

Authors Doug Lenat, George Miller, and Toshio Yokoi critique and defend one another's systems, ideas, and approaches to AI development.

CYC, WordNet, and EDR: Critiques and Responses

Lenat on WordNet and EDR

I applaud Miller's WordNet project and feel that there is much in common in our approaches, even though there are fundamental differences in the two expressions of that spirit. Here, I list the four differences I noted, closing with a crucial observation concerning the common spirit in our work.

The differences are profound, stemming from four sources:

- Precision in defining terms and methodology;
- The importance of words in natural languages like English;
- A small number of key semantic relationships; and
- Disambiguating word senses to distinguish the linguistic contexts in which the words are used.

How can anyone oppose these four points? Surely they are common sense. Here, I address why I have a problem with each of them. Even after a decade of building CYC, it seems clear to me that the definitions of terms—and even the methodology used to codify and represent knowledge—are incomplete. They presume various assumptions and contexts, so we are never sure the definition is fully decontextualized; there can always be some hidden assumption we overlooked.

When you and I speak or write to each other, the most we can hope for is a sort of incremental approach toward agreement, toward communication, toward common usage of terms. Consider words like “sarcasm” and “expensive”; we share only most of the meaning of the words, yet still we can communicate with each other effectively. Moreover, words (in a natural language) are often more the result of historical accident than a grand plan.

Miller offers too few types of semantic relations. Where WordNet has about six such relations, CYC has thousands. Every one of them was entered into our ontology regretfully—we wanted as small a vocabulary as possible—but they had to be entered because things had to be said about them or about using them or they could not be easily and tersely expressed any other way. We found that having about a tenth as many distinct named semantic relations as the number of other (non-relation) concepts makes the statement of definitions and other information quite compact.

My fourth and final point of departure between Miller's and my expectations is in how polysemous words and pronouns are disambiguated. The techniques he cites are important, but how would they identify what the word “they” refers to in the following two sentences: “The police arrested the demonstrators because they feared violence,” and “The police arrest-

ed the demonstrators because they advocated violence”? How would they disambiguate the word “pen” in the following two sentences: “The pen is in the box,” and “The box is in the pen”? How would they decide that a tree doctor is a person who treats trees rather than a tree that practices medicine? In these examples, we draw on our commonsense knowledge of everyday objects, their size, where they are commonly found, the motives and tactics employed by various groups, the capabilities of various types of objects, and the mental and physical requirements for performing various occupations and activities.

I view our common spirit as the paradigm of tackling important large problems head-on, even if it means a lot of elbow grease brought to bear constructing an artifact, rather than a lot of elegance to prove a theorem. In this, Miller and I wholeheartedly agree. Fierce adherence to this tenet has resulted in WordNet’s being a useful, even necessary, artifact. Humanity is fortunate that Miller has had the courage and the tenacity to build it for us.

Many of my comments apply to both EDR and WordNet. For example, both take words seriously—in natural languages, specifically English and Japanese.

However, words are often red herrings. They cut up the world (ambiguously) along lines drawn for reasons mostly of historical accident, of cognation to other languages, of the need for words to be short to allow humans to breathe regularly, and other reasons. Many-to-many mapping between words and concepts is worth having in a developer’s ontology, so why not include the useful concepts? It is the word senses that need hyponyms, entailment, and other characteristics. EDR has partially adopted this point of view in its creation of a concept dictionary. WordNet also partially adopted this point of view through its replacement of words with synsets.

But both systems’ developers must still take one final step—to include concepts that are worth naming but cannot be described by a single word or synset. Examples are “bachelor party,” “JFK’s assassination,” and “the first place one can remember calling home.” Concepts are worth having as distinct first-class objects, but only if things like rules of thumb can be used to describe them. Indeed, their usefulness increases if their definition is imprecise. For example, “white cat” is an unnecessary concept because not much can be said about white cats other than the compositional facts entailed in their being cats and their being white. If we have the concepts of “white” and “cat,” we do not need to define “white cat.” However, “white-collar worker” is a much more useful concept because many things can be said about white-collar workers and yet the term is hard to define precisely.

Another point of contention between CYC and both EDR and WordNet is the number of semantic relations worth distinguishing.

I agree with many of Yokoi’s points, including the need for a large sharable knowledge base and the use of both manual and automatic methods for growing

the knowledge base and its linguistic front end. Moreover, I emphatically agree with him on the importance of including extralinguistic knowledge.

Miller on CYC and EDR

I do not like to criticize others doing work similar to my own. Creating a hand-crafted knowledge base is a labor-intensive enterprise that reasonable people undertake only if they feel strongly that it is necessary and cannot be achieved any other way. Such motivation is rare and should not be diminished by unappreciative criticism.

It is possible, however, to point out features that may help readers understand more clearly the different goals and assumptions implicit in the three examples presented here. The most ambitious goal is to provide a commonsense knowledge base adequate for any AI system. Only slightly less ambitious is the goal of providing a lexical knowledge base adequate for machine translation between Japanese and English. The goal of providing a lexical knowledge base adequate for English is modest by comparison.

Lenat discusses several fundamental assumptions underlying CYC. He and his co-workers were among the first to recognize that practical AI systems need commonsense knowledge. The assumption that commonsense knowledge is propositional and that a large but finite number of factual assertions, assisted eventually by some kind of learning machine, will satisfy this need is articulated clearly in his article. The assumption that the truth of factual assertions depends on context is also articulated; a context greatly facilitates the search for relevant factual assertions. It is also apparently assumed that some finite number of contexts will suffice. But more adventurous than the assumptions that commonsense knowledge is propositional and finite (and that generative devices are unnecessary) is the assumption that a commonsense knowledge base can be assembled in ignorance of the use to which it will be put, imagining that a single base will serve any AI problem. But Lenat’s confidence that CYC will soon provide solutions to a list of longstanding problems is admirable. One can only hope he is right.

In his account of the EDR Electronic Dictionary, Yokoi explores several assumptions that EDR shares with the WordNet project. For example, both assume that linguistic knowledge can be separated from encyclopedic or commonsense knowledge, that lexical knowledge (knowledge of words and their meanings) is necessary for computers to process human languages, and that a large but finite number of factual propositions (definitions and linguistic relations) can satisfy this need. Both also assume that word forms can be distinguished from the concepts they are used to express and that lexicalized concepts can be organized by semantic relations.

Since the EDR Electronic Dictionary and WordNet



were conceived and developed independently and in ignorance of one another, these common assumptions may reflect something basic about the nature of human languages and the human minds that use them.

Because the EDR dictionaries treat two languages, it is assumed that a concept dictionary can be constructed independently of the dictionaries of Japanese and English word forms. For example, in the EDR Dictionary, the lexicalized concepts in Japanese and English outnumber the word forms in either language by a ratio of four to three, whereas the WordNet experience is that English word forms outnumber lexicalized concepts by a ratio of four to three. Either the semantic overlap of Japanese and English words is surprisingly small or the EDR Dictionary draws finer distinctions between lexicalized concepts than does WordNet. No doubt this comparison is testimony to the difficulty of determining the nature of a lexicalized concept.

Yokoi assumes that encyclopedic concepts can be added to the large bilingual dictionary of lexicalized concepts already assembled but stops short of assuming that the kind of commonsense knowledge that CYC tries to represent can also be added to this extralinguistic knowledge base. One can only wish him success in this ambitious extension of the EDR Dictionary.

Yokoi on WordNet and CYC

The EDR Electronic Dictionary project, the WordNet project, and the CYC project share a common understanding that natural language is the key to organizing general knowledge or world sense.

Each of these three projects, however, focuses on different ways of handling knowledge. Knowledge discussed in this context includes knowledge about natural language and how it is used, as well as knowledge about what natural language represents. Suppose this knowledge is divided into three levels—the surface level, the concept level, and the knowledge level. The differences in focus for each project are:

- For the EDR Electronic Dictionary, the surface level and the concept level are considered, and only synonymy, hyponymy, and entailment are examined.
- For WordNet, the concept level is considered, and synonymy, hyponymy, entailment, antonymy, meronymy, and troponymy are examined.
- For CYC, the knowledge level is considered.

The knowledge-base products of the three projects need further improvement, expansion, and development, although the endeavor depends on first clarifying three points:

- The breadth of knowledge, including amount and type of knowledge, as well as knowledge col-

lection methods;

- The functions that may be realized through such accumulated knowledge in various ways; and
- The application systems that may be developed through such functions.

The EDR Electronic Dictionary aims broadly at realizing two functions:

- The syntactic processing function for judging sentences syntactically right or wrong; and
- The semantic processing function for taking each component word of a sentence, specifying its corresponding meaning (the concept), and deciding how deeply the components relate.

These functions are not yet completed. For example, for the syntactic processing function, efforts to expand and enrich the dictionary's vocabulary and refine its accompanying information have yet to be made. For the semantic processing function, continuous improvement and expansion will be necessary. On the whole, the quality of the English section of the dictionary is inferior to that of the Japanese section. Other functions more advanced than these functions and worthy of consideration include:

- A function that completes incomplete sentences through contextual processing that includes anaphora resolution and other operations;
- A function that verifies the basic proposition expressed in a simple sentence; and
- A function that verifies the compound proposition expressed in a compound sentence.

Meanwhile, WordNet's efforts to define a concept through various conceptual relationships are moving knowledge bases closer to the realization of a function that completes incomplete sentences. And CYC development is working toward the realization of functions that verify the propositions in simple and compound sentences. However, no discussion related to these functions is included in my article on the EDR project. When considering the general development of knowledge bases, however, it might be beneficial for the readers of *Communications* if WordNet, CYC, and EDR developers would delineate their policies, research, and future plans regarding these functions.

Lenat's Response to Miller

Miller is absolutely right on in his comments about CYC. I want to respond to only one point. We believe that educating a program—independent of specific tasks we later ask it to perform—is doable and worth doing. Although

unprecedented when building software, such education has precedents in child rearing, being the basis for most K-12 education, religious education, and parenting in general. Is our paradigm a gamble? Yes—a high-risk but high-payoff gamble that has a chance of working. One gets to make only a few decade-sized bets in one's research career, and as with Miller and Yokoi, I intend to make every one of them count for something. Miller, Yokoi, and I will likely either see our work succeed and revolutionize the world or fail spectacularly. Only one thing is certain—these three projects are not bumps on a log. It saddens me how few software-related projects I can say that about these days.

Lenat's Response to Yokoi

CYC covers the concept level and the surface level, though Yokoi is correct in assuming that we feel the knowledge level is the primary source for the other two. Lexical knowledge is necessary, but animals roamed the earth long before any of them had the concept of animal or before any of them used words to express that concept to each other. Since 1986, CYC has contained distinct terms for Red (the concept) and TheEnglishWordRed. CYC would lack common sense if it knew nothing about red, but we feel no such urgency about whether it knows the word for red in, say, Algonquin, Basque, English, or Japanese. We do not agree with Yokoi's view that natural language is the key when organizing general knowledge or word sense. To the contrary, we consider natural language to be a red herring, though one we cannot ignore forever because most people use it so heavily to communicate.

CYC must eventually understand and generate natural language, but this is an add-on to its knowledge base and reasoning engine, not an integral organizing part of them. It took millennia for society to realize that so-called deaf-and-dumb people are no more stupid than anyone else; we should not make the same mistake about our computer programs. Work is simplified if people, or programs, communicate in natural language, but that is just a veneer—not the source—of their intelligence.

Miller's Response to Lenat

I agree with Lenat that more semantic relations between word senses are desirable. WordNet uses what we regard as a minimal number, in part because we were originally interested in seeing whether relational semantics is a viable alternative to compositional semantics. By the time we were convinced that relational semantics is adequate, we had invested too much work in it to

abandon it and start over with a richer set of semantic relations. However, the set we stayed with has proved to be quickly intelligible to users with limited linguistic sophistication.

I also agree with Lenat and Yokoi that much more than a WordNet-type lexical database is required for processing natural language satisfactorily. We assumed the ability to distinguish among the syntactic categories of noun, verb, adjective, and adverb, but we have not tried to resolve problems of co-reference; nor have we suggested any way to recognize predicate-argument structures or to parse grammatical sentences. All that and more will be required for successful natural language processing. Our assumption, which may prove too optimistic, has been that the lexical component of language can be studied more or less independently of other components.

Even tasks limited to lexical knowledge involve enormous dimensions. Without some 300 years of gradually accumulated lexicographic reference works, we would be unable to imagine how lexical knowledge can be captured in a computer database. Viewed in that way, WordNet is a short chapter in a long story—an example of what is becoming possible. Future developers will surely improve on what we have begun.

Yokoi's Response to Miller and Lenat

I want to add two points:

- **Levels of knowledge.** The difference between CYC and EDR and WordNet stems from the difference between the levels of knowledge each deals with. Handling the type of knowledge that CYC has been treating entails the type of scheme adopted by its developers. Will EDR and WordNet (with their structures and future enhancement) become more like CYC? Furthermore, if it is possible for EDR and WordNet to more closely resemble CYC, what steps are needed to achieve this result? The answer is not yet clear. However, it seems an appealing theme for future research. Studying this theme seems to be a promising approach to scientifically identifying what people call common sense.
- **Word forms and concepts.** The number of word forms and the number of concepts in the EDR Electronic Dictionary are about the same. The overlapping of Japanese concepts and English concepts is smaller than we expected. However, I think this overlapping will gradually change as the dictionary is improved. Hyponymy varies depending on the viewpoint from which it is formed. If hyponymy is viewed as semantic co-occurring relations between words and predicates—as it is in EDR—I think it is more natural for Japanese hyponymy and English hyponymy to fall into separate systems of classification. □