[*Mechanical Translation and Computational Linguistics*, vol.11, nos.1 and 2, March and June 1968]

Paraphrase Generation and Information Retrieval from Stored Text

by P. W. Culicover, IBM Boston Programming Center

First the notion "paraphrase" is defined, and then several different types of paraphrase are analyzed: transformational, attenuated, lexical, derivational, and real-world. Next, several different methods of retrieving information are discussed utilizing the notions of paraphrase defined previously. It is concluded that a combination keyword-keyphrase method would constitute the optimum procedure.

1. Introduction

This paper deals with the use of paraphrase relationships to effect the retrieval of stored English text and the information contained in stored English text. Throughout the discussion I will be dealing with the various options available in terms of desirability and practicability. My concern will be primarily with determining the optimum value of each of the following theoretical parameters: form of storage, form of input, method of matching, and form of response. More specifically, I shall be investigating the possibility of utilizing known paraphrase relationships to expedite the retrieval process and to reduce the quantity of linguistic processing involved in eliciting the proper responses from the data base.

Part 2 of this paper deals with paraphrases *in vacuo;* that is, it deals with the problem of isolating various well-formed types of paraphrase and the method of generating paraphrases. Part 3 is concerned with the general problem of retrieval from text and discusses and evaluates various logical possibilities in this direction. The uses of paraphrase relationships are evaluated with regard to the degree with which they affect the various possibilities discussed here.

2. The Meaning of "Paraphrase"

Paraphrase is generally considered to be a meaning-preserving relation. If we are dealing with text as a primarily linguistic form, then we may say that text *a* and text *b* are "paraphrases" of one another if text *a* and text *b* have the same "meaning." This is not particularly illuminating from the point of view of a discipline that can make no use of human intuition, namely, information or text retrieval, since "meaning" or "same in meaning" are undefined in the absence of a linguistically competent individual. Insofar as a mechanical device is linguistically naive, it is necessary to define in structural terms what "meaning" is, so that it may be evaluated by a mechanical procedure. Preparatory to this it will be necessary to define in as precise terms as possible various types of paraphrase. This is due to the fact that certain paraphrase relationships are inaccessible to a generalized recognition procedure. Hence it is desirable to isolate the accessible relationships from the inaccessible ones in order to remove from the investigation any problems which are to all extents and purposes insoluble.

TYPES OF PARAPHRASE

It is possible to define a number of special cases of paraphrase. This means, in effect, that there are certain cases of meaning equivalence which are definable in purely formal terms, with no consideration of the intrinsic meaning of the items under consideration.

Transformational Relationships

In linguistic theory instances of a formal relationship between synonymous linguistic elements (phrases or sentences) are called "transformational relationships" [9, 11, 13, 15, 24, 29]. For my purpose, a transformational relationship is one having the following properties: given a set of sentences or phrases which are transformationally related, (I) the grammatical relationships which obtain between the words of each member of the set, and (II) the words which form one member of the set and which have cognitive significance are the same as the words having cognitive significance which form any other member of the set.

Grammatical Relationship

The term "grammatical relationship" is defined as a structural condition on the sentence [13]. Given a set of grammatical relationships,

$$G_1, G_2 \ldots G_n,$$

we say that A bears the relationship G_i to B ($G_i(A,B)$) if and only if B dominates A (B > A):

B.
|
A

Each G_i is defined in terms of syntactic types X, Y, Z . . . so that $G_i(A,B)$ if and only if A is an X, B is a Y, and B> A. Given the primary grammatical relationships, we may define secondary relations; thus $G_j(A,C)$ if and only if $G_i(A,B)$ and $G_k(B,C)$ and A is an X, B is a Y, and C is a Z. The domination relationship need not hold between A and C in such cases.

Cognitive Significance

The term "cognitive significance" is an expression which refers to the intrinsic meaning of the word. For all intents and purposes the cognitive significance of a word is a linguistic primitive. Insofar as mechanical manipulations of linguistic structures are concerned, the meaning of the word is the word itself. Some examples which illustrate Condition I on transformational relationships are as follows:

The dog bit the postman. (1)

The postman bit the dog. (2)

The postman was bitten by the dog. (3)

Defining the Grammatical Relationship

Let us define three primary grammatical relationships in terms of the syntactic types "S," "NP," "VP," and "V." G_1 = "subject of," G_2 = "object of," G_3 = "predicate of." $G_1(A,B)$ if and only if A is an "NP," B is an "S," and B>A. $G_2(A,B)$ if and only if A is an "NP," B is a "VP," and B>A. $G_3(A,B)$ if and only if A is a "VP," B is an "S," and B>A. A secondary grammatical relationship is "object of S." If we call this relationship G_4, then $G_4(A,B)$ occurs if and only if $G_2(A,B)$ and $G_3(B,C)$, where A is an NP, B is a VP, C is an S, and C>B>A. Parsing (1), (2), and (3) we get

$$S[_{NP}[\text{The dog}]_{VP}[_{V}[\text{bit}]_{NP}[\text{the postman}]]]. \quad (4)$$

$$S[\text{The postman}]_{VP}[_{V}[\text{bit}]_{NP}[\text{the dog}]]]. \quad (5)$$

$$S[_{NP}[\text{The postman}]_{VP}[_{V}[\text{was bitten}]?[\text{by the dog}]]]. \quad (6)$$

Although (4) and (5) contain the same words, we see that they fail to meet Condition I, since "the dog" is the "subject of (4)" and the "object of (5)," while "the postman" is the "object of (4)" and the "subject of (5)." A similar, yet more complex, situation obtains between (5) and (6), since we have yet to define the phrase "by the dog."

Since, however, it is known by speakers of English that (4) and (6) represent synonymous sentences, this fact must be represented in some way. A level of representation is created on which it is noted that the "subject of (4)" and the "subject of (6)" are the same item. This level of representation indicates the "logical" or "deep" grammatical relations obtaining between the elements of the sentence, so that it is now possible to demonstrate that the "deep object of (6)" is the same as the "object of (4)." On the other hand, the "subject of (5)" is the "deep object of (4)" and the "deep object of (6)," and so on.

In effect, the syntactic grammatical relationships do not always reflect the semantic relationships. In order to identify the notions indicated above, let us define the semantic relationships in terms of *(a)* the surface syntactic relationships, and *(b)* the relevant syntactic characteristics of the sentence. For example, let "agent of" be represented as M1 and "proposition" as M_2. The relationships which then hold are as follows: A is the agent of $B(M_1(A,B))$, if A is an NP; B is a proposition represented by S ($M_2(B,S)$), and A is the subject of S if S is an active sentence, or A is in the by-phrase of S, if S is a passive sentence. In formal terms, $M_1(A,B)$ if $[M_2(B,S)]$ and [G1(A,S) if S is active] or $[G_4(A,S)$ if S is passive]. Similarly, representing "goal of" as M_3, we have $M_3(A,B)$ if $[M_2(B,S)$j and $[G_2(A,S)$ if S is active] or [G1(A,S) if S is passive].

Illustrating Condition II

Condition II may be illustrated by similar examples. Consider the following sentences:

$$S[_{NP}[\text{The postman}]_{NP}[_{VP}[\text{bit}]_{NP}[\text{the dog}]]]. \quad (7)$$

$$S[_{NP}[\text{The dog}]_{VP}[_{V}[\text{was bitten}]_{AGT}[\text{by the postman}]]]. \quad (8)$$

$$S[_{NP}[\text{The dog}]_{VP}[_{V}[\text{was bitten}]?[\text{in the leg}]]]. \quad (9)$$

We say that (7) and (8) meet Condition II because neither contains words of cognitive significance that the other does not contain. In the case of (8), the words "was" and "by" are of "grammatical significance," that is, they signal the particular syntactic form of the sentence, namely, that it is a passive sentence. However, they are devoid of meaning in the sense in which meaning is defined by such terms as "reference," "activity." "means," "manner," etc. Sentence (9) fails to contain "the postman," which has cognitive significance, and (8) fails to contain "in the leg," which also has cognitive significance. Hence (9) fails to meet Condition II with respect to (7) or (8).

Some Transformations

Given the above definitions and conditions we see that they in turn define a class of transformational relationships which we list below in part [9, 11, 29, 42]. It should be pointed out that our conditions are in fact fairly loose and permit a wide range of structures to claim transformational relationship.

a) Declarative, yes-no question: The dog bit the postman; did the dog bite the postman?
b) Extraposition, nonextraposition: It strikes me as funny that the dog bit the postman; that the dog bit the postman strikes me as funny.
c) Active, passive: The dog bit the postman; the postman was bitten by the dog.
d) Determiner, relative clause: The dog bit the angry postman; the dog bit the postman who was angry.
e) Adverb, final; adverb, not final: The dog bit the postman yesterday; yesterday the dog bit the postman.

Some Apparent Exceptions to Condition II

A second level of transformational relationship occurs in cases where Condition II is not met but where what would be needed in order for Condition II to be met is predictable on a formal basis. Such cases appear where ellipsis or some form of pronominalization has taken place. Some examples of these phenomena are given below:

Pronominalization.—Consider the following sentences:

If Mary wants a book she'll take one from the library. (10)

If Mary_i wants a book Mary_j will take one from the library. (11)

In sentence (10) "she" refers to Mary if in sentence (11) j = i. Since, however, the most normal interpretation of (11) is that j = i, it can be concluded that whenever j = i, the second noun phrase must be represented by a pronoun. Thus the cognitive significance of "she" in (10) is the same as the cognitive significance of the corresponding NP in a position where such an NP cannot exist. Although (10) is not transformationally related to an occurring sentence, it and (11), where j = i, are semantically identical.

Wh questions, declaratives.—We note that the sentences below also fail to meet Condition II:

What bit the postman? (12a)

The dog bit the postman. (12b)

Where did you sleep? (13a)

You slept in a bed. (13b)

Why did the dog bite the postman? (14a)

The dog bit the postman because the postman kicked him. (14b)

The (a) sentences are identical to the (b) sentences except for the fact that where the former contain Wh-words (what, where, why, etc.) the latter contain phrases with cognitive significance. However, since the Wh-words are in the syntactic position of a phrase with cognitive significance, and since they do not add any meaning to the sentences which contain them, the failure of the above examples to meet Condition II is a special type of failure. The Wh-words represent a phrase with cognitive significance, with the additional feature that they indicate that a question is being asked about the nature of such a phrase. As we have seen, a question does not alter the cognitive significance of a sentence, and so these cases may be thought of as a combination of pronominalization and a question, where the characteristics of the pronoun are being questioned.

Ellipsis.—A similar type of pronominal relationship is found to obtain between sentences such as those below:

The dog bit the postman but the cat won't. (15a)

The dog bit the postman but the cat won't bite the postman. (15b)

In sentences like (15a) and (15b) we see again that Condition II does not hold, since (15b) contains a phrase of cognitive significance which (15a) does not contain. However, we observe that in no case can a sentence have a modal verb (will, can, could, shall, may, must, etc.) without having a main verb as well and still be grammatical. Since (15a) is grammatical, even though it contains the phrase "the cat won't," it must be the case that it is possible to determine what the missing verb phrase is. Clearly the missing verb phrase is "bite the postman," and in general when a modal stands along the missing verb phrase is identical to the one in the preceding sentence.

Violations of Condition II

A third level of syntactic relationship exists when Condition II is not met but when a variant of Condition II called Condition II' is met.

Condition II': Given sentences A and B, every word in A is in B, but there are words in B which are not in A.

If Conditions I and II' are met by sentences A and B then we shall say that "A is an attenuated paraphrase of B." Some examples of attenuated paraphrase are given in (16) below:

The dog bit the postman. (16a)

The dog bit the postman on the hand. (16b)

The dog with fangs bit the postman on the hand. (16c)

The relationships which obtain between these sentences is one of "entailment." That is, if (16c) is true, then (16b) and (16a) are true. If (16b) is true, (16a) is true, but (16c) need not be true. Similarly, if (16a) is true, then (16c) and (16b) may, but need not be, true. Let us call "with fangs" and "on the hand" "qualifying phrases." We observe that a sentence which has a qualifying phrase entails any sentence which is identical to it but for the fact that it lacks a qualifying phrase in that position. Conversely, either of two such sentences will satisfy as answers to a question which does not refer to the qualifying phrase. Thus, to the question "Did the dog bite the postman?" all of (16a), (16b), and (16c) are correct answers. However, to the question "Did a dog bite the postman on the hand," only (16b) and (16c) are correct answers, since (16a) has no information pertaining to where the postman was bitten. Similarly only (16c) can be a satisfactory answer to the questions "Did the dog with fangs bite the postman on the hand" or "Did the dog with fangs bite the postman," assuming (16a-16c) constitute the entire extent of information which we possess about the event.

Since Condition II is met, it doesn't matter whether the question is asked in the active or in the passive, or if the information is maintained in the active or in the passive.

Lexical Paraphrase

It is possible to identify an entirely different type of paraphrase, which we shall call "lexical paraphrase." Lexical paraphrases can be differentiated into two categories on formal grounds: "word" and "idiomatic." The latter category may be further subdivided into "continuous" and "discontinuous."

Synonyms and entailment [27, 28].—Word paraphrases are commonly called "synonyms." In this case we wish to refer only to classes of synonyms whose members contain one word only. Innumerable examples of such paraphrases can be found in any thesaurus, and we shall not trouble to list any here. Of greater interest is the relationship between sentences which meet Condition I but which fail Conditions II and II'. Let us define a subclass of such sentences by mean of Condition III.

Condition III: Given sentences A and B, all the words in A which have cognitive significance are either identical
to the words in B which have cognitive significance or are "word" paraphrases of words in sentence B.

If two sentences meet Conditions I and III, then we say that they are "exact paraphrases" of one another. Such an example is:

The dog bit the irate postman. (17a)

The dog bit the angry postman. (17b)

Depending on our definition of paraphrase, we may define as rich or as sparse a field of exact paraphrases as we desire. Consider, for example, the words "box," "hatbox," "ashtray," and "container." The same relationship of entailment discussed in Violations of Condition II above can be shown to obtain between certain pairs of the words above, precisely as a result of the degrees of qualification which each represents. If we place each word into the frame "This thing is a . . ." the entailment relationship becomes very clear.

This thing is a box. (18a)

This thing is a hatbox. (18b)

This thing is an ashtray. (18c)

This thing is a container. (18d)

If (18a) is true, then (18d) is true. If (18b) is true, then (18a) and (18d) are true. If (18c) is true, then (18d) is true. If (18d) is true, then (18a)-(18c) may be true but need not be. Similarly, certain sentences will be satisfactory answers to certain questions, depending on whether the sentence entails the declarative counterpart of the question.

In the case of exact paraphrases we may therefore define an "entailed paraphrase" when one sentence entails the other, and a "full paraphrase" when each sentence entails the other.

Idiomatic paraphrases.—These occur when one or both members of the relation consist of more then one word. Some examples are: enter-go in, discover-find out; return-go back; fall asleep-doze off; look for-look up (information, etc.). Semantically they may be treated in precisely the same way as word paraphrases are. Discontinuous idioms are idioms which contain, rather than concatenate with, other elements in a syntactic structure. For example, "John lost his way," "Mary lost her way" but not "Mary lost John's way." Most complex are those where the variable element is not predictable from the rest of the sentence: "Bill sold X down the river," "John looked X up."

Derivational Paraphrase

Another completely different type of paraphrase is "morphological" or "derivational." The English language contains a number of very productive rules for deriving one lexical category from another, or for deriving new members of a lexical category from members of the same category combined with members of other categories. By far the most productive of these processes are subsumed under the name "nominalization" [12, 34]. Consider the following examples:

orient—orientation
circumvent—circumvention
instruct—instruction
deceive—deception (19)
instigate—instigation
compute—computation
destroy—destruction

believe—believer
ride—rider
compute—computer
instruct—instructor (20)
write—writer
destroy—destroyer
give—donor
receive—recipient

Observe that the relationship between the verbs and the nouns in (19) is "N_i = the act of V_i-ing," while the relationship involved in (20) is "N_i — one who V_i's." However, "computer" more frequently means "machine which computes," and many of the nominalizations of the type in (19) often have a passive connotation, as in "the building's destruction" versus "the landlord's destruction of the building." Similar relationships obtain between verbs and adjectives, thus "believe-believable," "like-likable," "permit-permissible," etc. An examination of such pairs also shows no constant relationship between the type of morphological relationship and the semantic relationship between paired elements. Because syntactic types are related in processes of nominalization, adjectivization, etc., none of the conditions discussed above can be met by sentences which contain these pairs. It is still possible, however, to isolate certain frequently occurring morphological relationships. For each such relationship we may state a Condition (x) such

that, if sentences A and B meet Condition (x), the sentences are "morphological paraphrases of type (x)" of one another. For example:

Condition A: Given sentences A and B, sentence A contains an agent nominalization (writer, computer, etc.), and sentence B contains the noun phrase

$$\text{the} \left\{ \begin{array}{l} \text{one who} \\ \\ \text{thing which} \end{array} \right\} V_i\text{'s}$$

Condition B: Given sentences A and B, sentence A contains a factive nominalization NP_1's N_j of NP_2 (the landlord's destruction of the building), and sentence B contains the noun phrase "(the fact) that NP_1 V_j-ed NP_2."

Condition B_2: Given sentences A and B, sentence A contains a factive nominalization NP_1's N_j (by NP_2) (the building's destruction by the landlord), and sentence B contains the noun phrase "(the fact) that NP_1 was V_j-ed (by NP_2)." If sentences A and B meet one of Condition (x), and one of Conditions II and III, then they are "morphological paraphrases."

"Real-World" Paraphrase

The last type of paraphrase which we shall consider is called "real-world paraphrase." It is this type which is the most inaccessible to general mechanical treatment because it is independent of linguistic structure. Real-world paraphrase may be divided into two types: logical paraphrase and informational paraphrase.

The first is characterized by the use of mathematics and rules of inference. For example, the (a) sentences below are paraphrases of the (b) sentences:

New York is larger than any other city except for Tokyo. (21a)

New York is the second largest city in the world, and Tokyo is the largest. (21b)

John has a car and his wife Mary has a car. (22a)

John and his wife Mary have two cars between them. (22b)

It is possible this type of logical paraphrase may be accessible to highly sophisticated techniques of data manipulation and paraphrase generation, although such techniques would seem to be far beyond the range of present-day capabilities.

The second type of real-world paraphrase, informational paraphrase, is characterized by a highly refined knowledge of the historical, sociological, cultural, and scientific structure of society. Such knowledge in its entirety can only be manipulated and utilized by a human being. For example, consider the following sentences:

The President of France laid a wreath on Marshal Petain's grave. (23a)

Charles de Gaulle laid a wreath on Marshal Petain's grave. (23b)

These two sentences are exact paraphrases of one another if and only if it is the case that the president of France is Charles de Gaulle. More sophisticated examples might be constructed, but this suffices to suggest that identifying reference and co-reference in the absence of linguistic clues, such as stress and pronominalizations, is hopeless for a general mechanical procedure.

The Generation of Paraphrases

Having isolated significant classes of paraphrases it is now a fairly direct matter to translate this into a method for generating paraphrases. We shall consider each type in turn and sketch out the method in brief.

Transformational relationships.—If the sentence has not undergone a particular transformation, apply the transformation. If the sentence has undergone the transformation, apply the reverse of the transformation.

Attenuated paraphrase.—Identify the qualifier and generate sentences which fail to contain one or more of the qualifiers. The full sentence is an answer to any questioned attenuated paraphrase.

Entailment word or idiomatic paraphrases.—Substitute for the word all words which it entails. The entailing sentence is an answer to any questioned entailed paraphrase.

Morphological paraphrase .—Substitute for the nominalization, etc., the phrase which paraphrases it. If a phrase is recognized, substitute the nominalization, etc., which represents it.

3. Information Retrieval from Texts

Let us now turn to the problem of retrieving information from stored texts. As mentioned in Part 1, we will be concerned with the form of the various parameters involved: storage, input, matching, and response. Needless to say, several of these are functions of the others: Once the storage and input format have been selected, the form of matching follows from it; if the form of matching and input are selected, the format of storage follows, and so on.

RELEVANT PARAMETERS

The various parameters should have the following values relative to the theoretically envisionable "worst possible case."

Matching

The matching process should be as fast as possible, and the time taken to find a match should be reduced to the minimum. These two criteria are not equivalent, since a very fast process might conceivably be called upon to match highly complex items.

Response

The response should contain all the information desired and no more.

Input

The input should have to undergo the minimal amount of processing in order to elicit the desired response from storage, but it should be specific enough to guarantee that overly large amounts of unelicited response are not generated. The input should be of a form which will expedite the matching process.

Storage

The storage should also be of a form which will expedite the matching process. The amount of processing required to identify what is stored should be minimal, since any processing whatever of large amounts of stored text would require a considerable investment in time.

TYPES OF PROCESSING

The most significant variable to be considered in this discussion is the degree of processing. The following types of processing are given in order of increasing complexity: keyword, keyphrase, keysentence, deepstructure. To each type we may also apply paraphrase generation.

Keyword

The keyword method identifies a likely keyword in the input sentence, matches the keyword with every occurrence of the same word in the unstructured text, and delivers as a response every sentence in the unstructured text which contains the keyword. There are several possible refinements of such a technique available.

Syntactic analysis.—A minimal syntactic analysis may be performed in the input sentence to insure that the keyword will always bear a particular relationship, in grammatical terms, to the input sentence. An equivalent, but alternative goal, is to perform a minimal syntactic analysis on the input sentence to insure that the keyword does *not* bear a particular relationship to the input sentence. For example, in the question "Were any suspension bridges built before the First World War?" the desired keyword "bridge" (or "suspension bridge") is the "goal," while the time adverbial "before the First World War" contains a noun phrase "the First World War" which is not the desired keyword and which does not contain a desired keyword.

Paraphrases.— Another refinement would be to generate all those words which the keyword entails. For example, any text which constitutes a satisfactory answer to the order "Tell me about the manufacture of containers for vegetables" also constitutes a satisfactory answer to the order "Tell me about the manufacture of boxes for vegetables," since a box is necessarily a container by definition, although the converse is not true.

This type of paraphrase generation, which applies equally well to more sophisticated methods of processing, would require the development of a highly structured "lexicon." Each word in the lexicon would be indexed in some way which would reflect the entailment relationship it has to other lexical entries. A simple example which illustrates the method by which such an indexing could be established is as follows:

We first construct a tree which schematically represents the entailment relationship (see fig. 1).

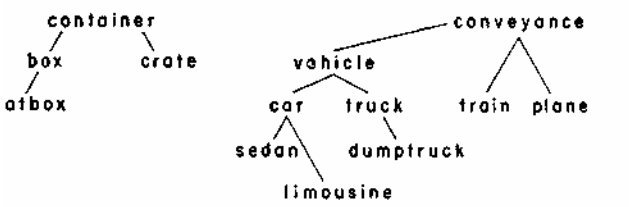


FIG. 1.—Trees representing the entailment relationship

Each topmost entry is then numbered as follows:

container: $a_1.0$ conveyance: $a_2.0$

Items one level down are indexed with one decimal place and the integer which represents the tree in question.

box: $a_1.b_1$ vehicle: $a_2.b_1$
crate: $a_1.b_2$

A similar process applies to the remaining elements, so that the final result of indexing is as follows:

container:	$a_1.0$	conveyance:	$a_2.0$
box:	$a_1.b_1$	vehicle:	$a_2.b_1$
crate:	$a_1.b_2$	train:	$a_2.b_2$
hatbox:	$a_1.b_1c_1$	plane:	$a_2.b_3$
		car:	$a_2.b_1c_1$
		sedan:	$a_2.b_1c_1d_1$
		truck:	$a_2.b_1c_2$
		limousine:	$a_2.b_1c_2d_2$

Given the index of the keyword in question, it is a simple matter to identify all those items which entail it. If the keyword is "vehicle," one finds all those items on the tree below it by generating the indices $a_2.b_10$, $a_2.b_1c_1$, $a_2.b_1c_2$, $a_2.b_1c_3$, $a_2.b_1c_4$. . . . Each of these indices may be used to generate items lower down on the tree by the same process: $a_2.b_1c_10$, $a_2.b_1c_1d_1$, $a_2.b_1c_1d_2$, $a_2.b_1c_1d_3$. . . . The extent to which paraphrases are generated by such a process may be arbitrarily limited.

One theoretical difficulty with a procedure of this type is that of structures being found which converge, so that a single item would have two indices (see, e.g., fig. 2).

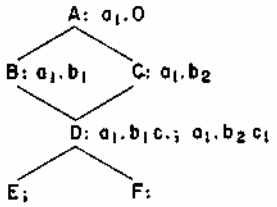


FIG. 2.—Entailment relationships which converge would raise a difficulty.

If such a case arose where the point of convergence itself branched, we would be faced with the problem of generating two sets of indices for each of the lower items (E and F) so that they could be reached by the path through B or by the path through C. This difficulty can be avoided by limiting the generation of paraphrases to the first level down from the item initially selected. Such a situation is of theoretical interest, since no examples of this involving word paraphrase alone have been found to date, although we do not eliminate the possibility that they may exist.

A similar problem exists when structures are found of the type shown in figure 3.

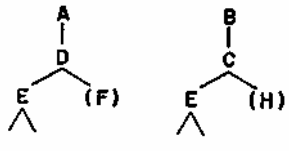


FIG. 3.—Entailment relationships involving an ambiguous word.

This represents a case where a word is ambiguous, for example, ball-toy-amusement and ball-affair-event. These are real cases where the word must have more than one index so that its different meaning may be identified.

It might be pointed out here that to limit the paraphrase generation to one level down from the selected item is not completely arbitrary. For example, if the question asked was "Tell me about the manufacture of containers," a reasonably specific answer might refer to boxes, cartons, crates, etc., but not to hatboxes, cigarette cases, or garbage cans.

Keyphrase

In Syntactic Analysis (above), we mentioned the possibility of doing a syntactic analysis of the input sentence. It should be obvious that having performed this analysis one could expect to use profitably the information gleaned from it. The keyword method does not take full advantage of this type of analysis, since it selects a single word from the major constituent selected by the analysis. In fact, having performed the analysis we can utilize the information gained from it to generate "keyphrases" which have the virtue of being considerably more specific than keywords, thus reducing the number of undesired responses.

An unfortunate consequence of using keyphrases is that the chances of matching the keyphrase with an identical phrase in the text are considerably lower than the chances of matching a keyword with an identical word in the text. Thus a phrase "the process of manufacturing steel" will not match with any part of the sentence "Basically steel is made by. . . ."

The advantage of the keyphrase method, as already pointed out, is the greater degree of specification it affords. The keyphrase method may be combined with the keyword method to increase the chances of finding a match. One straightforward method of doing this would be to select from the keyphrase the word with the greatest specification, that is, the word with the longest decimal index, in this case presumably "steel." This word is then matched with identical occurrences in the text. As a response to the question we select any text under a predetermined length which contains both the keyword (or its paraphrases) and occurrences of the other words in the keyphrase (or their paraphrases).

This process, although admittedly more cumbersome than the keyword method alone, has two other major advantages. First, it is not necessary to generate structural paraphrases of the keyphrase (e.g., "John's computer," "the computer of John's," "the computer that John had") since the syntactic relationship would be irrelevant to such a procedure; the relative order of the words in the keyphrase plays no role. Second, it takes into consideration the fact that the words which make up the input sentence may be strewn about the text quite far from the "target" word selected from the keyphrase. It should also be mentioned that this method, like the keyword method alone, requires no processing whatsoever of the stored text.

Keysentence and Deepstructure

A third method available is called "keysentence." Keysentence, unlike keyword or keyphrase, requires not only

processing of the input sentence with a moderately sophisticated recognition device but also requires processing of the stored text, which is a distinct disadvantage. In the case of keysentence, the syntactic analysis is not used only as a means of eliminating unlikely keywords or keyphrases but is used also to limit the field of possible responses by identifying the nature of the question or order.

The keysentence method consists primarily of reducing both the input and the stored text to a set of pointers to the identifiable semantic categories of the proposition. The deep subject is labeled "Agent," the verb "Action," the deep object "Goal," the adverbials "Time," "Place," "Manner," and so on. If a question is being asked about one of these categories, then the question word is labeled "Q-Agent," "Q-Verb," "Q-Goal," "Q-Time," etc. The keysentence is matched with any sentence in the stored text to which it is identical, of which it is an attenuated paraphrase, or which differs from it only by a labeled category instead of the Q-labeled category in the keysentence.

The major advantage of this method is that responses would be elicited which precisely corresponded to the input sentence. It can be seen without much investigation that most sentences in contiguous text will never be questioned. Hence, the processing of every sentence in the stored text would be useless, as well as impractical, uneconomical, and time-consuming.

Deepstructure Method

A similar argument can be made against the "deepstructure" method, which entails a complete syntactic analysis of every sentence in the stored text as well as every input sentence. No benefit can be seen to result from structurally identifying every word which appears in a stored text, since very rarely, if at all, will a query be so specific as to require such a considerable degree of detail. Furthermore, deepstructures are so much larger than strings, being two-dimensional rather than one-dimensional, that any envisionable storage capabilities would be greatly exceeded by reasonable quantities of text. Matching problems would also be expected to arise if such a technique were ever seriously implemented.

RESPONSE

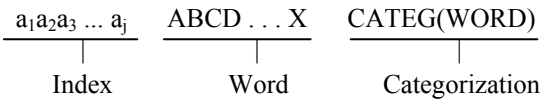
One parameter which I have discussed very little is "response." Ideally the desirable response is the one which exactly answers the question asked. However, since a stored body of text cannot be safely relied upon to contain all sentences which are possible answers to all questions, one must aim at somewhere below the ideal situation. The keyword method will return responses consisting of all those sentences in the text which contain at least one occurrence of the keyword. The keysentence and deepstructure techniques would be able to return only single sentences as responses and would therefore be insensitive to cases where the proper response was in paragraph form ("paragraph" meaning two or more sentences). As we have seen, however, the combination keyword-keyphrase method searches for occurrences of the words in the keyphrase that all fall within a limited segment of the text. With such a method it would be feasible to experimentally vary the maximum segment length to ascertain the optimum length—response ratio. This is to say, the longer the segment, the more unelicited information will appear in the response; the shorter the segment, the more elicited information will not appear in the response. It would also be possible, it would seem, to vary the length of the relevant segment according to the number of keywords in the keyphrase. The greater the number of keywords, the greater the number of sentences which would be allowed to constitute a proper response.

4. Keyword-Keyphrase Structure

This last section discusses the outlines of a possible implementation of the keyword-keyphrase method.

THE DICTIONARY

The dictionary entries as presently envisioned would contain three well-defined segments: the word, the index, and the categorization. The format of a typical dictionary would be as shown below.



DICTIONARY LOOK UP

Dictionary look up (DLU) matches the word in the input sentence with a word in the dictionary and replaces the word in the sentence with the corresponding categorization.

PARAPHRASE GENERATION

In generating paraphrases, paraphrase look up (PLU) matches the keyword with a word in the dictionary, generates indices used on the index of the word, and looks up the words corresponding to the generated indices.

Syntactic Analysis

The purpose of the syntactic analysis is to delimit the various phrases which compose the input sentences and to determine the grammatical functions of the various phrases. The analysis relies on certain grammatical generalizations.

Signaling Keyphrases

If the sentence begins with a noun phrase of a sentential qualifier, such as an adverbial, then it is neither a question nor a command, and may be accepted as data by the system. If the sentence begins with a question word it is a factual question, and if it begins with a verb it is a command. If the sentence begins with a modal, then it is either a yes-no question as a request, depending on certain contextual conditions which we need not discuss here.

If the question word is an adverbial, such as "when," "where," "how," or "why," then we must consider both the subject and the object of the sentence to be relevant. The same is true of yes-no questions. However, if the question word is "what" or "who," then we know in advance that either the subject or the object of the sentence is being questioned and that the other will constitute the keyphrase.

Similarly, other keyphrases may be signaled by their position in the sentence or by their function. Thus a prepositional phrase introduced by "about" signals the presence of significant keywords in the following noun phrase.

Matching

After the keyphrase has been isolated, it is reduced to a set of keywords. The index of each keyword is looked up in the dictionary, and the keyword with the longest decimal index is matched against the text. If no match is found, then paraphrases of this keyword are matched against the text. If a match is found, then the remaining keywords are matched against a portion of the text consisting of *n* sentences in the neighborhood of the sentence containing the most specific keyword. The section of text containing all or some of the matched keywords is retrieved as a response.

Variations

A notable characteristic of a method such as the one described immediately above is the existence of variables whose values may be changed in order to effect a change in the operation of the system. These variables may be defined in mathematical terms, so that a change in the system would not require a major programming change.

Given a sufficiently rich dictionary, it is possible to generate virtually unlimited quantities of paraphrases to be used in matching. Thus one degree of flexibility comes from being able to determine how much more specific than the keywords a paraphrase may be.

Given also that the target area for the keyphrase match is defined as a certain number of sentences in the environment of the primary keyword, this affords us a further degree of flexibility. As mentioned previously, the size of the environment could be a function of the number of keywords.

Alternatively, the size of the environment could be fixed, while the variable could be the percentage of the total number of keywords required to constitute a satisfactory match. The percentage could also be a function of the number of keywords.

The first method is actually more flexible than the second, although probably slower. It is possible to envision a target technique whereby if the environment number is *n,* and if the second keyword falls in the ith sentence from the first keyword, then the third keyword must fall within *n-i* sentences of either of the previous keywords, or between them. In theory such a technique seems rather cumbersome, but it should be pointed out that it affords one the option of specifying precisely the sharpness of definition desired in the matching process. It is quite possible that, rather than being a benefit, such a high degree of sophistication would be a liability in view of its relative slowness in an area where speed is as important a factor as precision. When the paraphrase option is taken into consideration, the difference in speed would be magnified, *ceteris paribus,* by the degree of increased specification one was willing to allow in the generation of paraphrases.

On the other hand, one might sacrifice precision for speed and define the environment number uniquely, either for the system, or for the input sentence. In other words, given *k* keywords, one could use a function (*k*) to determine the maximum number of sentences *i* on either side of the primary keyword within which all the keywords (or their paraphrases) must fall. Figure 4 illustrates a successful match.

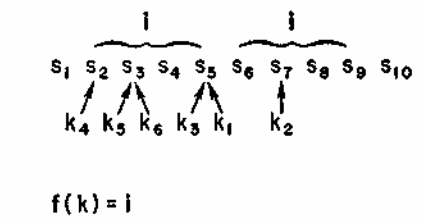


FIG. 4.—Example of a successful match

In this example the text returned as a response would be $s_2 \ldots s_7$.

One might find, while varying *i* according to the value of *k,* that a function which provides a relative maximum of precision with a relative minimum of procession is $f(k) = i = C$, where C is some constant. Given the extent of present knowledge in this field, anything like the correct values for these variables would be impossible to specify in the absence of empirical results in terms of precision and speed. Figure 5 is a flowchart which sketches out the general outlines of a procedure such as the one described in Part 4.

Received January 21, 1969
Revised September 5, 1969

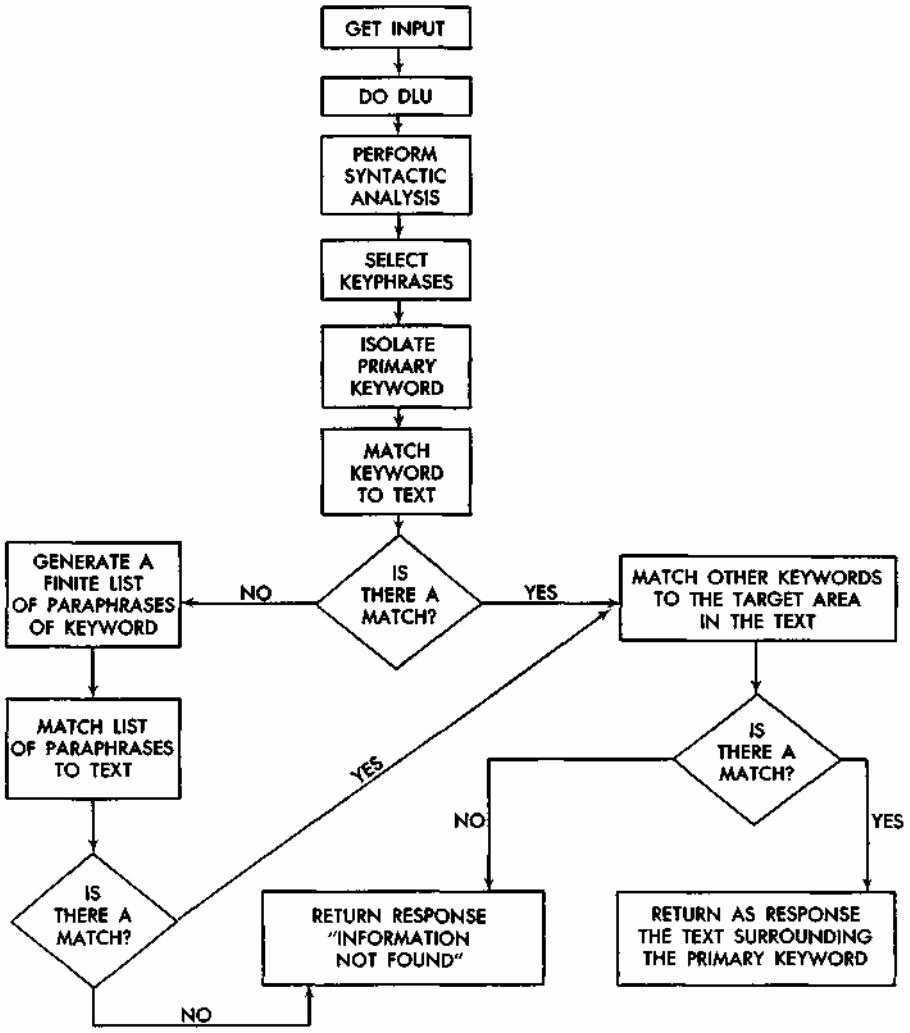


FIG. 5.—Flowchart of a possible implementation of the keyword-keyphrase method

References

1. Alt, F. L., and Rhodes, I. "Recognition of Clauses and Phrases in Machine Translation of Languages." In *Proceedings First International Conference on Machine Translation of Languages and Applied Language Analysis.* London: Her Majesty's Stationery Office, 1961.
2. Anderson, T. "A Model of Language Use." Communication to summer meeting of the Linguistic Society of America, University of California, Los Angeles, July 1966.
3. Bar-Hillel, Y. "Logical Syntax and Semantics." *Language,* no. 30 (1954).
4. Bar-Hillel, Y. "Discussion on Papers by Mr. R. H. Richens and Dr. L. Brandwood." In *Proceedings Symposium on Mechanisation of Thought Processes.* Vol. 1. London: Her Majesty's Stationery Office, 1959.
5. Bobrow, D. G. "Natural Language Input for a Computer Problem Solving System." Technical report MAC-TR-1. Ph.D. dissertation, M.I.T., Cambridge, Mass., September 1964.
6. Bohnert, H. "Logical-Linguistic Studies for Machine Text Perusal." Semi-annual status reports, IBM Thomas J. Watson Research Center, Yorktown Heights, N.Y., 1962-65.
7. Cambridge Language Research Unit. "Colloquium Report." In *Semantic Problems in Language.* Cambridge, 1962.
8. Ceccato, B. T., and Jones, P. E. "Automatic Derivation of Microsentences." *Communications of the ACM,* vol. 9, no. 6 (June 1966).
9. Chomsky, N. "The Logical Structure of Linguistic Theory." Mimeographed. Cambridge, Mass.: M.I.T. Library, 1965.
10. Chomsky, N. "Context-free Grammars and Pushdown Storage." *Quarterly Progress Report no. 65,* Research Laboratory of Electronics, M.I.T., Cambridge, Mass., 1962.
11. Chomsky, N. *Syntactic Structures.* The Hague: Mouton & Co., 1957.
12. Chomsky, N. "Remarks on Nominalization." In *Readings in Transformational Grammar,* edited by R. A. Jacobs and P. S. Rosenbaum. Waltham, Mass.: Blaisdell Publishing Co., in press.
13. Chomsky, N. *Aspects of the Theory of Syntax.* Cambridge, Mass.: M.I.T. Press, 1965.

14. Chomsky, N. *Cartesian Linguistics.* New York: Harper & Row, 1966.
15. Chomsky, N., and Miller, G. Three chapters in *Handbook of Mathematical Psychology,* edited by R. D. Luce, R. R. Bush, and E. Galanter. Vol. 2. New York: John Wiley & Sons, 1963.
16. Craig, J. A. "Dependency Aspects of the DEACON Context-Dependent Phrase-Structure Grammar." Communication to fourth annual meeting of Association for Machine Translation and Computational Linguistics, University of California, Los Angeles, July 1966.
17. Darlington, J. "Machine Methods for Proving Logical Arguments in English." *Machine Translation* (June 1965).
18. Dubois, J. "An Experimental Approach to Generating Kernel Sentences in French." Contribution to 1966 Linguistic Institute Conference, Bunker-Ramo Corp., Los Angeles, Calif., July 1966.
19. Fromkin, V. A. "Some Requirements for a Model of Speech Performance." Communication to summer meeting of the Linguistic Society of America, University of California, Los Angeles, July 1966.
20. Green, B. F., et al. "Baseball: An Automatic Question Answerer." In *Proceedings Western Joint Computer Conference,* May 1961.
21. Gross, M. "On the Equivalence of Models Used in the Fields of Mechanical Translation and Information Retrieval." In *Information Storage and Retrieval.* Vol. 2. London: Pergamon Press, 1964.
22. Halliday, M. A. K. "Some Aspects of Thematic Organization of the English Clause." Communication to summer meeting of the Linguistic Society of America, University of California, Los Angeles, July 1966.
23. Harris, Z. "Discourse Analysis." *Language,* vol. 28, no. 1 (January 1952).
24. Harris, Z. "Co-occurrence and Transformation in Linguistic Structure." *Language,* no. 33 (1957), pp. 293-340.
25. Hays, D. G. "New Rule Forms and Characterization Procedures for Phrase Structure and Transformation Rules." Communication to summer meeting of the Linguistic Society of America, University of California, Los Angeles, July 1966.
26. Irons, E. T. "Structural Connections in Formal Languages." In Proceedings IDA Conference on Language Structure, Princeton, N.J., 1963.
27. Katz, J. J. *The Philosophy of Language.* New York: Harper & Row, 1966.
28. Katz, J. J., and Fodor, J. A. "The Structure of a Semantic Theory." *Language* 39 (1963): 170-210.
29. Katz, J. J., and Postal, P. *An Integrated Theory of Linguistic Descriptions.* Cambridge, Mass.: M.I.T. Press, 1964.
30. Kuno, S. "A System for Transformational Analysis." In Proceedings International Conference on Computational Linguistics, New York, 1965.
31. Kuno, S. "Computer Analysis of Natural Language." To appear in Proceedings of the Symposium on Mathematical Aspects of Computer Science, American Mathematical Society, New York, April 1966.
32. Kuno, S., and Oettinger, A. "Multiple Path Syntactic Analysis." In Proceedings IFIPS Congress, Munich, Germany, 1962.
33. Kuno, S., and Oettinger, A. "Syntactic Structure and Ambiguity of English." In *Proceedings Fall Joint Computer Conference.* Baltimore: Spartan Press, November 1963.
34. Lees, R. *The Grammar of English Nominalization.* The Hague: Mouton & Co., 1960.
35. Lieberman, D. "Computer Support for Lexicon Development." In *Specification and Utilization of a Transformational Grammar,* edited by D. Lieberman. AFCRL-66-270. Yorktown Heights, N.Y.: IBM Thomas J. Watson Research Center, March 1966.
36. Lindsay, R. K. "Inferential Memory as the Basis of Machines which Understand Natural Language." In *Computers and Thought,* edited by E. A. Feigenbaum and J. Feldman. New York: McGraw-Hill, 1963.
37. Matthews, G. H. "Analysis by Synthesis of Sentences of Natural Languages." In *Proceedings First International Conference on Machine Translation of Languages and Applied Language Analysis.* London: Her Majesty's Stationery Office, 1961.
38. Moyne, J. A. "A Workpaper Describing Concepts, Scope and Implementation Considerations for Proto-RELADES." Unpublished. IBM Boston Programming Center, August 1966.
39. Moyne, J. A. "A Simulated Computer for Natural Language Processing." IBM Systems Development Division technical report TR 00.1463, Poughkeepsie, N.Y., July 1966.
40. Raphael, B. "SIR: A Computer Program for Semantic Information Retrieval." Ph.D. dissertation, Mathematics Department, M.I.T., Cambridge, Mass., 1964.
41. Rhodes, I. "A New Approach to Mechanical Syntactic Analysis of Russian." *Machine Translation,* vol. 6, no. 2 (1969).
42. Rosenbaum, P., and Lochak, D. "The IBM CORE Grammar of English." In *Specification and Utilization of a Transformational Grammar,* edited by D. Lieberman. AFCRL-66-270-IBM. Yorktown Heights, N.Y.: IBM Thomas J. Watson Research Center, March 1966.
43. Sammet, J, "The Use of English as a Programming Language." *Communications of the ACM* (March 1966), pp. 228-32.
44. Simmons, R. F. "Natural Language Processing" *Datamation* (June 1966).
45. Tabory, R., and Peters, S., Jr. "Ambiguity, Completeness and Restriction Problems in the Syntax-Based Approach to Computational Linguistics." Technical report TR 86-78, IBM Federal Systems Division, Cambridge, Mass., November 1966.
46. Weizenbaum, J. "ELIZA: A Computer Program for the Study of Natural Language Communication between Man and Machine." Cambridge, Mass.: Department of Electrical Engineering, M.I.T., August 1965.
47. Woolley, G. H. "Syntax Analysis Beyond the Sentence." Communication to fourth annual meeting of Association for Machine Translation and Computational Linguistics, University of California, Los Angeles, July 1966.
48. Yngve, V. "A Model and a Hypothesis for Language Structure." *Proceedings of the American Philosophical Society,* vol. 194, no. *5* (1969).