding that a single word may have many senses, and even makes rather fine distinctions in some cases, there are many cases where it seems that two or more intuitively distinct senses have been confounded in a single node. This makes it difficult to determine whether a given hyponymy link is true subsumption or not, since it may be true given one possible meaning of a node but false given another.

For example, take the node for sense 1 of the word 'hair.' This node is a hyponym of 'body covering' and a hypernym of the nodes 'hairball,' 'mane,' 'mustache,' and 'eyebrow,' among others, and its gloss is "any of the cylindrical filaments characteristically growing from the epidermis of a mammal and covering the body or parts of it." The gloss would seem to indicate the countable sense of the word 'hair,' as in "I found a hair on my plate," but the hypernym and hyponym links suggest the mass sense of the word, as in "his hair was matted." Single hairs, which are referred to by the first sense, have entirely different properties than collections of many hairs, which are referred to by the second sense, so a concept in a knowledge base must represent only one or the other. In this case, the bulk of the evidence points towards the mass sense, but since that evidence comes from the hypernymy and hyponymy links themselves, it is difficult to justify using it in preference to the information from the gloss when deciding whether or not the hyponymy links are valid. (There is also a separate question of whether a mustache, for example, is identical with a quantity of hair, or is a distinct individual that is merely composed of that quantity of hair. We will not address this question here; see [6].)

We showed earlier that there is evidence for taking the meaning of 'gold' in WordNet to be an individual rather than a predicate, because it has the hypernym 'noble metal.' But the same WordNet node happens to have the hyponym 'gold dust.' The relationship between 'gold' and 'gold dust' is that of a subsuming predicate to a subsumed one—every quantity of gold dust is a quantity of gold—so in fact there is conflicting evidence about the denotation of the node 'gold.' This ambiguity has immediate consequences: the hyponymy relationship is supposed to be transitive, but if it were, then it would mean that gold dust is a noble metal, and therefore an element, by two more transitive steps.

Ontological Obscurity

Whereas in the examples we have just seen, the meaning of a node is overconstrained, *i.e.* information about its meaning from various sources conflicts, in other cases the meaning is underconstrained, *i.e.* there is not enough information to identify the node's intended denotation. Consider the node 'abstraction,' which has the gloss "a general concept formed by extracting common features from specific examples." This node's hyponyms are 'time,' 'space,' 'attribute,' 'relation,' 'amount,' and 'set.' Taking any of the hyponyms on its own, one may be able to convince oneself that it is a predicate subsumed by 'abstraction,' although perhaps not with great confidence; considering them all together, the group seems so heterogeneous that it is difficult to accept that they are all siblings in a taxonomy, all subclasses of a single natural class. Gangemi *et al.* attribute the heterogeneity of siblings that is evident in WordNet to the contrast between what they call "types" and "roles," two different kinds of properties. We remain uncommitted for the moment about the validity of that analysis; regardless of whether it is correct, it seems that the problem is a lack of precision in defining the meaning of the node 'abstraction.' This sort of imprecision is particularly prevalent in the upper levels of the hierarchy, where rather abstract concepts have been invented to try to collect more self-evident lower-level concepts. It is difficult to identify the intended meaning of many high-level nodes from their words and glosses alone, and the heterogeneity of their hyponymy links only add to the confusion. An explicit description of an ontological theory on which the hierarchy's design was based would have been a useful disambiguating tool.

As we will see in the next section, difficulty in identifying the intended denotation of nodes turned out to be such a common problem that we were unable to reach confident quantitative estimates of the proportion of hyponymy links that represent true subsumption.

3 Measuring the Quality of Extracted Information

We evaluated some random samples to estimate the quality of the knowledge in WordNet when interpreted according to a formal semantics based on predicate subsumption.

In the first experiment, one of us was presented a list of 200 hyponym/hypernym pairs, generated as follows: for each pair, the hyponym was chosen from all of WordNet with a uniform distribution, and then one of that node's hypernyms was chosen, with a uniform distribution over the set of hypernyms. For each of the two nodes, the annotator recorded whether or not he could identify a unique predicate that the node expressed. He also recorded whether the predicate expressed by the hyponym was subsumed by the predicate expressed by the hypernym. If he felt that one or both nodes were ambiguous, but there was a choice of interpretations under which the hyponymy link expressed subsumption, he answered yes to this question. In 130 of the cases (65%), the answer to all three questions was 'yes.' Of the 400 nodes, the answer about whether the annotator could identify a unique predicate interpretation was 'yes' in 346 cases (85%).

Two problems became apparent in this first sample. First, the distribution of words was clearly not that of ordinary text. Medical terms and plant and animal names, including many Latin names of species, were dramatically prevalent, and there were a number of other very uncommon words in the sample. Second, while the annotator answered 'yes' or 'no' to each question, there were a significant number of them on which he was not confident of his answer, for reasons such as those we discussed in the previous section.

Of course, since we chose the words with a uniform distribution over all of WordNet, their distribution reflected the distribution of WordNet in general. Since it has very broad coverage, WordNet necessarily contains many infrequent words—in fact, *most* of its words are infrequent ones—and consequently our sample consisted mostly of infrequent words.

Since we wanted a measurement that would be likely to be a good predictor of the usefulness of the knowledge for language processing tasks, and since we were concerned that the quality of the information involving common words might differ from the average quality over all words, in the next experiment we chose the sample of node pairs with a distribution biased by frequency, using the frequency counts provided with WordNet. These counts are from a relatively small sub-corpus of fewer than 400,000 words, selected from the Brown corpus. Half of the words in WordNet did not occur at all in the corpus; to make the chance of those words occurring in the sample non-zero, we raised their occurrence count to 0.1 for the purposes of calculating sampling probabilities.

We also felt that it was important to have a number of concepts from the upper levels of the hierarchy in the sample, since they are likely to be used frequently in inference applications. Many of the higher-level concepts are artificial multi-word concepts that were constructed as abstractions of lower-level concepts, and therefore do not occur in the corpus, and would have a low probability of occurring in the sample. We considered using an adjusted frequency count equal to the sum of a node's actual frequency count plus the adjusted counts of its children, but this made for too much of a bias towards the top levels. We settled on a scheme in which a node's adjusted frequency count is its actual count plus the greatest adjusted count of its hyponyms.

To address the problem of the annotator's lack of confidence in some of the judgments, we devised a more formal set of guidelines for the process. The instructions are summarized here:

- A proper noun expresses a predicate that is true of only one thing.
- If a common noun expresses a kind (an individual), then it also expresses the predicate which is true of individuals of that kind, and that predicate is the one to consider. For example, in "gold is an element," 'gold' expresses a kind, but in "the gold in my fillings picks up radio waves" it expresses a predicate.

- Deverbal nouns such as ‘running,’ ‘translation,’ and ‘governance’ express predicates over actions/events/episodes.
- The following schemata indicate a subsumption relationship:
  - If x is a P, then x is a Q.
  - If x is a quantity of P, then x is a quantity of Q.
  - If x is a P event, then x is a Q event.
- These quantifiers may also be mixed, *e.g.* an affirmative answer to

  If x is a quantity of P, then x is a Q

  is evidence that Q subsumes P.
- The schema “P is a Q” is *not* evidence that Q subsumes P. For example, “gold is an element” is not an assertion that if x is a quantity of gold, then x is an element; rather, it is an assertion that the kind gold is an element.
- Glosses are to be taken only as secondary information. They often suggest things of a different ontological sort than the synonyms themselves; trust the synonyms, and only use the gloss to help disambiguate.

In the second experiment, we generated a new set of 100 hyponym/hypernym pairs according to the adjusted-frequency-biased distribution described above, and each of us annotated that set. One annotator answered ‘yes’ to all three questions 74 times, and the other 77 times. While this would seem to be reasonable agreement on the question of proportion, in fact our agreement on the question of which links were the false ones was rather poor. We had a kappa score of 0.38, with a 95% confidence interval of between 0.14 and 0.62.[3] This low agreement score coincided with our feeling that we we still lacked confidence in many of the judgments. The increased preciseness of the annotator instructions was offset by the fact that in the frequency-biased sample, nodes from higher levels in the hierarchy appeared more frequently, and as we have discussed, ambiguous and imprecisely-defined nodes are concentrated in the upper levels.

The frequency-biased sample contains a lower proportion of links that are clearly instantiation rather than subsumption. As reported in in Section 2, about 20% of the links in the uniformly distributed sample were of this type, but only between 1% and 3% of the frequency-biased sample were. We attribute this to the fact that nodes representing individuals, or more precisely nodes that can *only* be interpreted as individuals and not as predicates, are usually leaf nodes, and so were much less frequent in the second sample, which was biased towards higher-level nodes.

Since that experiment, we have continued refining the annotator rules, in an effort to come closer to defining an interpretation that assigns a clear truth value to every link, but at this point we feel that they don’t yet provide enough guidance to make another experiment worthwhile.

Annotator instructions necessarily limit the annotator’s attention to local information. In other words, although to identify the most appropriate interpretation of the network it might be necessary to consider all 66,000 nodes and their interrelationships, it isn’t reasonable to instruct the annotator to do this, so we have limited the information the annotator considers to the list of synsets and the gloss of a hyponym and a hypernym. Perhaps we have been overly restrictive; it may be that we could increase the confidence and consistency of judgments by presenting the annotator with a list of the hyponym’s siblings, to provide extra context.

---

[3]The kappa score compensates for the probability of agreeing by chance, which is rather high if both annotators usually answer ‘yes;’ a kappa score of zero indicates no more agreement than would be expected by chance, and a score of one indicates perfect agreement.

4 Detecting Systematic Errors

We have seen that many of the links in WordNet don't represent predicate subsumption relationships, and that there are a few systematic phenomena that account for many of these non-subsumption links. We have therefore considered the possibility of heuristics for automatically identifying non-subsumption links, so that WordNet can be used with higher confidence as a source of taxonomic knowledge. Preliminary work on automatically differentiating instantiation links from subsumption links looks particularly promising.

The observation that makes heuristic detection of instantiation links seem possible is that most of such links fit one of two patterns: either the hyponym is a proper noun (assuming that one takes proper nouns to represent individuals, which is contrary to the annotator instructions listed in Section 3, but probably desirable as argued in Section 2.1), or it is a mass noun and the hypernym is a count noun, for example 'gold' as a hyponym of 'rare metal.' This pattern often indicates that the hyponym names a kind and the hypernym names a predicate over kinds.

Proper nouns can be detected with high recall by simply looking for capitalization, but this feature alone results in low precision, because WordNet contains many capitalized words which are derived from proper nouns, but are not themselves proper nouns, for example 'Asian' (a person from Asia), 'Americana,' 'Europeanization,' 'Dewar flask.' It should be possible to screen out many of these derived common nouns by morphological analysis, but we have not yet experimented to quantify how much the precision can be improved.

There are a number of cues that can provide information about whether a node represents a count or a mass concept. One is morphology: certain suffixes almost always indicate a mass noun, *e.g.* -ness, -tion, and -ing. Two other useful resources are the CELEX[4] and Alvey [15] lexical databases, which list for each noun whether the word has a mass sense, and whether it has a count sense. The Alvey lexicon has entries for 23,492 nouns, and CELEX 29,264 nouns, with a total of 31,115 unique nouns in their union. This combined Alvey/CELEX database only overlaps WordNet by 23,748 nouns, which means that it covers 25% of the nouns in WordNet; however, since many WordNet nodes contain multiple synonyms, these 23,748 words are enough to cover at least one word in 46% of the nodes. Furthermore, Alvey and CELEX were built to contain the most common English words, so in a frequency-biased sample coverage would be much higher.

The CELEX database can also help with morphological analysis: it indicates the morphological structure of polymorphemic words, which can help eliminate spurious analyses by which, for example, 'thing' would be labeled a mass noun because it ends in -ing. For nouns that don't occur in CELEX, such spurious analyses could be reduced by looking for the postulated root in WordNet.

A problem with all of these sources of information is that they provide information about words, not about WordNet nodes. Many words have both mass and count senses listed in WordNet, and word-based heuristics can't distinguish which is which. As part of the SENSUS project, a mapping was created semi-automatically between the Longman Dictionary of Contemporary English (LDOCE), which indicates for each defined word sense whether it is mass or count, and WordNet [11]. This merged resource would be able to indicate, for each WordNet node that was successfully mapped to an LDOCE word sense, whether it was mass or count. Unfortunately, because of LDOCE's restrictive licensing agreement, the SENSUS group is unable to make this resource available to us.[5]

5 Conclusions and Future Work

We set out to measure the quality of the information in WordNet when viewed as a source of formally interpretable taxonomic knowledge. From a uniformly distributed sample, we estimated that about 35% of

[4] Available from the Linguistic Data Consortium

[5] Richard Whitney, personal communication

WordNet's hyponymy links do not represent a subsumption relationship between two predicates. More than half of those represent an instantiation relationship between a predicate and an individual; there is reason to believe that these instantiation links can be rather reliably detected by heuristics that make use of morphology and mass/count information from other lexical databases.

Our confidence in these numbers is limited by two factors. First, when evaluating the sample to make this estimate, the annotator felt that the questions he was attempting to answer were often ill-defined, and second, the sample was dominated by very infrequent words. On a second sample, we used a biasing scheme by which the probability of a word sense appearing in the sample was proportional to its frequency in a corpus, with some smoothing so that all words had non-zero probability. The evaluation of the second sample demonstrated two things. First, it showed that the proportion of non-subsumption links, and of instantiation links in particular, was affected by the move to a frequency-biased sample. Second, in this sample we used two independent annotators, and inter-annotator agreement was poor, indicating that the questions being asked were still often ill-defined, despite work undertaken following the first experiment to make them more precise.

The kind of subsumption information we have been looking at so far is useful for giving positive answers to type inference questions (*e.g.* "if x is a dog, is x a mammal?") and for inheritance of information down the taxonomic hierarchy ("if x is a dog, is x warm-blooded?"). Taxonomies also often include information about exclusivity of their classes, *e.g.* that the classes 'physical object' and 'abstraction' are disjoint, which can be used to give negative answers to questions like "if x is a dog, and y is an idea, is it possible that $x = y$?" WordNet doesn't include exclusivity information, but naturally important exclusivity relationships exist between some of its nodes (in a sample of 250 randomly-chosen pairs of siblings, we judged 175 (70%) to be mutually exclusive), so we intend to look into heuristics for identifying exclusive pairs of nodes. In particular, since sibling natural kinds tend to be mutually exclusive (*e.g.* bird *vs.* mammal), while roles tend not to be (*e.g.* musician *vs.* father), it would be useful to be able to distinguish these two kinds of predicates. Morphology is one cue that could be useful towards that end. Differentiating natural kinds from roles, would also be of use in identifying links that are problematic according to Guarino *et al.*'s framework, since natural kinds and roles, as we describe them here, seem similar to what they call types and roles, respectively.

There is more to WordNet than the network of nouns induced by the hyponymy relationship. It also includes other relationships between nouns, *e.g.* the part-of relationship, as well as other parts of speech. Similar techniques to those we have discussed here could be used for evaluating and eventually extracting and using those other kinds of information.

An important question that is as yet unanswered is how well our frequency-biased sample approximates the distribution of information that a real language processing application would use. It seems clear that the uniform distribution of our first sample is not the right thing to measure, since almost all of the words in the sample were very infrequent ones; but our method of distributing some of the frequency of hyponyms to their hypernyms is admittedly somewhat *ad hoc*.

Our evaluation was made without a particular application for the information in mind. A more task-based evaluation might also be interesting. Given an application that uses taxonomic knowledge and a corpus of problem instances, we could (a) collect statistics on which items of knowledge extracted from WordNet were used in each instance, giving a more defensible basis for a distribution to be used in statistical sampling, and (b) given information about the correct answers to the problem instances, use system performance as a metric for the improvement gained by heuristic modifications of WordNet.

Besides providing a more well-founded basis for statistical analysis, a task-based evaluation would eliminate the problems that stem from our inability to agree reliably on the intended denotation of some nodes. Evaluation would be by the pragmatic criterion of whether a piece of information leads to the right answer, rather than a partly subjective evaluation of whether it is true under some interpretation. We are ambivalent about this prospect. It is true that the task-based approach to evaluation would yield data that are more reliable and more obviously relevant than the data we have been able to obtain so far. On the other hand, the approach would mean giving up on one of the main attractions of the symbolic approach to AI, namely the idea that a system's internal representations can be interpretable by a human. If WordNet contains information that is useful for some language processing task, then ultimately it should be possible to state what that information is, by defining a formal interpretation of WordNet. This is what we have been attempting to do, and we still hope to make further progress in this direction.

References

[1] J. Bateman. Upper modeling: Organizing knowledge for natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Generation*, Pittsburgh, PA, 1990.

[2] Ronald J. Brachman. What is-a is and isn't: An analysis of taxonomic links in semantic networks. *Computer*, 16(10):30–36, October 1983.

[3] Sharon A. Caraballo. Automatic construction of a hypernym-labeled noun hierarchy from text. In *37th Annual Meeting of the Association for Computational Linguistics: Proceedings of the Conference*, pages 120–126, 1999.

[4] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[5] Aldo Gangemi, Nicola Guarino, Alessandro Oltramari, Domenico Pisanelli, and Geri Steve. Conceptual analysis of lexical taxonomies: The case of wordnet top-level. Internal Report 06/2001, LADSEB-CNR, Padova, Italy, April 2001.

[6] Nicola Guarino and Christopher Welty. The role of identity conditions in ontology design. In C. Freksa and D. M. Frank, editors, *Spatial Information Theory: Cognitive and Computational Foundations of Geographic Information Science*. Springer Verlag, 1999. Amended version at http://www.ladseb.pd.cnr.it/infor/Ontology/Papers/IJCAI99.pdf.

[7] Nicola Guarino and Christopher Welty. Ontological analysis of taxonomic relationships. In A. Laender and V. Storey, editors, *Proceedings of ER-2000: The 19th International Conference on Conceptual Modeling*. Springer-Verlag, October 2000.

[8] Sanda Harabagiu and Dan Moldovan. Knowledge processing on extended wordnet. pages 379–405.

[9] Marti A. Hearst. Automated discovery of wordnet relations. In Fellbaum [4], chapter 5, pages 131–152.

[10] Aaron N. Kaplan. Towards a consistent logical framework for ontological analysis. Technical Report 748, University of Rochester, Rochester, NY, May 2001.

[11] K. Knight and S. K. Luk. Building a large-scale knowledge base for machine translation. In *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94)*, 1994.

[12] James Pustejovsky. *The Generative Lexicon*. MIT Press, Cambridge, MA, 1995.

[13] James Pustejovsky. Type construction and the logic of concepts. In P. Bouillon and F. Busa, editors, *The Syntax of Word Meaning*. Cambridge University Press, 2001.

[14] Philip S. Resnik. *Selection and Information: A Class-Based Approach to Lexical Relationships*. PhD thesis, University of Pennsylvania, 1993.

[15] G. J. Russell, S. G. Pulman, G. D. Ritchie, and A. W. Black. A dictionary and morphological analyser for english. In *Proceedings of 11th International Conference on Computational Linguistics*, pages 277–279, Bonn, 1986.

[16] Lenhart K. Schubert and Chung Hee Hwang. Episodic logic meets little red riding hood: A comprehensive, natural representation for language understanding. In L. Iwanska and S. C. Shapiro, editors, *Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language*. MIT/AAAI Press, 2000.