Technical Correspondence

The Extraction of a Minimum Set of Semantic Primitives from a Monolingual Dictionary is NP-Complete

Within the last 15 years, a variety of unsolved problems of interest primarily to operations researchers, computer scientists, and mathematicians have been demonstrated to be equivalent in the sense that a solution to any of them would yield a solution to all of them. This class of problems, known as NP-complete, contains many long-standing problems of scheduling, routing, and resource allocation. This note contains a demonstration that a problem of interest to applied linguistics also belongs to this class – namely, the process of extracting a minimum set of semantic primitives from a monolingual dictionary is NP-complete, implying that the task is currently computationally insoluble.

A particular linguistic problem has found applied relevance in three areas: natural language comprehension, bilingual dictionary construction, and reading theory. The problem is that of maximally simplifying the cross-referential lexicon known as the dictionary. For a variety of purposes of those who wish to make dictionaries computer readable, the problem of finding a "base" set of semantic primitives from which other lexical entries may be defined has been of interest. If, for example, a set of 60 lexical entries (each unwavering in semantic denotation, connotation, intent, and content) could be found, out of which all other entries could be satisfactorily defined, then certain problems of circularity of definition and of algorithmic complexity could be solved. Simiarly, the speed of construction of a truly "bidirectional" bilingual dictionary could be enhanced if those persons engaged in the construction were aware of such a minum set of monolingual primitives. This paper demonstrates that, in general, such problems are computationally intractable by virtue of their isomorphism to a group of problems known as NP-complete.

DEFINITIONS

1. A lexicon is a set.
2. The elements of a lexicon are called words.
3. A string is a sequence of words.
4. The string universe, X*, of a lexicon X is the set of all strings composed of elements of X.
5. A language (over a lexicon, X) is a subset of X*.
6. Within a language, L, a definition of a word, w, is some string in L of words within the lexicon (excluding w itself). That is, a definition associates a string with the word being defined.
7. A dictonary for a lexicon X is a set of definitions such that each word in X is defined.
8. A directed graph G=(N,L) consists of a set N of nodes, together with a set L of ordered pairs of elements of N.
9. In a graph G, a directed cycle is a sequence of nodes, $(n_1, n_2, \ldots, n_K)$ in which each of the lines (n_i, n_{i+1}) as well as (n_K, n_1) is in L.

DISCUSSION OF DEFINITIONS

It may appear backwards to define a language based on its words, since in agglutinating languages, for example, the determination of what is or is not a word is based on extensive prior knowledge of the language. However, this notation is fairly common, which is the reason for its adoption. One might take the morpheme as the primary element rather than the word without loss of applicability of these remarks.

Also, we might be tempted to define a sentence as a string contained in a given language, though such is not needed here.

Definition 6 excludes the possibility that a word might appear within its own definition. This restriction can be relaxed under certain circumstances.

The problem of semantic reduction to a minimum set of lexical primitives may now be stated thusly: Given a dictionary, we seek to rewrite that dictionary, substituting definitions for words freely so as to minimize that portion of the lexicon occurring as members of defining strings. That is, what is the smallest number of words in the lexicon such that all other words may be defined from this select set?

Karp (1972) demonstrated that the problem "Feedback Vertex Set", FVS, is NP-complete. [For readers unfamiliar with the concept of NP-completeness, Garey and Johnson (1979) present an overview of the topic.] The following shows that our semantic reduction problem is equivalent to FVS. FVS is stated as follows by Garey and Johnson: Given a directed graph G=(N,L) and a positive integer k, is there a subset of N consisting of k or fewer nodes that contains at least one vertex from every directed cycle in G?

To show the problems equivalent, we first note that for the sake of simplifying the dictionary we are not concerned, per se, with the order of words within defining strings. That is, so long as we keep track of this ordering, it will not affect the ultimate size of the defining lexicon. Let D be a dictionary for the lexicon X. We now construct a directed graph G based upon D: let each w in X be a node of G. Now for each w construct a line leading from w to any word occurring in the string defining w. The construction of G is now complete.

We now observe that asking the question of the size of the smallest set of entries from which D may be reconstructed is computationally equivalent to asking whether or not there is a set of k such entries (and re-asking this question for a new value of k). Next, in redefining a word w, we may explore any path leading from w. If all such paths terminate in primitives, then w has been defined in terms of primitives. Yet, if any such path returns to w, then w has not been appropriate defined. We are therefore concerned with finding the smallest set of nodes that will "stop" any directed cycle. This is precisely the problem of FVS.

Example: Let D be given as follows:

X = {ama,ba,di,enig,gala,ki,li,tso,ub,zomir}

ama = di gala li tso
ba = li zomir ki enig
di = ub enig ki zomir ba tso
enig = di zomir ba ki
gala = ub tso ub li
ki = ba tso ub tso li
li = ub ki di gala ba enig
tso = ba zomir ki li gala
ub = di li zomir
zomir = di gala ba tso

Clearly, no human language would be representable by such a small dictionary; any larger lexicon, though, would not be easily exemplified herein. The transformation of D to a directed graph is shown below.

A SIMPLIFIED DICTIONARY WITH FIVE PRIMITIVES

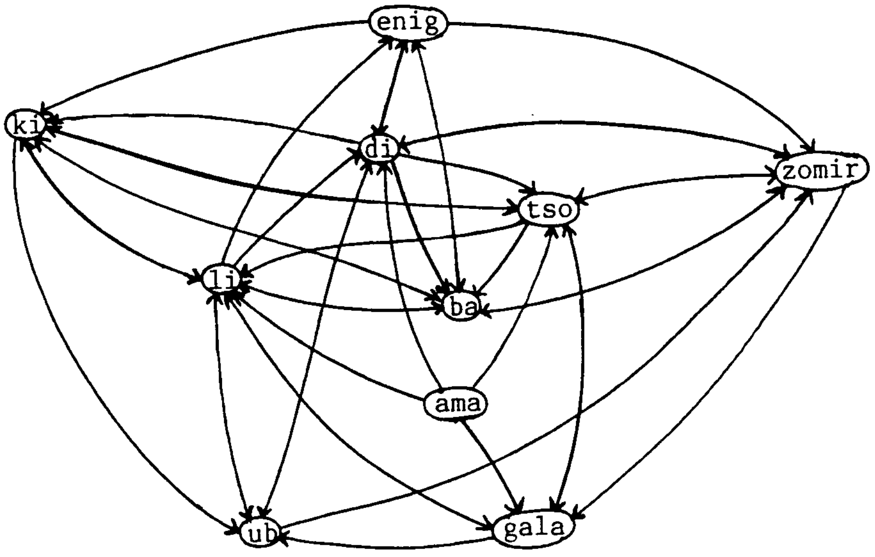
ama = di ((di li zomir) tso (di li zomir) li) li tso
ba + primitive +
di + primitive +
enig = di zomir ba (ba tso (di li zomir) tso li)
gala = (di li zomir) tso (di li zomir) li
ki = ba tso (di li zomir) tso li
li + primitive
tso + primitive +
ub = di li zomir
zomir + primitive +

David P. Dailey
Department of Psychology
University of Alaska
Fairbanks, Alaska

References

Garey, M.R. and Johnson, D.S. 1979 *Computers and Intractability.* W.H. Freeman and Company, San Francisco, California.

Karp, R.M. 1972 Reducibility among Combinatorial Problems. In Miller, R.E. and Thatcher, J.W., Eds., *Complexity of Computer Computations.* Plenum Press, New York, New York: 85-103.


enig
ki
di
zomir
tso
li
ba
ama
ub
gala