8-22-1989

Extraction of Knowledge about the Cognitive Process of Browsing from Discourse and Thinking-aloud Protocols. Also Published as Part of: Kwasnik, B. et al Automatic Extraction from Dictionary Text: Project Development.

Barbara H. Kwasnik
Syracuse University

Elizabeth D. Liddy
Syracuse University

Myaeng H. Sung
Syracuse University

Extraction of Knowledge about the Cognitive Process of Browsing from Discourse and Thinking-aloud Protocols. Also Published as Part of: Kwasnik, B. et al Automatic Extraction from Dictionary Text: Project Development.

Keywords

Browsing, functional browsing, Classification, Classifying documents, Explorable vocabulary.

Disciplines

Library and Information Science

Automatic Knowledge Extraction from Dictionary Text: Project Development

Barbara H. Kwasnik
Elizabeth D. Liddy
Sung H. Myaeng
Information Studies and CASE Center
Syracuse University
Syracuse, New York 13244-4100

August 1989

CASE Center Technical Report No. 8911

These papers describe a research project supported by a CASE Center R&D Grant (G-31) during May/ June 1989:

Automatic Knowledge Extraction from Dictionary Text.

The papers were presented at the 11th IJCAI Conference in Detroit and are to be published in the individual workshop Proceedings:

Barbara Kwasnik, "Extraction of Knowledge about the Cognitive Process of Browsing from Discourse and Thinking-Out-Loud Protocols," presented at the Workshop on Knowledge Acquisition, August 22, 1989.

Elizabeth DuRoss Liddy, "Explorable Vocabularies," presented at the First International Lexical Acquisition Workshop, sponsored by AAAI in conjunction with IJCAI-89, August 21, 1989.

Sung Myaeng, "Automatic Knowledge Extraction from Dictionary Text Using Conceptual Graphs," presented at the Fourth Annual Workshop on Conceptual Structures, August 20-21, 1989.

CASE Center Technical Report Series
New York State Science and Technology Foundation
Centers for Advanced Technology Program

EXTRACTION OF KNOWLEDGE ABOUT THE COGNITIVE PROCESS OF BROWSING FROM DISCOURSE AND THINKING-OUT-LOUD PROTOCOLS

Barbara H. Kwasnik
School of Information Studies
Syracuse University

May 28, 1989

PROBLEM

This work is part of an ongoing project at Syracuse University School of Information Studies whose working title is "Automatic Knowledge Extraction from Dictionary Text." Drs. Elizabeth Liddy, Sung Myaeng, and I have as our goal a computational approach for extracting semantic relationships from dictionary entries for use in building a conceptual representation of knowledge that can be browsed interactively. Achieving this goal requires parallel attention to the linguistic, computational, and behavioral/cognitive aspects of the problem.

The behavioral aspect of special interest to us is browsing, the surfacely simple act of nondirected searching for something of interest to the individual. Since we would like our representations of relationships among concepts to be accessible for interactive browsing by users, we must first better understand browsing as a cognitive behavior.

There are two major questions:

1. Can browsing behavior be described functionally? That is, is there a method by which we can tap the cognitive process of browsing and describe it in terms general enough to account for a variety of specific behaviors? and
2. Can this behavior be described at a fine enough level of granularity to be useful in specifying the characteristics of an automatically produced explorable vocabulary?

APPROACH

The project discussed above is still in the preliminary stage and has not yet been reported anywhere in the literature. Related work in studying cognitive processes (Kwasnik 1987; Kwasnik 1989) suggests that it is possible to elicit information about cognitive processes from participants within the context in which the behavior takes place, that people are able to articulate considerable information about these processes, and that the data produced by this articulation lend themselves to analysis at a level which can yield general rules about the behavior.

In the study cited above, eight university faculty members were asked to describe their own offices in terms of the organization of documents. Each subject was also asked to sort a day's mail. The method of data collection was interview and thinking-out-loud protocols. The resulting discourse was analyzed and coded. Each classificatory decision (i.e., each instance of naming or talking about a document's placement or disposition) was coded using a set of inductively derived descriptive dimensions. It was possible to describe a wide range of specific behaviors using a relatively small set of descriptive dimensions along which decisions had been made. That is, although human behavior varies, there is enough regularity to make descriptions possible at a general level.

Next, the discourse was reformulated into rules or IF.... THEN statements. Each rule represented a classificatory decision. Based on a knowledge of these rules, the researcher then attempted to sort and classify the documents of four of the subjects in the same way each of them might have done it. Based on the analysis of a one-time interview only, the researcher succeeded in classifying documents correctly almost two-thirds of the time.

The findings of this study suggest that it is possible to describe cognitive behavior (such as classificatory decisions) at a functional level using a small set of descriptive dimensions. The study also demonstrates that in modelling such behavior it is extremely important to develop some method of modelling the context, that is the circumstances, the person's goals, knowledge and so forth.

We intend to use a similar set of techniques to study and describe the process of browsing. In many ways browsing is similar to the process of classification. Both involve cognitive processes that are made manifest through observable behavior. Both are everyday behaviors readily understood by most people. We think people can tell us about the process of browsing while they are engaged in it. Based on the study briefly outlined above and others, we think that a variety of behaviors called "browsing" can be described by a manageably small set of descriptors.

What remains an open question is whether or not such descriptions of what tasks people accomplish when they browse and what links they establish among the various components of the domain being browsed can be specified at a fine enough level to be of use in an interactive explorable vocabulary. The answer to this question must await the results of work on all three aspects of the project: the linguistic, the computational, and the behavioral. In any event, the intermediate process of learning more about browsing will shed light on questions of importance in artificial intelligence and information science.

REFERENCES

KWASNIK, BARBARA H. The influence of context on classificatory behavior. Ph.D. Dissertation. Rutgers University, May, 1989.

KWASNIK, BARBARA H. A method of studying classificatory behavior. In: Proceedings of the Fifth International Conference on the Empirical Foundations of Information and Software Science, Riso National Laboratory, Roskilde, Denmark, November 23-25, 1987.

EXPLORABLE VOCABULARIES

Elizabeth DuRoss Liddy[1]
School of Information Studies
Syracuse University, Syracuse, New York

ABSTRACT

A project is described whose goal is to automatically extract knowledge in the form of concepts and relations from a machine readable dictionary, specifically Longman's Dictionary of Contemporary English. Defining formulae will serve as indicators to the nature of the relations among concepts in a definition and conceptual graphs will be used as the knowledge representation scheme. An explorable vocabulary consisting of nodes and relationally-labelled arcs will be developed for use as a browsing tool, based on results from a behavioral study of the cognitive behavior of browsing.

PROBLEM

Individuals who are attempting to learn about a new domain of knowledge are obviously at a disadvantage due to their relative unfamiliarity with the semantic structure of the concepts in the new field of interest. The most appropriate learning strategy in such a situation might be to conduct an exploratory foray into the subject area in an attempt to learn the vocabulary of the field and to discover how the knowledge is conceptually organized. One could envision some type of a graphical semantic representation (e.g. semantic network) being a useful tool at this moment; presenting the concepts of the field and specifying the semantic relations among them. This "explorable vocabulary" could serve to clarify an individual's original notions; to provide appropriate terminology for further investigation of the topic; and to suggest richer connections between concepts which had not yet occurred to the novice.

Drs. Sung Myaeng, Barbara Kwasnik and myself have embarked on such a project. We are devising a computational approach for extracting semantic relations from dictionary entries for use in building a conceptual representation of knowledge that can be browsed interactively. Such a tool might be used by the novice in a field, as described above, or in a variety of other cognitive browsing situations to be discussed later. In order to accomplish the general goal of automatically creating such an "explorable vocabulary", our research takes us into three inter-related areas of investigation. These can be categorized as linguistic, representational, and behavioral aspects and each will be explored in this paper.

---

[1]The research reported herein was supported by a seed grant from the Center for Advanced Technology in Computer Applications and Software Engineering at Syracuse University.

LINGUISTIC ASPECTS

Knowledge can be thought of as concepts and the relations between these concepts. Although the concepts which comprise knowledge are of significance, we are particularly interested in investigating the nature of the relations that exist among the concepts. In particular, we will focus on knowledge encoded in language, and more specifically on how the language used in dictionary definitions implicitly reveals the semantic relations that exist among concepts.

We have chosen dictionaries as our lexical database because dictionaries are culturally validated sources of knowledge in that they contain information that has been accepted by native language speakers over many years. In addition, they provide a more complete coverage of the language than would be available in either narrative or expository samples of text. They provide an excellent database when the goal is to build a semantic representation that contains the concepts and relations across a number of subject or topic areas. Of the dictionaries that are available in machine readable form, we are planning to use Longman's Dictionary of Contemporary English (LDOCE). There are a variety of reasons for this choice, including the fine work done on LDOCE by other researchers [Alshawi, 1987; Boguraev & Briscoe, 1987; Wilks et al, 1987], but one of our primary reasons is LDOCE's use of a base defining vocabulary of 2000 words. As will be explained later in this paper, this factor is of prime significance to the approach taken in this project.

In order to discuss the process of extracting semantic relations from dictionary entries, the nature of relations must first be presented. Relations are properties that hold between two or more entities. The entities may be people, events, objects, situations, actions, words, places, etc. Relations define the nature of the interaction, dependency, influence or simply co-occurrence that holds between the entities. Relations have been of research interest in a variety of disciplines such as psychology, cognitive science, ethnography, and linguistics. The recent text by Evens [1988] provides an excellent introduction to and summary of the work done on relational models in linguistics, psychology, and computer science.

Although there has been no consensus among researchers as to what the set of useful relations is, with some models based on as few as three [Werner, 1988] and other models such as Evens [1981] including more than one hundred, many of the models do have a good number of relations in common [Evens et al, 1980]. Some of the more frequently included relations are provided below, along with example definitions from LDOCE. Capitalized terms exist in the relation being exemplified. The underlined terms are the phrases which suggest the nature of that particular relationship and will be discussed later as "defining formulae". For example, BECOME THINNER exists in the goal relationship to DIET and this relation is indicated by the defining formula in order to.

taxonomy - WHIRLIGIG: a TOY

synonymy - DOCKYARD: a place where ships are built; SHIPYARD
cause - DEPRESS: to make LESS ACTIVE OR STRONG
characteristic - SMOOTH: having AN EVEN SURFACE
agent - SHOWMAN: a person whose business is PRODUCING PLAYS
instrument - STAB: to strike forcefully with THE POINT OF SOMETHING SHARP
goal - DIET: to eat according to a special diet, esp. in order to BECOME THINNER
location - SHOOT: an area of land where ANIMALS ARE SHOT FOR SPORT
purpose - NAIL FILE: a small instrument with a rough surface for SHAPING FINGER NAILS

As can be seen in these sample definitions, the relations that exist between concepts in the definition and the term being defined communicate much of the meaning that is conveyed by the definition. These relations indicate how the concepts are structured in regard to each other. In fact, it is difficult to envision how a semantic representation would communicate if only the concepts that appear in the definition were included. Perhaps because the terms being used here as examples are commonplace ones, this is not quite so obvious. But if one were not fluent in the English language or the definitions were from an unfamiliar field of knowledge, the vital role played by inclusion of the specific nature of the relations would be more dramatically clear.

Given that we view relations as essential to truly useful semantic representations, our first task is to determine which of the myriad semantic relations used in other relational models are either explicitly or implicitly present in dictionary definitions. The second task is the delineation of the specific lexical constructions which reveal these relations. These lexical constructions have been referred to as "defining formulae" [Smith, 1981; Ahlswede, 1985] and can be thought of as the sublanguage of dictionary definitions. Sublanguage theory [Sager et al, 1987] suggests that any type of text that is used within a group of individuals for a common purpose will develop characteristic syntax and semantics. This notion is reflected in the "defining formulae" of dictionary entries, since lexicographers are a specialized group working on a common task. For example, the 'characteristic' relation can be indicated by the presence of one of the following defining formulae:

having _____
marked by _____
characterized by _____
possessing _____
showing _____

Much of the research which has made use of defining formulae in lexicographic analysis has either focused on only one part of speech (e.g. nouns or verbs or adjectives) or attempted to identify defining formulae for only a few of the possible relations (e.g. taxonomy or synonymy). Our preliminary analysis of LDOCE suggests

that a wide range of relations are implicitly present in its definitions and our project will identify all relations that occur in the definitions and extract their defining formulae.

Our preliminary analysis suggests that there is a scale of complexity in the nature of defining formulae, with some levels requiring more semantic information to be available about the terms in the defining formulae. The base level of defining formulae require simple lexical string matches. For example: 'used as....' indicates the purpose relation. However, defining formulae may be classified at a more general level. For example, the taxonomic relation can be expressed by any of the following lexical constructions:

a kind of _____
a type of _____
a branch of _____

which can be more generically captured in the defining formula:

'a'/'an' + general noun + 'of'

Since the set of general nouns in the English language is of reasonable size, a simple list of these nouns could be developed for look-up and substitution in this defining formula.

At the next level of complexity, an even more semantically based formula might be written, for example, for the agent relation as:

'a'/'an' + animate noun + 'who'

Defining formulae at this level of complexity would require that terms in the dictionary (which will be consulted when processing other definitions) be tagged for whatever features (e.g. animate) that are specified in any of the defining formulae. Our work will make use of the full range of defining formulae enumerated above.

In addition, our preliminary analysis indicates that there is a one-to-many ratio between relations and defining formulae, that is, there is more than one way to express a relationship. This poses no problem. However, for the defining formulae to function as needed, it is necessary that there not be a many-to-one ratio between relations and any single defining formula. That is, it should not be the case that one defining formula is used to indicate more than one relation. This remains an empirical question to be investigated in our project.

REPRESENTATIONAL ASPECTS

Having surveyed the variety of representational schemes suggested in the literature,

we have chosen Sowa's conceptual graphs [Sowa, 1984]. Although there were numerous reasons for our choice, only a few of the most essential ones will be presented here.

Conceptual graphs form a knowledge representation language in which concept nodes represent entities and relation nodes show how the concepts are interconnected. When this definition of conceptual graph is compared to the description of the knowledge contained in dictionary entries that needs be represented in the tool we are developing, the appropriateness of the conceptual graph representation can be seen. More important, however, is the fact that conceptual graphs are formally defined, with theorems and proofs that dictate how the meaning of a propositional statement (e.g. dictionary definition) can be interpreted and translated into a conceptual representation. In addition, Sowa's scheme provides a set of standard operations that can be performed on conceptual graphs. These operations provide the tools necessary to construct the interconnected, explorable vocabulary we are aiming for. [See Sowa and Way, 1986 for a concise description of these operations, or Sowa, 1984 for full detail].

There are four basic formation operations; *copy, restrict, join, and simplify*. As an example of their usefulness for our purposes, consider the *join* operation which creates a single graph by merging two graphs on a single matching concept. For our purposes, this means that if definition A has a matching concept with definition B, the *join* procedures will add the relations and concepts in the graph of B to the matching concept in A and the pointers in B that point to this concept will be reset to point to the matching concept in the graph of A. Then the *simplify* operation routines will check each relation connected to the newly joined concept for duplicate concepts and relations and eliminate the redundant ones. There are other, more complex operations available in this scheme, such as *relational expansion* and *relational contraction* which will be useful in our research but will not be further detailed here.

Our methodology is to first develop a canonical graph (conceptual graph for a concept or relation type) for each of the approximately 2000 terms in the defining vocabulary of LDOCE. These terms themselves are defined in LDOCE and from their definitions the canonical graphs will be built. Then, using these graphs in conjunction with the defining formulae described above, a conceptual graph for each definition in LDOCE can be constructed. The advantage of combining the conceptual graph approach with the knowledge contained in defining formulae comes into play in lexical disambiguation. As stated in Sowa and Way [1986, p. 67], some words have multiple senses each with a different canonical graph. For example, the preposition 'to' may indicate either the destination or recipient relation, and the preposition 'by' may indicate either the instrument, location, or agent relation. Sowa and Way [1986] suggest that when such words are encountered, the semantic interpreter must consider multiple graphs until they are blocked by failing to find an acceptable *join* . Multiple graphs may still result . However, the defining formulae (e.g. 'by' + animate noun = agent) can be used to provide

information as to what the specific relation, and therefore word sense, is in this definition. The result would be one unambiguous graph.

As the above discussion might suggest, the approach we are taking could produce an overly connected representation when applied to the LDOCE definitions. This is an empirical question. Decisions need to be made as to whether all or only a subset of relations should be included in the explorable vocabulary when presented to a browser, and if so which relations. In addition, there are other questions as to the presentation of the explorable vocabulary which require investigation.

BEHAVIORAL ASPECT

The specifics of the display of the explorable vocabulary as it will be presented to users will be determined by what our research into the nature of the cognitive behavior of browsing reveals. We define browsing as the nondirected search for something of interest to the individual doing the browsing. Browsing is a surfacely simple cognitive behavior about which little is known. Much that is available in the literature on browsing is concerned with system browsing, that is, the navigational techniques built into a system which determine the order or pattern in which the system guides the user through the data or an application [Palay and Fox, 1981; Cove and Walsh, 1988]. Our perspective on browsing is quite different. We are interested in browsing as a cognitive behavior evidenced in a variety of situations in which an individual cannot specify in advance precisely what it is he is searching for, but which he will recognize as 'a find' when he discovers it. We will study browsing behavior in order to describe it functionally, at a general enough level to account for a variety of specific behaviors, and simultaneously at a fine enough level of granularity for it to be useful in specifying the characteristics of an optimally useful explorable vocabulary.

We will make use of observation and thinking-out-loud protocols on a set of subjects involved in the process of browsing. Since browsing is a relatively common everyday behavior and earlier research [Kwasnik, 1987] has demonstrated that other cognitive behaviors (i.e. classification) can be tapped through these two procedures, we will observe and have individuals tell us about the process of browsing while they are engaged in it. An interesting study by Canter, Rivers, and Storrs [1985] analyzed users routes of navigation through a network-structured database. Command and menu interfaces were analyzed and users' movements were characterized by 'pathiness', 'ringiness', 'loopiness', or 'spikiness'. Although the environment in that study was quite different from what the explorable vocabulary will be, this is the only study we have found which characterized the functional browsing behavior of the user.

We will make use of browsing studies in two ways. First, to identify the relations which browsers in a language-based task identify as most desirable and useful to them in augmenting the creative discovery process. These results will suggest which

of the relations that are explicitly or implicitly contained in dictionary definitions should be included in the explorable vocabulary. Second, we will study browsers as they actually explore the vocabulary representation as it evolves during the project. This continual input will aid us in devising the optimum presentation.

CONCLUSIONS

The research project delineated in this paper has a very practical goal - the development of a tool, an explorable vocabulary. This tool will make the wealth of semantic information which is explicitly and implicitly present in a dictionary accessible and navigable to anyone interested in browsing through the entities and relations which are connected through links in their definitions. The conceptual and relational information accessible in a richly connected network-like representation in which the nature of the relationships between concepts are labelled, could serve as a powerful tool for augmenting the cognitive processes involved when individuals attempt to access or add to their current knowledge structures. The tool might be used for a variety of purposes:

1. Familiarization with a new field in which more richly structured knowledge is needed,thereby serving as an orientation device.

2. Establishment of links between hitherto unrelated concepts in a fairly new field of inquiry in which connections still need to be found, thereby serving as a means for discovering novel connections.

3. Assistance in the relatively simpler task of writing about and explaining ideas to someone not so familiar with the field and the need is to make obvious all the connections and relations with which the writer is overly familiar and therefore likely to leave out.

In all of these situations, the type knowledge representation that can be automatically elicited from dictionary definitions by a methodology which makes use of the two theoretically tested notions of defining formulae and conceptual graphs, offers great potential as a useful cognitive tool. Rather as a writing aid, a cognitive map for further exploration of ideas, or as a technique for improving queries put to a retrieval system [Wang, Vandendorpe and Evens, 1985], an explorable vocabulary containing conceptual representations based on semantic relations in dictionary entries has the potential for offering substantive cognitive assistance.

In addition, the nature of the investigation promises to shed light on three vital issues of concern to researchers in artificial intelligence, namely, the automatic extraction of knowledge from dictionary text; the use of conceptual graphs as a representational technique for capturing and then displaying concepts and relations in a useful semantic representation, and; the increased understanding of the

cognitive behavior of browsing.

REFERENCES

[Ahlswede, 1985]. Thomas E. Ahlswede. A linguistic string grammar of adjective definitions from *Webster's Seventh Collegiate Dictionary.* In *Humans and machines; Proceedings of the 4th Delaware Symposium on Language Studies*. Ablex, Norwood, New Jersey, 1985.

[Alshawi, 1987]. Hiyan Alshawi. Processing dictionary definitions with phrasal pattern hierarchies. *Computational Linguistics,* 13 (3-4), 195-202, July-December, 1987.

[Boguraev and Briscoe, 1987]. Bran Boguraev and Ted Briscoe. Large lexicons for natural language processing: Utilising the grammar coding system of LDOCE. *Computational Linguistics,* 13(3-4): 203-218, July-December, 1987.

[Canter, Rivers, and Storrs, 1985]. David Canter, Rod Rivers, and Graham Storrs. Characterizing user navigation through complex data structures. *Behavior and information technology,* 93-102, 1985.

[Cove and Walsh, 1988]. On line text retrieval via browsing. *Information Processing & Management,* 24(1):31-37, 1988.

[Evens, 1981]. Martha W. Evens. *Structuring the lexicon and the thesaurus with lexical-semantic relations.* Final report to the National Science Foundation on grant IST-79-18467.

[Evens, 1988]. Martha W. Evens, editor. *Relational models of the lexicon: representing knowledge in semantic networks.* Cambridge University Press, Cambridge, England, 1988.

[Evens et al, 1980]. Martha W. Evens, Bonnie E. Litowitz, Judith A. Markowitz, Raoul N. Smith, and Oscar Werner. *Lexical-semantic relations: a comparative survey.* Linguistic Research, Edmonton, Canada, 1980.

[Kwasnik, 1987]. Barbara H. Kwasnik. A method of studying classificatory behavior. In *Proceedings of the Fifth International Conference on the Empirical Foundations of Information and Software Science,* Roskilde, Denmark, November, 1987.

[Palay and Fox, 1981]. Andrew J. Palay and Mark S. Fox. Browsing through databases. In R. N. Oddy (Ed.), *Information retrieval research,* pages 311-324. Butterworth's, London, 1981.

[Sager et al, 1987]. Naomi Sager, Carol Friedman, and Margaret S. Lyman. *Medical language processing: Computer management of narrative data.* Addison-Wesley, Reading, Mass. 1987.

[Smith, 1981]. Raoul N. Smith. On defining adjectives - Part III. *Dictionaries: Journal of the Dictionary Society of North America* , 3:28-38, 1981.

[Sowa, 1984]. John F. Sowa. *Conceptual structures: Information processing in mind and machine.* Addison-Wesley, Reading, Mass, 1984.

[Sowa and Way, 1986]. John F. Sowa and Eileen C. Way. Implementing a semantic interpreter using conceptual graphs. *IBM Journal of Research Development,* 30(1):57-69, January, 1986.

[Wang, Vandendorpe, and Evens, 1985]. Yih-Chen Wang, James Vandendorpe, and Martha Evens. Relational thesauri in information retrieval. *Journal of the American Society for Information Science.* 36(1):15-27, 1985.

[Werner, 1988]. Oscar Werner. How to teach a network: minimal design features for a cultural acquisition device. In M. W. Evens (Ed.), *Relational models of the lexicon: representing knowledge in semantic networks,* pages 141-166. Cambridge University Press, Cambridge, England, 1988.

[Wilks et al, 1987]. Yorick Wilks, Dan Fass, Cheng-Ming Guo, James E. McDonald, Tony Plate, and Brian M. Slator. A tractable machine dictionary as a resource for computational semantics. In *Proceedings of the Workshop on Natural Language Technology planning* , pages 1-27, Blue Mountain Lake, New York, 1987.

Automatic Knowledge Extraction from Dictionary Text Using Conceptual Graphs

Sung H. Myaeng[1]
School of Information Studies
Syracuse University

shmyaeng@rodan.acs.syr.edu

ABSTRACT

Dictionaries are a rich source of domain-independent, culturally validated knowledge. Our attempt is to extract and represent such knowledge automatically from a machine-readable dictionary for interactive browsing, with parallel attention to three inter-related aspects: linguistic, computational, and behavioral. Our approach is to apply Sowa's semantic interpreter based on the theory of conceptual graphs, in conjunction with defining formulas to be generated from our linguistic study.

INTRODUCTION

The interest and need to represent concepts and their associations have been witnessed by continuing research since Quillian's intent (1966) to represent the semantics of English words in a network model (hence the term "semantic network"). Regardless of whether the interest is in modeling human memory or in representing knowledge to be exploited by artificial intelligence systems, the process of developing such semantic representation of concepts and their relationships has been mostly done manually and limited to a narrow domain, restricting its use in applications. Our work has begun with a belief that such information exists not only in the individual's mind, but also in culturally validated bodies of knowledge such as dictionaries and that the information is extractable and representable through an automatic, computationally feasible process. Our goal is to devise a computational approach to analyzing dictionary text, extracting relationships among concepts, and representing them for interactive browsing.

One can think of a range of applications of a semantic network where concepts correspond to lexical items and their relationships to lexical relations. As in [Evens 1988, pp 16-22], two fundamental categories of applications are: computer models constructed by anthropologists, linguists, and psychologists in order to investigate the implications of their theories; and lexicons and knowledge bases built by computer scientists for applications such as information retrieval, natural language

[1]The preliminary study of this project was supported by a seed grant from the CASE Center at Syracuse University.

processing, or other attempts to build intelligent systems. While we do not eliminate any of these areas as a potentially promising application of the representation we are constructing, one particular area of our interest is browsing, an activity of moving about an information source with or without a search criterion in an attempt to satisfy relatively unspecified information needs.

Browsing is different from those application areas where semantic networks have been used in AI research. Most notably, we consider browsing as an activity usually performed by humans, which needs to be distinguished from a goal-driven search performed by an AI reasoning system. Although one can think of a computer program browsing a database nondeterministically for a certain task, typical "browsing" systems simply provide an interface that gives human users freedom to navigate a database, knowledge base, etc., which would otherwise be searched algorithmically by a computer (see, for example, [Palay, Fox 1981] and [Campagnoni and Ehrlich 1989]). In other words, computer systems have been developed to enhance human browsing but not to replace it. As in the literature ([Bawden 1986] and [Cove, walsh 1988]), we view human browsing as a cognitive as well as physical activity and thus believe that the selection and the presentation of the concepts and relations to be embedded in the representation should be based on the criteria reflected in human behaviors, not just based on the criteria used for computer reasoning in AI research.

We have undertaken the project with parallel attention to three inter-dependent aspects: linguistic, computational, and behavioral. In the linguistic aspect, we first attempt to identify semantic relations present, either explicitly or implicitly, in machine-readable dictionary definitions. The regularity of definition text is then exploited to delineate the lexical clues essential to the process of automatically extracting the semantic relations among concepts. In parallel with the linguistic analysis, we have embarked on the study of issues related to the computational aspect: how the semantic relations can be translated to a representation and what representational scheme would be most appropriate. Another issue is related to development of a user interface for interactive browsing of the conceptual representation we are building. The behavioral aspect comes into play naturally to guide the two aspects described hitherto. The result from the study of human cognitive behaviors in a variety of browsing tasks will specify desirable characteristics of the conceptual representaion, which will eventually determine the kind of semantic relations and the level of granularity. It will also provide a set of design guidelines for a user interface to the conceptual representation.

USE OF DICTIONARY

Unlike many previous attempts to build a conceptual representation for a well-defined, specialized domain, our focus in this project has been on a general-purpose dictionary. In addition to the wealth of rich, detailed semantic information that has attracted some work on semantic feature extraction (see, for example, [Chodorow,

Byrd 1985] and [Alshawi 1987]), there are other features that motivated our work. First, general-purpose dictionaries have been standardized and culturally validated through their existence among native speakers for many years. The accepted definitions of concepts can serve well as a rich source of information when an intelligent agent is in search of a unbiased link between what's been understood and what needs to be understood. Because of their cultural acceptance and general-purpose use, one can envision a dictionary-based conceptual representation serving as a linkage to a variety of special-purpose representations.

Another dictionary feature attractive to knowledge representation is from the view point of natural language processing. The definitions are relatively compact, concise, and regular, making it easier and less ambiguous to automaticlly extract semantics than from ordinary natural language text. This aspect is well indicated by the work done by Ahlswede and Evens (1988), where they automatically extracted information about lexical-semantic relations existing in adjective definitions from a machine-readable dictionary.

Not only do there exist machine-readable versions of general-purpose dictionaries, but also some of them contain valuable information not found in their printed form. For example, the Longman Dictionary of Contemporary English (LDOCE) contains two kinds of special codes in its machine-readable version. One is the semantic codes that provide information on semantic markers and selectional restrictions and, separately, identify subject fields for particular senses of words in the dictionary. The subject codes were used by Walker and Amsler (1986) to develop a procedure to disambiguate the appropriate sense of words and phrases in a given context. The other kind is a system of grammatical codes that describes a particular pattern of behavior of a word. This information was used in constructing a large lexicon for natural language processing [Boguraev and Briscoe 1987]. We opt for LDOCE, at least initially, in our project because of the aforementioned properties and the additional feature that a defining vocabulary of 2000 words has been chosen carefully through a study of frequency lists of English words and only the most central sense of the words have been used in definitions [Procter 1978].

In our context, there are at least two justifications for using a machine-readable dictionary. First, some of the features such as the regularity of definitions and the availability of grammatical codes facilitate the process of automatically extracting and representing domain-independent knowledge. Second, the nature of a dictionary lends itself to the browsing application. Unlike a computer program designed to perform well in a specific domain, human browsers with common sense and a broad base of knowledge would benefit most from a conceptual representation whose content is as rich and comprehensive as a dictionary. In fact, knowledge found in the dictionary has been culturally validated and is expected to be blended well with the knowledge possessed by browsers of the culture. Furthermore, this approach of using a dictionary as a knowledge source complements other approaches to constructing a domain-dependent explorable vocabulary to be used in identifying and expanding a user's query [Myaeng and

Korfhage 1987] in information retrieval. The effectiveness of thesauri containing lexical-semantic relations, although not constructed from a dictionary, has been reported in the literature [Wang et al. 1985].

ROLE OF CONCEPTUAL GRAPHS

There are requirements of a good representational language for semantics. As in [Woods 1975],

- It should represent any particular interpretation of a sentence precisely, formally, and unambiguously.

- There must be an algorithm or procedure for translating the original sentence into this representation.

- There must be algorithms which can make use of the representation for the subsequent inferences and deductions that the human or machine must perform on them.

It appears at first that a representational language to be used by human browsers would not need all the power since humans are obviously able to do inferencing in a rather informal system and good at understanding things that might appear ambiguous to a computer program. Since we are envisioning a tool, not just a database, for human browsers in a variety of situations, however, the target representational language should be amenable to all the requirements desired by an intelligent computer system. The position of the tool in a spectrum of inferencing capability will vary depending on the degrees of autonomy of the tool and of human involvement in a particular man-machine interaction environment.

There are a number of reasons why the conceptual graph developed by Sowa (1984) was chosen as the representational language. As alluded to in Quillian's original study (1965) of semantic networks, the use of concept nodes and links seems to be the most natural way of representing the semantics of English words. Although slightly different, the notation of the conceptual graph with concept nodes and relation nodes is a simple variant of that of the semantic networks. This, in a sense, provides a psychological basis for the use of conceptual graphs since the primary users of the target representation of the dictionary are human browsers. Moreover, it gives the formalism and full representational power of the first-order logic and can handle higher order logic, making our work theoretically sound.

Another reason is related to the applicability of the conceptual graph framework for semantic interpretation of natural language sentences. With the existence of a lattice for concept types and canonical graphs in a lexicon and a syntactic parser, English sentences can be translated into corresponding conceptual graphs [Sowa 1986]. This capability seems particulary useful and applicable to our project since the aim is essentially to construct a conceptual graph for each concept type definition from its

counterpart in the dictionary. Following the formalism, in other words, lambda abstractions will be used to define concepts types for all the words and phrases to be included in the final representation.

Reasoning capability provided by the theory of conceptual graphs is also important to note. The operations on the lattice and the four formation rules, copy, restrict, join, and simplify, allow for the construction of a conceptual graph for an English sentence; canonical graphs for individual words occurring in the sentence are used as building blocks and combined by a series of operations to form a larger conceptual graph. Two other operations, type expansion and type contraction, also seem useful in manipulating graphs at a more practical, application-oriented level. In assisting browsers, for instance, the expansion operation would give a "zoom-in" effect when applied to a node in a definition conceptual graph. Although expensive, the contraction operation would achieve a "zoom-out" effect when a subgraph is identifed to be equivalent to a graph defining a type and replaced by its corresponding node. All these operations are expected to contribute to the ease of transforming individual conceptual graphs to actual presentation in a browsing tool where some sort of inferencing has to take place.

In processing the LDOCE, what should be available a priori are the following: a set of canonical graphs for the defining vocabulary (approximately 2000 words), a concept type hierarchy (lattice) for the defining vocabulary, and a set of relations with their canonical graphs. We expect that our linguistic study will play an essential role in generating the hierarchy and canonical graphs and identifying the relations. Research reported in the literature ([Chodorow and Byrd 1985], [Amsler 1980], [Ahlswede and Evens 1988]) shows the feasibility although they are not directly related to the theory of conceptual graphs. The output of the dictionary process will be a set of conceptual graphs corresponding to the type definitions. What has not been decided in detail is the process itself. One possible extreme is to follow the approach of Sowa's semantic interpreter [Sowa 1986], limiting the defining formulas to the generation of the grammar and input such as the canonical graphs and the type hierarchy. The other extreme is to use the defining formulas only, without even generating canonical graphs, to extract relational information to be represented subsequently in a conceptual graph form. How these two extremes can be blended is a question that we are currently pursuing.

CONCLUSION

The use of conceptual graphs for the purpose of extracting and representing relational information from dictionary definitions seems promising. Especially with the LDOCE, Sowa's semantic interpreter approach can be applied elegantly in conjunction with defining formulas to be generated from our linguistic study. In deciding how and where to use the defining formulas, we must consider the trade-off between computational tractability and expressiveness of the resulting representation. Although our primary goal is to construct explorable vocabulary for

human browsers, we are, in a sense, striving to automatically generate type definitions for a large lexicon that can be used in a variety of applications. The resulting conceptual graphs will serve as type definitions, not just schemata, since the source of information is from a culturally validated body of knowledge.

REFERENCES

Ahlswede, Thomas and Evens, Martha (1988). Generating a relational lexicon from a machine-readable dictionary. International Journal of Lexicography, 1 (3), 214-237.

Alshawi, H. (1987). Processing dictionary definitions with phrasal pattern hierarchies. Computational Linguistics, 13 (3-4), 195-202.

Amsler, R. A. (1980). The Structure of the Merriam-Webster Pocket Dictionary, Doctoral Dissertation, TR-164, University of Texas, Austin, TX.

Bawdem, David (1986). Information systems and the stimulation. Journal of Information Science 12, 203-216.

Boguraev, Bran and Briscoe, Ted (1987). Large lexicons for natural language processing: utilizing the grammar coding system of LDOCE. Computational Linguistics 13 (3-4), 203-218.

Brachman, R. J. and Levesque, H. J. (1985). Readings in Knowledge Representation, Morgan Kaufmann, Los Altos, CA.

Campagnoni, F. R. and Ehrich, K. (1989). Information retrieval using a hypertext-based help system. In Proceedings of the 12th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, June, Cambridge, MA.

Chodorow, Martin and Byrd, Roy (1985). Extracting Semantic Hierarchies from a large on-line dictionary. Proceedings of the 23rd Annual Meeting of the ACL, Chicago, June, 299-304.

Cove, J. F. and Walsh, B. C. (1988). Online Text Retrieval via Browsing. Information Processing and Management, 24 (1), 31-37.

Evens, Martha (1988). Relational Models of the Lexicon: Representing Knowledge in Semantic Networks, Martha Evens (Ed.), Cambridge University Press, Cambridge.

Myaeng, S. and Korfhage, R. (1987). A concept version. In Methodologies for Intelligent Systems, Proceedings of the Second International Symposium, October, Charlotte, NC, 185-192.

O'Connor, B. (1988). Fostering Creativity: Enhancing the Browsing Environment. International Journal of Information Management, 8, 203-210.

Palay, A. J. and Fox, M. S. (1981). Browsing through databases. In Information Retrieval Research, Oddy, R. (Ed.), Butterworths, London.

Proctor, P. (Ed.) (1978). Longman Dictionary of Contemporary English, Longman Group, Harlow & London, England.

Quillian, R. (1966). Semantic Memory. Unpublished doctoral dissertation, Carnegie Institute of Technology. Abridged version: Word concepts: a theory and simulation of some basic semantic capabilities. Behavioral Science, 12, 410-430, 1967.

Sowa, John (1984). Conceptual Structures: Information Processing in Mind an Machine. Addison-Wesley Publishing Co., Reading, MA.

Sowa, John (1986). Implementing a semantic interpreter using conceptual graphs. IBM J. Res. Develop. 30 (1), 57-68.

Sowa, John (1988). Using a lexicon of canonical graphs in a semantic interpreter. In Relational Models of the Lexicon: Representing Knowledge in Semantic Networks, Martha Evens (Ed.), Cambridge Universiy Press, Cambridge.

Walker, Donald and Amsler Robert (1986). The use of machine-readable dictionaries in sublanguage analysis. In Analyzing Language in Restricted Domains: Sublanguage Descriptions and Processing, Grishman, R. and Kittredge R. (Eds.), Lawrence-Erlbaum Associates, Hillsdale, New Jersey..

Wang, Yih-Chen; Vandendorpe, James; and Evens, Martha (1985).
Relational thesauri in information retrieval. Journal of the American Society for Information Science, 36 (1), 15-27.

Woods, W. A. (1975). What's in a link: foundations for semantic networks. In Representation and Understanding: Studies in Cognitive Science, Bobrow, D. G. and Collins, A. M. (Eds.), Academic Press, New York.