TOWARDS BUILDING A WORDNET NOUN ONTOLOGY

GEORGE A. MILLER, FLORENTINA T. HRISTEA

Abstract. WordNet, a lexical database for English that is extensively used by computational linguists, has not previously distinguished hyponyms that are classes from hyponyms that are instances. This work describes an attempt to draw this distinction and reports the way in which the results were incorporated in the last version (2.1) of WordNet.

1. INTRODUCTION

If you were to say, "Women are numerous", you would not wish to imply that any particular woman is numerous. Instead, you would probably mean something like "The class of women contains numerous members". To say, on the other hand, "Rosa Parks is numerous", would be nonsense. As Quirk *et al.* (1985: 288) point out, proper nouns normally lack number contrast.

This important distinction underlies the present discussion of WordNet (WN) nouns. Some nouns are understood to refer to classes; membership in those classes determines the semantic relation of hyponymy that is basic for the organization of nouns in WN. Other nouns, however, are understood to refer to particular individuals. "Rosa Parks", for example, ordinarily refers to a particular individual.

The distinction to be discussed here is between words ordinarily understood as referring to classes and words ordinarily understood as referring to particular individuals and places. This distinction was not drawn in initial versions of WN, which used the "is a" relation in both cases. That is to say, both "A heroine is a woman" and "Hillary Clinton is a woman" were considered to be occurrences of the "is a" relation and were encoded in the WN database in the same manner.

2. WORDNET

WN (Miller 1990; Fellbaum 1998) is a lexical database that currently contains approximately 147,000 English nouns, verbs, adjectives, and adverbs organized by semantic relations into 117,500 meanings, where a meaning is represented by a set of synonyms (a synset) that can be used (in an appropriate

RRL, **LI**, *3–4*, p. 405–413, Bucureşti, 2006

context) to express that meaning. An entry in WN consists of a synset, a definitional gloss, and (sometimes) one or more phrases illustrating usage. The semantic relations used to organize words and entries are synonymy and antonymy, hyponymy, troponymy and hypernymy, meronymy and holonymy.

WN can be equally viewed as a lexical database, as a semantic network or as a knowledge base. It has been recognized as a valuable resource in the human language technology and knowledge processing communities. Many researchers who use WordNet view it primarily as a lexical knowledge base and make subsequent use of it. Its applicability has been cited in more than 300 papers and systems have been implemented using it. Many groups of researchers expressed their interest in WordNet applications in various fields, such as: Information Retrieval, Information Extraction, Word Sense Disambiguation, Text Inference, Natural Language Generation, Learning, Knowledge Acquisition and others.

Requests to incorporate the distinction between classes and particular instances into WN have come from ontologists, among others. In their discussion of WN, for example, Gangemi *et al.* (2001) and Oltramari *et al.* (2002) complain about the confusion between concepts and individuals. They even suggest that if there was an "instance of" relation, they could distinguish between a concept-to-concept relation of subsumption and an individual-to-concept relation of instantiation. This is, essentially, the suggestion we try to follow in the present work.

Incorporating this distinction was resisted for many years because WN was not originally conceived as an ontology but rather as a description of lexical knowledge. It includes verbs, adjectives, and adverbs in addition to nouns. Early in its development the nouns in WN were divided into 25 categories corresponding to general topics: act, animal, artifact, attribute, body, cognition, communication, event, feeling, food, group, location, motive, object, person, phenomenon, plant, possession, process, quantity, relation, shape, state, substance, and time, with a file for each topic. And a few very generic nouns were used in a Tops file that related the several topical files.

Although no ontology was intended, the organization of nouns in WN bore many similarities to an ontology. As the importance of ontology became more apparent, requests to convert the WN noun hierarchy into an ontology could no longer be ignored. Version 2.1 of WN takes a step in that direction: the noun.Tops file is reorganized so as to have a single unique beginner: *entity*. In a reasonable ontology, however, all terms might be expected to conform to the membership relation of set theory, and would not contain particular individuals and placenames. The fact that classes and instances had been confounded in WN posed a problem; the obvious way to solve that problem was to distinguish between them.

Another reason to distinguish classes from particular instances arose when Beth Sundheim (personal communication) proposed that placenames found in WN might be linked to a gazetteer. WN contains many geographical terms for

individual placenames (Boston, Germany, Africa, etc.) that are also described in gazetteers (along with longitude, latitude, population, etc.). WN also contains some geographical information (New England, the Balkans, the Confederacy, etc.) that gazetteers do not include, so linking the two would strengthen both. But first the individual instances - the placenames - need to be identified.

3. IDENTIFICATION OF INSTANCES

There are three characteristics that all words denoting instances share. (1) They are, first of all, nouns. WN contains some 147,000 unique words (both simple and compound), of which approximately 117,000 can be used as nouns. (2) Moreover, nouns denoting instances are proper nouns, which means that they should be capitalized. WN contains some 40,000 capitalized nouns which are contained in approximately 24,000 synsets. (3) Finally, the referent should be a unique entity, which implies that they should not have (or should rarely have) hyponyms.

Unfortunately, these three characteristics are shared by many words that are not particular instances. In clear-cut cases, such as persons or cities, there is little problem identifying instances. For example, every entry that has *city* as a hypernym is an instance of a city in WN; every entry in the person file between *Alvar Aalto* and *Vladimir Kosma Zworykin*, whether an architect or zoologist, names a particular individual. Such instances, whether cities or persons, can be easily identified. In addition to the biographical section of the person file and all the entries with *city* as a hypernym, anything with *river*, *range*, *peak*, or *terrorist organization* as a hypernym can be identified as an instance. Almost anything with *lake* as a hypernym is also an instance, (e.g., Lake Erie) except for words like *bayou*, *lagoon*, *loch*, *lough*, *pond*, *oxbow lake*, *pool*, and *tarn* that denote classes of lakes, not particular instances.

Those are the easy cases. There are many other proper nouns without hyponyms, however, that are not instances. There seemed to be no alternative to inspecting all the synsets that contained candidate nouns, one at a time, in order to identify all the instances. This was performed by two persons.

The manual tagging of instances was done in the form of an experiment for which an interface was prepared that would present the capitalized nouns in WN, along with their definitional glosses, one synset at a time, and would keep a record of the decisions that were made regarding the nouns' classification. Two manual taggers, FH and GM, went through some 24,073 items and labeled them as classes or instances.

The results of the experiment can be summarized in a fourfold table, where the diagonal values represent agreements (21,302) and off-diagonal values (2,771) represent disagreements:

		FH		
		Classes	Instances	Totals
	Classes	14,167	2,673	16,840
GM	Instances	98	7,135	7,233
	Totals	14,265	9,808	24,073

The relatively large number of disagreements was taken to indicate that the taggers were working with different conceptions of the task. Nevertheless, the coefficient kappa was calculated to be 0.75, with a very small variance (0.004), indicating substantial correspondence.

The strategy that GM tried to follow for assigning "instance" tags was to concentrate on a word's referent. When he knew of a unique referent, he considered it a clear case of an instance; when he was unsure that his criterion was met, his tendency was not to assign an "instance" tag; and when a class was clearly indicated, the "instance" tag was, of course, not assigned. For example, when *Beethoven* is used to refer to the German composer, it is an instance, but when *Beethoven* is used to refer to the composer's music (as in "She loved to listen to Beethoven"), the same word refers to a class of musical compositions. Moreover, just to be clear, when there were two unique referents, both were tagged as instances. For example, *Bethlehem* in the Holy Land and *Bethlehem* in Pennsylvania were considered to be unique referents and both were tagged as instances. And when an instance had two or more hypernyms, it was tagged as an instance of all of them. For example, *Mars* is an instance of a terrestrial planet (having a compact rocky surface) and also an instance of a superior planet (its orbit lies outside the Earth's orbit).

Whereas GM's strategy was conservative in recognizing instances, FH tried to follow a more even-handed strategy. That is to say, this tagger did not think of the task as one of recognizing instances but rather as one of deciding, on the basis of available evidence, which of two categories was most appropriate. Such evidence usually referred to uniqueness and specificity (of location, moment in time, author, etc.). According to such criteria, *Geneva Convention* was tagged as an instance of *convention* since it is unique and the gloss refers to a specific date. Similarly, *North Atlantic Treaty* was considered an instance of *treaty* since the gloss said it was signed in a particular year by 12 particular countries. Thus, FH found many instances that GM had been unsure having met his criterion and so rejected.

However, uniqueness is not the sole criterion that was taken into account. In computer science, for example, although each operating system is unique, both taggers considered DOS and UNIX to be a class of operating systems (with MS-DOS and Linux as instances). Similarly, LISP, Prolog, COBOL, C, and BASIC were taken to be classes of programming languages. The same agreement between taggers did not occur, however, in their treatment of natural languages, as will be discussed below.

One problem that bothered both taggers was the occasional occurrence of capitalized and lower-case words in the same set of synonyms. For example, one synset contained { North, northland, Septentrion }, another contained { diazepam, Valium }, etc. The occurrence of words beginning with lower-case letters seems to indicate a class, whereas the capitalized words left open the possibility of an instance. The problem, of course, is that the relation must be assigned to the whole synset according to WN conventions. It makes no sense for a word to refer to an instance and for its synonym to refer to a class.

4. CONFLICT RESOLUTION

In order to bring some system into the resolution of tagger differences, it was decided to look at these differences as a function of the 25 general topics that were used to organize WN nouns. Three files with regard to which substantial disagreement was observed were the communication, location, and object files, and in the case of the quantity file the disagreement was complete (Kappa = 0). While resolving the disagreements between taggers it was decided, therefore, to examine some files more closely than others.

This discussion illustrates the kind of disagreements that arose and indicates the ways they were resolved. As WN grows in future versions an effort will be made to maintain the distinction introduced here.

4.1. Double classification using noun category

In the MUC-7 task definition (Chinchor 1997), three types of named entities were to be identified: organizations, persons, and locations. Dates, times, money, and percentages were subtasks. With WN, there are potentially many more types of named entities available, at least as many as there are categories corresponding to noun classes. For example, *al-Qaeda* can be identified as an instance of an organization, *Marie Curie* can be identified as an instance of a person, and *Boston* can be identified as an instance of a location by simply combining the instance tags with the noun's category: group, person, and location, respectively.

4.2. Instances in the communication file

When WN was first developed, criteria for including a word in the communication file were lenient, involving spoken and written messages, the languages they were spoken or written in, the expressive styles of speaking or writing, different writing systems, sacred texts, treaties, legal documents and contracts, types of publications, and all the special terminology used to discuss these topics. In most cases the two taggers agreed, but some interesting disagreements emerged.

For example, the two taggers disagreed in their treatment of sacred texts. Whereas they agreed that *Adi Granth*, *Zend Vesta*, *Bhagavadgita*, *Mahabharata*, and others were particular instances of sacred texts, when they came to the Christian Bible they disagreed: GM called it a class term whereas FH felt it was a particular instance, no different from the other sacred texts. GM defended his choice by pointing out that WN contained many hyponyms of *Bible*: *Vulgate*, *Douay, King James, Revised Version, American Revised Version*, etc. But GM's decision seemed to make the Bible a special case, which may have resulted from WN's compilers knowing more about the Bible than about other sacred texts. It was decided that this was a case in which a sacred text could be a class: *Bible* was tagged as a class of a sacred text and its hyponyms were tagged as instances.

Most of the disagreements in the communication file, however, resulted from differences regarding natural languages. GM felt that no language is an instance, although the use of a language on some particular occasion might be considered an instance; for this tagger all languages were classes. FH, on the other hand, felt that specific languages should be particular instances. *Old Italian*, *Sardinian*, and *Tuscan* were therefore tagged as instances of *Italian*. Generally speaking, this tagger tended to distinguish among the official languages of a given country, a group or family or branch of languages, dialects, etc. Thus, most of the disagreements between the two taggers concerning the communication file resulted from their treatments of the many different languages. In the end, it was decided that, from an ontological point of view, languages are not instances, only the speech acts are so.

For the convenience of comparative linguists, however, it should be pointed out that the hyponyms of the WN entry for *natural language*, *tongue* give a classification of the major languages of the world.

4.3. Instances in the location file

The location file overlaps somewhat with the (natural) object file, but they will be considered separately here.

The location file contains a variety of words used to locate objects or events in space as well as the names of regions and countries and their political divisions and inhabited areas. With most of these entries the taggers agreed. There was one case where the value of having more than one tagger was clearly demonstrated, the case where GM considered nearly all of the regions included as hyponyms of *geographical area, geographic region* to be classes, not instances. These hyponyms included *Andalusia*, *Appalachia*, *Antarctic Zone*, *Badlands*, *Barbary Coast*, *Bithynia*, *Caucasia*, *Finger Lakes*, *Gulf States*, *New England*, *Nubia*, and many more. For various reasons, mostly historical, these regions may no longer have well-defined political boundaries but the terms still have geographical significance and are in general use. Although vague in denotation, they will be considered as instances in WN.

The location file also contains the signs of the zodiac: *Aries, Taurus*, *Gemini*, *Cancer*, etc., which will also be considered as instances.

4.4. Instances in the object file

The object file includes natural objects, not artifacts. It includes the names of islands and continents, rivers and lakes, mountain peaks and ranges, seas and oceans, planets and satellites, stars and constellations, electrons and mesons, etc. Here again the taggers agreed except for constellations, which GM called classes and FH identified as instances. In WN 2.1, as in future versions, constellations will be considered instances.

If WN is to be supplemented with links to a gazetteer, it is the instances in the location and object files together that will comprise all of WN's placenames.

4.5. Instances in the quantity file

The quantity file is relatively small. It contains words used in the various systems of weights and measures, the many monetary units of the world, and a sampling of digits, etc. The reason for considering it here is that the two taggers disagreed so completely.

FH considered a number of metric units (especially those named for the scientist they honored) to be instances; this tagger also considered many monetary units to be instances (e.g., the *Hong Kong dollar* as an instance of *dollar* or the *Cuban peso* and as instance of *peso*, etc.); and the numbers (6 for example) were considered instances of digits. Since GM found no instances in the quantity file, disagreement was complete. After reviewing the file, both taggers came to the conclusion that none of the words in the quantity file will be considered to denote instances. The multiple occurrence of cases like digits and monetary units led to their classification as types.

4.6. Instances in the artifact file

The artifact file is large (more than 11.400 entries) including such impressive instances as the seven wonders of the ancient world. The artifact file contains the names of man-made things including, in addition to ordinary names, the slang names, trade names, street names, etc., many of which were capitalized and so were offered as possible instances.

A few puzzles arose with the names of artifacts, the most frequent one resulting from the convention of combining the generic and trade names of medicinal drugs in the same synset, e.g., the bronchodilator {metaproterenol, Alupent} or the tranquilizer {chlordiazepoxide, Librium, Libritabs} and so on through a long list of drugs. After discussion, the taggers concluded that a chemical name like *acetylsalicylic acid* denotes a class of substances and that *Bayer aspirin* denotes the same class of substances, so the terms are synonymous. In short, giving something a trade name does not change it from being a class to being an instance.

This conclusion about trade names served to solve some other problems where only the trade name was given in WN: *Band-Aid*, *Catepillar*, *Dacron*, *Orlon*, *Ovulen*, *Tenoretic*, etc.; none will be called instances in WN. Nor will the street names of drugs, although often imaginative, be considered as instances.

5. CONCLUSIONS

Overall, there were 7,671 synsets in WN that the two taggers finally agreed should be tagged as instances.

The symbol used to code hypernyms has been ‘@.’ That is to say, { peach, drupe,@ } has represented “a peach is a drupe” or “all peaches are drupes“. This notation is appropriate for representing relations between classes but it is not appropriate for representing relations between instances and classes. That is to say, when {Berlin, city,@} is used to represent “Berlin is a city”, the particular instance “Berlin” is treated inappropriately as a class. A different symbol is needed to code instances. We have chosen, therefore, simply to add an ‘i’ to the ‘@’; to represent “Berlin is an instance of a city” by {Berlin, city,@i} in the new notation.

The release of WN 2.1 contains the distinctions between classes and instances described here, so that it is now possible to treat WN nouns as a semi-ontology by simply ignoring all entries tagged with ‘@i.’ Alternatively, by selecting only those entries tagged with ‘@i’ and contained in the location and natural object files, it will be possible to extract from WN all 3.062 placenames that can be related to a gazetteer.

We are convinced that the mentioned distinctions between classes and instances will be subject to helpful criticism by WN users, as are all the other lexical relations in WN. It is hoped that this modification, leading to a semi-ontology of WN nouns, will make WN even more useful to future users.

Acknowledgments. Florentina Hristea is grateful to the Romanian-U.S. Fulbright Commission for the Fulbright Grant that made it possible for her to collaborate in this research. Work by the Cognitive Science Laboratory was supported by a contract between Princeton University and the Advanced Research Development Activity (AQUAINT Program Phase 2, Contract No. NBCHC40012). The authors are indebted to Benjamin Haskell for developing the interface used to conduct the experiment and to Christiane Fellbaum, Helen Langone, and Randee Tengi for comments on the manuscript.

REFERENCES

Chinchor, N., 1997, *MUC-7 Named Entity Task Definition (version 3.5)*, unpublished manuscript.

Fellbaum, C**.** (ed.), 1998, *WordNet: An Electronic Lexical Database*, Cambridge, MA, MIT Press.

Gangemi, A., N. Guarino, A. Oltramari, 2001, *Conceptual Analysis of Lexical Taxonomies: The Case of WordNet Top-Level*, in: C. Welty, B. Smith (eds), *Proceedings of FOIS2001*, ACM Press, 285-296.

Miller, G. A**.** (ed.), 1990, *WordNet: An On-Line Lexical Database*, in the special issue of the *International Journal of Lexicography*, 3, 235-312.

Oltramari, A., A. Gangemi, N. Guarino, C. Masolo**,** 2002, *Restructuring WordNet's Top-Level: The OntoClean Approach*" in *Proceedings of LREC2002* (OntoLex workshop), Las Palmas, Spain.

Quirk, R., S. Greenbaum, G. Leech, J. Svartvik, 1985, *A Comprehensive Grammar of the English Language*, London / New York, Longman.