REQUIREMENTS OF TEXT PROCESSING LEXICONS

Kenneth C. Litkowski
16729 Shea Lane, Gaithersburg, Md. 20760

Five years ago, Dwight Bolinger [1] wrote that efforts to represent meaning had not yet made use of the insights of lexicography. The few substantial efforts, such as those spearheaded by Olney [2,3], Mel•Cuk [4], Smith [5], and Simmons [6,7], made some progress, but never came to fruition. Today, lexicography and its products, the dictionaries, remain an untapped resource of uncertain value. Indeed, many who have analyzed the contents of a dictionary have concluded that it is of little value to linguistics or artificial intelligence. Because of the size and complexity of a dictionary, perhaps such a conclusion is inevitable, but I believe it is wrong. To avoid becoming irretrievably lost in the minutiae of a dictionary and to view the real potential of this resource, it is necessary to develop a comprehensive model within which a dictionary•s detail can be tied together. When this is done, I believe one can identify the requirements for a semantic representation of an entry in the lexicon to be used in natural language processing systems. I describe herein what I have learned from this type of effort.

I began with the objective of identifying primitive words or concepts by following definitional paths within a dictionary. To search for these, I developed a model of a dictionary using the theory of labeled directed graphs. In this model, a point or node is taken to represent a definition and a line or arc is taken to represent a derivational relationship between definitions. With such a model, I could use theorems of graph theory to predict the existence and form of primitives within the dictionary. This justified continued effort to attempt to find such primitives.

The model showed that the big problem to be overcome in trying to find the primitives is the apparent rampant circularity of defining relationships. To eliminate these apparent vicious circles, it is necessary to make a precise identification of derivational relationships, specifically, to find the specific definition that provides the sense in which its definiendum is used in defining another word. When this is done, the spurious cycles are broken and precise derivational relationships are identified. Although this can be done manually, the sheer bulk of a dictionary requires that it be done with well-defined procedures, i.e. with a syntactic and semantic parser. It is in the attempt to lay out the elements of such a parser that the requirements of semantic representations have emerged.

The parser must first be capable of handling the syntactic complexity of the definitions within a dictionary. This can be done by modifying and adding to existing ATN parsers, based on syntactic patterns present within a dictionary. Incidentally, a dictionary is an excellent large corpus upon which to base such a parser.

The parser must go beyond syntactics, i.e., it must be capable of identifying which sense of a word is being used. Rieger [8,9] has argued for the necessity of sense selection or discrimination nets. To develop such a net for each word in the lexicon, I suggest the possibility of using a parser to analyze the definitions of a word and thereby to create a net which will be capable of discriminating among all definitions of a word.

The following requirements must be satisfied by such a parser and its resulting nets. Diagnostic or differentiating components are needed for each definition. Each definition must have a different semantic representation, even though there may be a core meaning for all the definitions of a word. Since the ability to traverse a net successfuly depends on the context in which a word is used, each definition, i.e. each semantic representation, must include slots to be filled by that context. The slots will provide a unique context for each sense of a word. Context is what permits disambiguation. Since the search through a net is inherently complex, a definition must drive the parser in the search for context which will fill its slots. These notions are consistent with Rieger•s; however, they were identified independently based on my analysis of dictionary definitions. Their viability depends on the ability to describe procedures for developing a parser of this type to generate the desired semantic representations.

As mentioned before, observation of syntactic patterns will lead to an enhancement of syntactic parsing; to a limited extent, the syntactic parser will permit some discrimination, e.g. of transitive and intransitive verbs or verbs which use particles. Further procedures for developing semantic representations are described using the intransitive senses of the verb "change" as examples. Procedures are described for (1) using definitions of prepositions for identifying semantic cases which will operate as slots in the semantic representation, (2) showing how selectional restrictions on what can fill such slots are derived from the definitional matter, and (3) identifying semantic components that are present within a definition. It is pointed out how it will eventually be necessary that these representations be given in terms of primitives. Procedures are described for building discrimination nets from the results of parsing the definitions and for adding to these nets how the parser should be driven. The emphasis of this paper is in describing procedures that have been developed thus far. Finally, it is shown how these procedures are used to identify explicit derivational relationships present within a dictionary in order to move toward identification of primitives. Such relationships are very similar to the lexical functions used by Mel•Cuk, except that in this case both the function and the argument are elements of the lexicon, rather than the argument alone.

It has become clear that semantic representations of definitions in the form described must ultimately constitute the elements out of which semantic representations of multisentence texts must be created, perhaps with two foci: (1) describing entities (centered around nouns) and (2) describing events (centered around verbs). If multisentence texts can then be studied empirically, the structure of ordinary discourse will then be based on observations rather than theory.

Although this paradigm may seem to be incredibly complex, I believe that it is nothing more than what the lexicons of present AI systems are becoming. I believe that more rapid progress can be made with an explicit effort to exploit and not to duplicate the efforts of lexicographers.

REFERENCES

1. Bolinger,D., Aspects of Language, 2nd ed., Harcourt Brace Jovanovich, Inc., New York, 1975, p.224.
2. Olney,J., C.Revard, and P.Ziff, Toward the Development of Computational Aids for Obtaining a Formal Semantic Description of English, SP-2766/001/00, System Development Corporation, Santa Monica, California, 1 October 1968.
3. Olney,J. and D.Ramsey, "From machine-readable dictionaries to a lexicon tester: Progress, plans, and an offer," Computer Studies in the Humanities and Verbal Behavior, Vol.3, No.4, November 1972, pp. 213-220.
4. Mel*Cuk,I.A., "A new kind of dictionary and its role as a core component of automatic text processing systems," T.A. Informations, 1978, No.2, pp.3-8.
5. Smith,R.N., "Interactive lexicon updating," Computers and the Humanities, Vol.6, No.3, January 1972, pp. 137-145.
6. Simmons,R.F. and R.A.Amsler, Modeling Dictionary Data, Computer Science Department, University of Texas, Austin, April 1975.
7. Simmons,R.F. and W.P.Lehmann, A Proposal to Develop a Computational Methodology for Deriving Natural Language Semantic Structures via Analysis of Machine-Readable Dictionaries, University of Texas, Austin, 1976 (Research proposal submitted to the National Science Foundation, Sept.28,1976).
8. Rieger,C., Viewing Parsing as Word Sense Discrimination, TR-511, Department of Computer Science, University of Maryland, College Park, Maryland, January 1977.
9. Rieger,C. and S.Small, Word Expert Parsing, TR-734, Department of Computer Science, University of Maryland, College Park, Maryland, March 1979.