QUESTION ANSWERING VIA CANONICAL VERBS AND SEMANTIC MODELS: PARSING TO CANONICAL VERB FORMS

Craig W. Thompson

January, 1973

Technical Report NL 11

Natural Language Research for CAI

Supported by

The National Science Foundation

Grant GJ 509X

The Department of Computer Sciences

and CAI Laboratory

The University of Texas

Austin, Texas 78712

ABSTRACT

In processing natural language, much of the inference problem disappears when a single meaning with very different surface forms can be mapped into a single conceptual structure. This paper describes a process whereby a set of verbs which are semantic paraphrases of one another can be mapped onto a single primitive "canonical verb" which describes the concept underlying the set of verbs.

A companion paper describes a STRIPS-like semantic modelling system which keeps track of the current state of the world, remembers past states, and makes changes in the current state of the world by interpreting the canonical semantic net representation of an input sentence.

A second companion paper describes a representation of discourse in the form of a semantic net and discusses regeneration of surface English from the discourse net.

Keywords: semantic net, grammar, deep case relations, paraphrase, canonical verb, syntactic paradigm, semantic paradigm.

QUESTION ANSWERING VIA CANONICAL VERBS AND SEMANTIC MODELS: PARSING TO CANONICAL VERB FORMS

Introduction:

Much recent work in computational linguistics has been concerned with representing factual material in the form of semantic networks for the purposes of answering questions, drawing conclusions, solving problems, etc. Much attention has been paid to the problem of representation and inference. In understanding natural language, a significant part of these problems is eliminated when identical meanings with vastly different surface forms can be represented by one conceptual structure. In order to create conceptual structures which uniquely and unambiguously represent the meaning of an utterance, it is necessary to establish primitive underlying actions and relations into which surface level verbs can be mapped. A verb can be thought of as an operation, its accompanying noun and prepositional phrases as arguments. Often, several verbs describe basically the same operation. For example, buy, sell, pay, cost, ... describe EXCHANGE; go, walk, run, drive, ... describe MOVE. The noun and prepositional phrase arguments of the primitive canonical verbs like EXCHANGE and MOVE are syntactically specified differently by different English verbs. The term "syntactic paradigm" has been used to describe the surface level correspondence between a verb's subject, object, and prepositional phrases. In this paper, the term "semantic paradigm" is used to describe the conceptual level correspondence between the syntactic subject, object and prepositional phrases and the canonical verb's arguments. For instance, in the semantic paradigm for the verb "sell", the syntactic subject is the conceptual "SELLER" argument of the canonical verb EXCHANGE; the syntactic indirect object is the "BUYER" argument; the syntactic direct object is the "THINGSOLD" argument. This paper describes the process of mapping simple English sentences into a form of semantic net in which the nodes are primitive word-sense meanings and the paths are arguments of the canonical verb.

Background:

In recent years many linguists have begun to represent sentences with attribute-value or semantic net representations instead of with the more traditional predicate calculus notations (see 7, 2, 10, 12, 14). It has been shown (13) that semantic network representations can not only offer simplicity of computational representation and easy readability, but also preserve meaning quite as precise as predicate calculus representations of quantification and scope of variables. For these reasons in this paper semantic structures of sentence meaning are represented in the form of networks of word sense nodes connected by a form of deep case relation.

Though various linguists have transformed sentences into semantic nets and regenerated well-formed, meaningful sentences from those semantic nets (see 7, 2, 10), their methods have been of limited generality. A few papers have represented a significant generalization of the linguist's transformational apparatus (16, 17). In this paper, Woods' approach is followed because it seems to best mirror the constituent structure of English sentences. Woods represents his English grammar in the form of a state transition network that is augmented with subnetwork subroutines and a series of conditions and structure building operations associated with each possible path. He demonstrates that the resulting augmented finite state transition network is an efficient and powerful device for both analyzing and generating natural language structures.

Several linguists have used deep case names such as AGENT, THEME, INSTRUMENT, SOURCE, GOAL ... to represent the relationships between the verb of a sentence and the noun and prepositional phrases that modify it (3, 12). Structures based on deep case relationships were intended to be language free unambiguous representations of meaning. But, in fact, such structures have been shown to bear more similarity to the surface properties of English than should exist (9). Recently some linguists have found that representing the common, generic concepts behind verbs facilitates paraphrase of conceptual structures without loss of information. The canonical verb TRANS was introduced (8) as a generic concept into which such words as "give" and "take" could be mapped, such that by specifying attributes of the cases of TRANS no information would be lost.

In this paper the ideas of semantic network representations of English meanings and of the augmented state transition network representation of transformations serve as background to the idea of a canonical verb. A simple algorithm is discussed for mapping an input sentence into canonical dependency structure. Some attention is given to the representation of questions in such structures and to the use of such structures for representing transitions in a STRIPS-like world model (4).

From Input Sentence to Semantic Net

The description of the mapping from input sentence to canonical semantic net will be developed in three sections: the first section outlines the form of the lexicon; the next describes an intermediate structure which consists of an ordered list of noun and prepositional phrase constituents; the third section shows how relationships are determined between the arguments of the canonical verb and the noun and prepositional phrase constituents.

The Lexicon: the lexicon contains syntactic and semantic information for each word in the English discourse to be processed. This information is used by the parsing grammar to translate an acceptable input string into an internal information structure.

Each lexical entry has certain syntactic and definitional properties associated with it. These properties are stored in lists of attribute-value pairs. Each lexical entry contains the attribute WORDCLASS on its property list. The property list associated with a noun contains singular, plural, singular possessive and plural possessive forms of the word, as well as a semantic attribute MKR, which contains superset properties like HUMAN, PHYSOBJ, PLACE, ... The property list associated with verbs contains infinitive, third person singular, past, ... forms of the verb, as well as a structural attribute SEM-PDGM which relates the syntactic surface structure of a simple sentence associated with that verb to the conceptual structure of the canonical verb associated with it. The form of the SEM-PDGM attribute is that of a list of triplets, each representing a constituent of the sentence. The first element of a triple is either a preposition for prepositional phrases or "OK" for noun phrases. The second element is a marker, like the MKR attribute of nouns, which identifies the nature of that constituent. The third element specified the name of the argument of the canonical verb that the constituent represents. The property list associated with interrogative pronouns and adverbs (QWORDS) contains a semantic attribute MKR analogous to that of nouns. Property lists for adjectives, adverbs, prepositions, conjunctions, quantifiers, determiners, etc., are not given here since this information is not relevant to building the conceptual structure of a sentence. Figure 1 shows lexical entries for a noun, a verb, and a QWORD. (see Figure 1) The arguments of the canonical verb EXCHANGE have the following meanings: A = SELLER, B = BUYER, C = THINGSOLD, D = VALUE EXCHANGED, E = DAY, F = PLACE, G = TIME OF DAY, H = WHY-REASON. In the semantic paradigm for the verb "buy", the arguments are ordered in the way they are expected to occur in an active sentence where "buy" is the main verb. As will be seen in the following sections, this ordering is by no means strict.

Intermediate Structure: The process whereby an English input string is recognized and translated into an internal constituent structure need not be described in any detail. The parsing grammar is a variant of the augmented finite state transition network described by Woods (17). The form of the intermediate translated structure will be described. The translated structures for a sentence input to the system are centered around the main verb of the sentence. The verbal constituent is the primary translated structure for the sentence. It contains the canonical verb and a modal relation specifying TENSE, MOOD, and CASE.

```
(L28    (WC VERB) (INF BUY) (SG3 BUYS) (PAST BOUGHT)
        (CANON-VB EXCHANGE)
        (SEM-PDGM  ((OK  (HUMAN ORGANIZATION)  B)  (OK (PHYSOBJ)  C)
                    (FROM (HUMAN ORGANIZATION · A)  (FOR (MONEY)   D)
                    (OK  (DAY)  E) (ON  (DAY)  E)  (AT  (PLACE)   F)
                    (IN  (PLACE) F)  (OK (TIMEDAY) G)  (OK (TIMEDAY)  G)
                    (IN  (TIMEDAY) G)  (OK (REASON) H)  (FOR (REASON)  H)   )) )
(L52    (WC N)  (SING MAN)  (PL MEN)  (S-POSS MAN'S)  (PL-POSS MEN'S)
        (MKR (HUMAN))  )
(L112   (WC QWORD) (PRINT-IMAGE WHEN)
        (MKR (TIMEDAY DAY))  )
```

<u>Lexical entries for a verb, noun, and interrogative pronoun</u>.

FIGURE 1

```
a)  INPUT SENTENCE:
            ON TUESDAY WHO DID THE OLD MAN BUY A CAR FROM.
b)  VERBAL CONSTITUENT:
            ((CANON-VB  EXCHANGE)
             (MODAL (TENSE PAST MOD INTEROG CASE AFFIRM)))
c)  NP-PP-QWORD CONSTITUENT LIST:
        ( (ON  (DAY)  (PI TUESDAY TOK L20 NBR S))
          (FROM (HUMAN) (PI WHO TOK L110))
          (OK (HUMAN) (PI MAN TOK L52 DET DEF NBR SING
                         MOD (PI OLD TOK L78 DEGREE POSITIVE)))
          (OK (PHYSOBJ) (PI CAR TOK L96 DET INDEF NBR SING)))
```

<u>Input Sentence and Parsed Sentence Structure</u>

<u>Before Semantic Analysis</u>.

FIGURE 2

Noun and prepositional phrases and QWORD constituents appear as triples in the order in which they are encountered. The triple representing a noun or prepositional phrase or a QWORD constituent has the following form:

(<PREP> | "OK" <MKR> <NP constituent of NP or PP> | <QWORD constituent>)

The first element is the preposition for a prepositional phrase or "OK" for a noun phrase. The second is the lexical marker MKR on the head noun of the NP or PP. Finally, either the NP constituent of NP or PP or a QWORD constituent appears. The NP constituent itself is a list of the following relations:

TOK:	a pointer to the lexical entry for the head noun
DET:	Definite or Indefinite determiner
NBR:	Singular or Plural
POSS:	a possessive modifier
MOD:	an adjectival modifier
QNTFR:	quantifier in a noun phrase
PI:	print image

QWORD constituents only have the TOK and PI relations. An input sentence and its translated structure are given in Figure 2.

Assigning Arguments: The final stage in the mapping from input sentence to canonical dependency structure consists of correlating the list of NP, PP and QWORD constituents with the semantic paradigm SEM-PDGM associated with the verb. Basically, for each NP or PP constituent, a search is made through the semantic paradigm for a matching preposition or "OK". When one is found, its MKR is matched against the NP or PP marker of the constituent. If the two lists of markers intersect and that argument is not already assigned, the argument letter (the third element of the SEM-PDGM triple) is associated with the constituent.

If a QWORD constituent is encountered, it is saved until all other constituents are matched. Then, a list of all left over arguments whose MKR's intersect with the QWORD's marker is returned.

In the example, after the constituent list from Figure 2 has been correlated with the semantic paradigm from Figure 1, Figure 3 shows the semantic net structure that is created for the sentence. This structure represents a question about a past instance of EXCHANGE. The seller A, the thingsold C, and the day E when the transaction occurred are known. The questioning constituent Q concerns the buyer B.

Paraphrase and Questions

As can be seen from the examples in Figure 4 and 5, this approach to analyzing input sentences yields two main dividends. First, as can be seen in Figure 4, by creating a canonical verb and a standard set of arguments, sentences which express the same meaning with different words can be mapped into identical structures. Thus sentences 4-a thru 4-d, which are obvious paraphrases, are represented by identical semantic nets.

Secondly, by treating QWORD constituents in almost the same way as NP and PP's are treated, it is possible to keep track of what canonical verb arguments are being asked about and whether those arguments are subject to limiting conditions. For example, to answer the question in Figure 5a, an instance of EXCHANGE must be found in which John was the

a) INPUT SENTENCE:

ON TUESDAY WHO DID THE OLD MAN BUY A CAR FROM.

b) SEMANTIC NET:

```
(C10      (CANON-VB EXCHANGE)
          (B (PI MAN TOK L52 DET DEF NBR SING
             MOD (PI OLD TOK L78 DEGREE POSITIVE))
          (C (PI CAR TOK L96 DET INDEF NBR SING))
          (E (PI TUESDAY TOK L20 NBR S))
          (Q (PI WHO TOK L110 ARG (B)))
          (MODAL (TENSE PAST MOOD INTERROG CASE AFFIRM)))
```

Input Sentence and its Semantic Net Structure in Canonical Form.

FIGURE 3.

a) BILL CHARGED JOHN $70 FOR THE BOAT.

b) FOR THE BOAT JOHN PAID BILL $70.

c) JOHN BOUGHT THE BOAT FOR $70 FROM BILL.

d) TO JOHN BILL SOLD THE BOAT FOR $70.

e)
```
(C5 (CANON-VB EXCHANGE)
    (A  (PI BILL . . . )
    (B  (PI JOHN . . . )
    (C  (PI BOAT . . . )
    (D  (PI $70 . . .  )
    (MODAL (PAST INDICATIVE AFFIRMATIVE)))
```

Sentences A, B, C, and D are Semantic Paraphrases, and each can be represented by Semantic Net E.

FIGURE 4

a) WHAT DID JOHN BUY FROM BILL.

```
(C8  (CANON-VB  EXCHANGE)
     (A  (PI BILL . . . ))
     (B  (PI JOHN . . . ))
     (Q  (PI WHAT ARG (C)))
     (MODAL (PAST INTEROG AFFIRM)))
```

b) WHAT MAN PAID BILL FOR THE CAR.

```
(C61  (CANON-VB EXCHANGE)
      (A  (PI BILL . . .
      (C  (PI CAR . . .
      (Q  (PI WHAT ARG (B)
          COND (PI MAN . . .)))
      (MODAL (PAST INTER AFFIRM)))
```

c) WHEN DID BILL SELL THE CAR.

```
(C20 (CANON-VB  EXCHANGE)
     (A  (PI BILL . . . ))
     (C  (PI CAR . . . ))
     (Q  (PI WHEN ARG (E F)))
     (MODAL (PAST INTEROG AFFIRM)))
```

d) WHAT DAY DID BILL SELL THE CAR

```
(C21  (CANON-VB EXCHANGE)
      (A  (PI BILL
      (C  (PI CAR
      (Q  (PI WHAT ARG (E)
          COND (PI DAY . . .)
      (MODAL (PAST INTEROG AFFIRM))
```

In representing questions, the QWORD constituent Q of the semantic net identifies which argument(s) of the canonical verb is in question and what conditions, if any, that argument must obey.

FIGURE 5

seller and Bill was the buyer; and from that EXCHANGE, the thing sold must be returned. In Figure 5b, a buyer who bought a car from Bill must be returned, with the condition that the buyer be a man. In Figure 5c, either the day or time of day of an exchange involving John selling the car must be returned. Finally, in Figure 5d only the day, not the time of day, of the same exchange must be retrieved.

Discussion and Conclusions

The value of the above approach to representing sentences lies in its simplicity and its ability to represent the meaning of a sentence with clarity. The semantic net data structures described are a natural way of representing connected word-concepts which make up a sentence. Such semantic nets are easily extended to connected discourse (1, 12, 14, 15).

It has been argued that certain information is lost when a sentence is mapped into canonical form. For example, if "John sold the car to Bill", then in some sense John is the initiator of the action. It may be argued that John is no more the initiator than Bill -- but rather, John and his actions are the subject of the discourse. Thus, a form of EXCHANGE in which John is foregrounded is chosen by the speaker. The sort of information which identifies syntactic subject, etc., is certainly important to thematic development and to anaphoric resolution (1, 3); and it must be maintained during the processing of multisentence discourse. Yet, it is not clear that it should appear at all in the final representation of the meaning of a discourse. The question of what information to retain in the semantic net of a connected discourse is a current area of research.

One approach to recording the meaning of a discourse is described in a companion paper (5). That paper described the set up and query language of a data base of binary relations, which makes use of the canonical semantic structures here presented to make changes in a current state of the world. Associated with each canonical verb are ADD and DELETE lists which operate as follows: if "John traded the car to Mary for a boat", then the relations (John own car) and (Mary own boat) are deleted and the relations (John own boat) and (Mary own car) are added to the model of the current state of the world. Such transitions are remembered and are retrievable by the system.

One continuing area of research is the extension of the syntactic type of sentence that the prototype system can parse. The system is currently limited to conjuncted simple, active sentences which describe concrete actions or states. It has proved quite successful having parsed several hundred sentences in an average time of .6 seconds per sentence. Presently the system is being refined to include passives and several types of embedded sentences. Adding passives to the parse only expands the realm of syntactic representations of a concept. But adding embedded sentences adds a new dimension to the "understanding" capabilities of the system. For instance, in an embedded sentence, one verb often serves to qualify the other. For example, in "John thought that Mary sold the car to Bill", whether or not the selling occurred is indeterminate. A question concerning the ownership of the car can only be answered tentatively.

The present system is written in GROPE, a graph processing language embedded in FORTRAN, and is currently in operation at the University of Texas.

Acknowledgements:

I thank Dr. Robert F. Simmons, Gary Hendrix and Jonathan Slocum for many discussions, many hours and many ideas.

REFERENCES

1. Baranofsky, S., "Some Heuristics for Automatic Detection and Resolution of Anaphora in Discourse, " Master's Thesis, The University of Texas, Austin, Department of Computer Sciences, January 1970.

2. Carbonnel, J. R., "Mixed Initiative Man-Computer Instructional Dialogues," BBN Report No. 1971, Bolt Beranik and Newman, Inc., May 1970.

3. Chafe, W. L., Meaning and the Structure of Language, University of Chicago Press, Chicago, Illinios, 1970.

4. Fikes, R. F. and Nilsson, N. J., "STRIPS: A New Approach to the Application of Theorem Proving to Problem Solving," Artificial Intelligence, II, 1971, 189-208.

5. Hendrix, G. G., "Question Answering via Canonical Verbs and Semantic Models: A Model of Textual Meaning," University of Texas, Department of Computer Sciences, Technical Report 12, January 1973.

6. Jacobs, R. A. and Rosenbaum, P. S., English Transformational Grammar, Blaisdell Publishing Co., Waltham, Mass., 1968.

7. Quillian, M. R., "The Teachable Language Comprehender," Comm. ACM, 12, 8 (August 1969), pp. 459-476.

8. Schank, R., "'Semantics' in Conceptual Analysis," Memo ACM-122, Computer Science Department, Stanford University, Stanford, Cal., May 1970.

9. Schank, R. C., Goldman, N., Rieger, C. J. and Riesbeck, C. K., "Primitive Concepts Underlying Verbs of Thought," Memo AIM-162, Computer Science Department, Stanford University, Stanford, Cal. February 1972.

10. Simmons, R. F., Burger, J. F., and Schwarcz, R. M., "A Computational Model of Verbal Understanding," Proc. AFIPS 1968, FJCC, Vol. 33, AFIPS Press, Montvale, N. J., pp 441-456.

11. Simmons, R. F., "Natural Language Question Answering Systems: 1969," Comm. ACM, 13, 1, January 1970. pp 15-30.

12. Simmons, R. F., "Some Semantics Structures for Representing English Meanings," University of Texas, Austin, Department of Computer Sciences, Preprint, November 1970.

13. Simmons, R. F., "Some Relations between Predicate Calculus and Semantic Net Representations of Discourse," University of Texas, Austin, Department of Computer Sciences, Preprint, November 1970.

14. Simmons, R. F. and Slocum, J. "Generating English Discourse from Semantic Networks," Comm. ACM, 15, 10, (October 1972) pp. 891-905.

15. Slocum, J., "Question Answering via Canonical Verbs and Semantic Models: Generating English for the Model," University of Texas, Austin, Department of Computer Sciences, Technical Report 13, January 1973.

16. Winograd, T., Understanding Natural Language, Academic Press, New York, 1972.

17. Woods, W. A., "Transition Network Grammars for Natural Language Analysis," Comm. ACM, 13, 10 (October 1970), pp. 591-606.