LEXICAL KNOWLEDGE BASES

Robert A. Amsler
Natural-Language and Knowledge-Resource Systems
SRI International
Menlo Park, California 94025, USA

A lexical knowledge base is a repository of computational information about concepts intended to be generally useful in many application areas including computational linguistics, artificial intelligence, and information science. It contains information derived from machine-readable dictionaries, the full text of reference books, the results of statistical analyses of text usages, and data manually obtained from human world knowledge.

A lexical knowledge base is not intended to serve any one application, but to be a general repository of knowledge about lexical concepts and their relationships. Thus natural-language parsers, generators, or other intelligent processors must be able to interface to the knowledge base and are expected to only extract those portions of its knowledge which they need for specific tasks. Likewise, the knowledge base is designed, built, and maintained primarily as a repository--rather than a tool serving the needs of other computational processors. Just as human memory, the knowledge base doesn't distinguish between 'useful' knowledge and information for which it at present doesn't have any functional use. In this manner the knowledge base is a test bed for concept representation mechanisms and data structures, rather than an adjunct to other computational processes.

Investigations of machine-readable dictionaries over the last decade have shown that they can be computationally useful for tasks such as parsing, computer-assisted instruction, speech generation, and content analysis. Sufficient knowledge of the contents of machine-readable dictionaries now exists to provide meaningful answers to questions concerning what additional information about lexical concepts will be needed to represent many aspects of human 'world knowledge.'

Machine-readable dictionaries are seen as providing an index into human knowledge. A dictionary definition provides the minimal information necessary to evoke the concept it defines in the mind of a human reader who already knows to what this concept refers. It is neither intended nor capable of serving as the actual 'meaning' of that concept. A lexical knowledge base is intended to provide a means of economically integrating not only dictionary definitions, but other types of lexical knowledge. The task of constructing a lexical knowledge base is seen as a goal in itself, distinct from the task of building natural language processing programs that will use that knowledge base.

Several of the components of a lexical knowledge base are already known and await assembly into one database. One component is the tangled-hierarchy of concepts compiled as part of an analysis of the kernels of the definitions in a dictionary. This 'tangled' hierarchy provides ISA arcs connecting 27,000 nominal concepts and 12,000 verbal concepts derived from the Merriam-Webster Pocket Dictionary [Amsler 1980]. Another component of the lexical knowledge base has been provided by the extraction of subject codes from the Longman Dictionary of Contemporary English. Some 17,000 concepts in the Longman dictionary possess subject designations that give the domain in which these concepts are used.

There is a subtle distinction between the ISA hierarchy and the subject classification that is worth mentioning. A word such as 'crossbow' is taxonomically linked to 'weapon' in the ISA hierarchy; but appears in the subject domain 'military history.' Subjects thus do not duplicate ISA linkage information, but add another facet to conceptual understanding.

There are a number of additional machine-readable dictionary properties that can of course be combined into a lexical knowledge base. Machine-readable dictionaries contain information regarding the appropriate level of usage of concepts; their geographic or chronologic associations; ,and semantic and syntactic restrictions on their potential arguments and combinations.

In addition to this immediatly available information listed for each concept in dictionary definitions, dictionaries contain much implicit information derivable from studying collections of definitions. For example, the verbs of motion can be analyzed to reveal much more about their core concept 'move' than would be seen from its definition alone.

Two major components of conceptual understanding which dictionaries fail to adequately describe are procedural knowledge and information derived from the mental inspection of visual imagery. Sources for procedural knowledge may exist in other types of special purpose reference books, such as encyclopedias; but information derived from conceptual visual images will require special encoding to be useful for computational reasoning. Many questions of relative and absolute size, position, and orientation are not answerable from definitions. While some sizes are available from reference books, there nevertheless remain many aspects of our understanding of tangible objects which can only be answered by examination of illustrations or scenes in which the objects appear.

Such illustrations are, however, an accepted part of many dictionaries and other lexical reference books. The famous 'Duden' series of pictorial dictionaries provide line drawings and illustrations of tangible objects, often collectively depicted in scenes which relate large amounts of information about their relative sizes, uses, etc. Such information will require encoding methods that bridge the gap between natural language understanding research and vision research.

Other line drawings often show the a series of images of human figures going through the steps of an athletic event, such as diving into a swimming pool, or performing a pole vault. The information shown is chronological and spatial, giving relative locations of the performer throughout time. Capturing this pictorial information in a lexical knowledge base will be necessary for it to contain the data needed to fully understand text.

These tasks are seen as providing the basis for building lexical knowledge bases. The fundamental question governing whether new information must be added to a lexical knowledge base shall be whether natural-language understanding problems demonstrate the need for the information and it can be shown to not be inferrable from existing material in the knowledge base.

[After July, 1984 the author will be joining the Artificial Intelligence and Information Science Group at Bell Communications Research in Morristown, New Jersey. Funding for this paper was provided in part by NSF grants IST-8208578, IST-8209346, and IST-8300940.]