Is Natural Language an Unnatural Query Language?

Christine A. Montgomery, Operating Systems, Inc.

ABSTRACT

The complex system of form/meaning correlations comprising a natural language presents a considerable challenge for automated interpretation. In order to deal with such complexity, attempts at automated interpretation of natural language queries have typically concentrated on limited subsets of natural language. However, such subsets are inevitably ill-defined in some way, adding another class of interpretive problems to the well-known ones of syntactic ambiguity and ungrammatical strings. While the other problems are practically resolvable in some sense, the ability of an automated system to resolve the many syntactic ambiguities of natural language is a function of the interpretive power of the linguistic model upon which the system is based. Although previous models have lacked the interpretive power to deal with natural language, a novel approach integrating the treatment of form and meaning may provide an effective basis for handling natural language as an instrument of communication in automated systems.

KEY WORDS AND PHRASES: natural language processing, query languages, syntax, semantics
CR CATEGORIES: 3.42; 3.60, 3.71

The use of natural language as a query language presents certain formidable problems due to its richness as an instrument of communication. Unless these problems can be solved in some systematic manner in the context of a given application, natural language is not a natural query language.

The complex system of form/meaning relations which comprises a natural language in theory allows humans to produce an infinite set of well-formed 'sentences', not to mention a presumably infinite set of ill-formed 'utterances', which are not as yet accounted for within the theory. This prolific generative capacity of language systems makes interpretation of natural language sentences a difficult enough task for humans, let alone computers.

Humans have the distinct advantage of somehow carrying around in their heads a representation of the system of a particular language (or languages), as well as representations of many possible 'worlds' or situational contexts which may provide differing interpretations of the same string of formal elements. A computer, on the other hand, is typically equipped only with the representation of a small fragment of the system of the language--i.e., a minigrammar capable of interpreting some ill-defined subset of the infinite set of well-formed sentences--and with some equally inadequate representation of the possible world relevant to the given application--i.e., a semantic interpretive device of some sort.

The problem of scale in the automated query systems thus far designed has been discussed elsewhere by this author [1,2]. It is clearly unrealistic to expect that the computer's fragmentary representations of grammatical and encyclopedic knowledge in the models discussed in references 1 and 2 will be able to deal even with the variety of natural language queries which may be used in a given situational context, quite apart from other types of well-formed sentences and ungrammatical strings, as well as other situational contexts.

As evidence for this contention, it is instructive to examine a few queries from the point of view of the computer attempting to interpret them. All of the queries listed below are syntactically and lexically ambiguous in various ways[1], and the computer must somehow select the correct interpretation from which to derive a search procedure.

(a) what proton energies are necessary to produce Bev pi meson beams?

---

[1] These are not the phantasmagorical product of a linguist's fevered imagination, but are taken from a set of 300 questions generated by nuclear physicists for some early experiments in information retrieval [3,4].

(b) how can coherent neutron scattering cross sections be measured for materials with high absorption cross sections?

(c) how can one measure low values of total coherent cross sections of the scattering of neutrons by heavy nuclei?

(d) how can the average energy required to make an ion pair in various gases be measured?

(e) what is Møller scattering?

Query (a) is fairly straightforward except for the problem of bracketing the final noun phrase properly. Using the thesaurus compiled for the given experiments, 'pi meson' can be identified as a 'phrase' or lexical unit, but--from the point of view of a non-nuclear physicist--'Bev' could modify either 'pi meson' or 'beams'.

Similarly, query (b) contains two noun phrases with multiple modifiers, where bracketing is also a problem, since 'cross section' is the only lexical unit identified in the thesaurus. In addition, the prepositional phrase beginning with 'with' could modify the verb 'measure' as well as the preceding prepositional phrase. Query (c) contains similar ambiguities, as well as an ambiguous 'by' phrase, which could modify the main verb 'measure', or--assuming a transformational interpretation--indicate the passive subject of the verb 'scattering' in the embedded sentence. Query (d) contains a major category ambiguity in the word 'pair', which could be either a noun or a verb.

Query (e) exhibits several different types of ambiguities, depending on whether 'scattering' is interpreted as the present active participle of the finite verb, or as a verbal noun. In the first case, the query would be structurally analogous to 'what is Horace doing?' In the second, a further ambiguity arises with respect to whether 'Møller scattering' is analogous to examples (f), (g), or (h) below: in other words, whether 'Møller' is essentially genitive, being derived from some embedded sentence relating to the discovery of scattering by Møller, or whether Møller is subject or object of another type of embedded sentence where 'scatter' is the main verb.

(f) Broca's area

(g) student teaching

(h) pilot training

Queries containing conjunctions may be ambiguous with respect to the scope of the conjunction. Note the ambiguity of 'and' in query (i) as conjoining either adjectives or noun phrases, and the related ambiguity of 'or'.

(i) find an explicit relationship between the polarization of the incident and emitted particles or photons in a capture process.

(j) what nuclear reactions are sensitive to the spin and parity of mesons and hence are useful in measuring those quantities?

Another dimension of difficulty emerges in query (j) where the notion that 'spin' and 'parity' are subsumed under 'quantities' (or vice versa) is required in the analysis of the question.

For all the queries except (a), this type of relational information is required in order to derive an answer. For example, if query (e) were simply 'what is a book?', the analysis procedure would be relatively straightforward; however, a complex relational structure defining set membership and equivalence relations within the given universe of discourse is required to produce an answer to the question.

Two interrelated issues arise in this connection. One concerns the problem of scale in systems using natural language as a query language and the other the interpretive power of the given system--i.e., its adequacy to analyze and represent the content of natural language queries.

Dealing first with scale, it is not technologically possible to construct a question-answering or fact retrieval system which can accurately analyze and represent the content of any input string of a given natural language. The size and complexity of such a system present a mind-boggling image. The models of query systems must be designed to handle some particular subset of a natural language, e.g. the language of bibliographic data in cybernetics [5]. In fact, it is only a minimal subset of that sublanguage which is actually handled in most query systems which have been designed to date; hence the question arises as to whether models constructed on so small a scale can be at all realistic [1, 2].

Another problem involved in scaling down the design of a natural language query system is that of defining the natural language subset which the system will be capable of interpreting. It is important to recognize that the boundaries of such a subset are inevitably ill-defined in some way. The system designer can never be absolutely certain as to what his natural language interpreter will interpret or what the interpretation will be--assuming that the system can supply some information for each item in an input string which has not previously been presented to it. On the other hand, some boundaries of the natural language subset may be rigorously defined: e.g., input strings containing conjunctions may be

unacceptable.

From the point of view of the prospective system user, the fact that the boundaries of the natural language subset are well-defined in some respects and ill-defined in others can be expected to cause no small amount of consternation. In the first case, the user may be stymied by a system message to rephrase his request--say, to generate a paraphrase without conjunctions. This raises the basic issue of 'habitability' of a natural language subset defined by Watt [6] in terms of the user's ability to communicate within the confines of the sublanguage acceptable to the machine.

Conversely, where the subset boundaries are ill-defined--essentially, where an ambiguity is not recognized by the interpretive system--the user receives a response to a question he did not intend to ask. For example, assume that the interpretive system always identifies an -ing form as either adjectival or nominal, which would provide a correct interpretation for query (e) above. If a user should then ask 'What is Teller studying?', the answer might be 'File contains no such entry', --when in fact the file may contain information on the current research activities of Teller. If the user has previously been requested to reformulate questions phrased in a form unacceptable to the system (i.e., he has strayed over a well-defined boundary), he may assume that since the form of his question was not challenged, it was correctly interpreted and answered by the system. On the contrary, he has in this instance strayed over an ill-defined boundary of the natural language subset which can be interpreted by the system. Because the boundary is ill-defined, neither the user nor the system designer is aware of the error.

Another issue related to the definition of a natural language subset is the problem of dealing with ill-formed queries. The following are examples of ill-formed queries taken from the set of queries described in footnote 1:

(k) does experimental data show whether anti-particles have negative inertial mass?

(l) in the measurement of single spectra Beta radiation, how must the absorptive material be placed...?

(m) what is the retardation affect of the mesonic potential in calculating neutron-proton interaction?

(n) how do boson particle and anti-particle differ?

(o) what is the ratio between the total cross section for 400 Mev neutrons in light elements?

(p) how may billion volts neutron beams be cleared of x-rays and low energy neutrons?

Queries (k) and (l) show lack of grammatical number agreement between subject and verb [(k)], and lack of lexical number agreement between adjective and noun [(l)], due to rather common errors in the usage of Latin plural forms. Query (m) also contains a fairly frequent erroneous substitution of 'affect' for 'effect'. If errors of this nature are frequent in a given natural language subset, then queries containing such errors should be accepted by the system as legitimate. Otherwise, entering query (k) would be equivalent to violating a well-defined boundary of the natural language subset, in that it involves a number agreement which is presumably recognized by the system, resulting in display of an error message. On the other hand, entering query (l) would be analogous to violating an ill-defined boundary of the natural language subset, since errors in lexical number agreement would presumably go unrecognized, and no information on 'single spectra Beta radiation' would be found.

Similarly, the lack of number agreement between subject and verb in query (n) and the error in number agreement between the preposition 'between' and its object in query (o) would most likely be recognized by the system and an error message displayed, while the error in number among the modifiers of the subject in (p) would probably not be detected. Although the errors in these queries are not due to particular lexical items, as are those in queries (k), (l), and (m), it is interesting to note that they all involve confusions in number agreement. Thus a practical means of coping with a potentially considerable percentage of ill-formed queries might be to relax the rules for number agreement in analyzing input queries and in representing them as search prescriptions in some formal language. Alternatively, all types of number agreement could be tested for, and a request for clarification presented to the user whenever agreement is lacking.

In any case, what is needed in order to provide a practical basis for making such a determination is a record of user-system interactions. Monitoring user inputs and system responses provides a means for developing information on the nature of ill-formed queries characterizing the particular universe of discourse as well as instances of interactions involving the subtle errors of straying across ill-defined boundaries of the natural language subset. This device to generate information for feedback improvement of a query system seems a particularly worthwhile investment, inasmuch as ill-formed queries will always constitute a certain percentage of the input and the insidious errors of

violating ill-defined boundaries can be expected to persist throughout an indefinitely long developmental stage.

The matter of interpretive power has been touched on in the preceding discussion on defining the natural language subset which a given query system will be capable of interpreting. However, the subject of interpretive power--or the capability of a system to analyze the content of natural language queries and to represent it in terms of formal search specifications--is obviously crucial, and merits a separate paper. In the interests of conserving time and space, my discussion of this topic is limited to one basic issue. Adapting an appropriate title from the pages of a linguistic journal, this issue might be called 'the Non-Uniqueness of Natural Language Information Processing Approaches'. If one examines the spectrum of models for processing natural language information, perhaps the most striking attribute of the entire inventory is its variety. There are some similarities,[2] but on the whole, each system designer has essentially done his own thing. The reason for this is, of course, to be found in the complexity of natural language, and in the lack of a theory of language to explicate the complexity.

The complex system of form/meaning correlations which constitutes a natural language is a phenomenon of awesome complexity taken in its entirety. Thus in order to deal with such a complex phenomenon, investigators have generally attempted to separate the treatment of form from the treatment of meaning. However, in so doing, they have in effect redefined the object of their investigations, since its most essential characteristic is that of form/meaning correlation. Given this tradition of disconnecting form from meaning, it is not surprising that no comprehensive theory of language has been developed.

A novel approach which integrates the treatment of form and meaning has recently been proposed by Montague [8]. The significance of Montague's theory derives from the unified description of form and meaning, such that for every syntactic rule there is a corresponding semantic rule which provides a translation of natural language phrases into expressions in intensional logic. The intensional logic is a metalanguage providing a coherent semantic interpretation for every expression in terms of a truth definition for a possible world. This theory appears to present a more powerful and consistent framework for the treatment of natural language than theories previously proposed by various linguists.

In an earlier paper dealing with natural language information systems [2], I suggested an approach which combines some concepts of current linguistic theory (i.e., Fillmore's role notions) and the concepts of relational logic embodied in several computational linguistic models. In future work on natural language information systems, I would like to integrate these concepts within the framework of Montague's theory. I feel strongly that this approach--together with the performance monitoring concept described above--can provide a systematic framework for dealing with natural language as an instrument of communication in automated systems, thus overcoming even my own reservations concerning the naturalness of natural language in an automated context.

[2] See Simmons [7] for a detailed discussion; see also [2].

REFERENCES

1. Montgomery, C. A., "Automated language processing," in Annual Review of Information Science and Technology, (ed. by Carlos A. Cuadra) Vol. 4: 145-174. Encyclopedia Britannica Press, 1969.

2. _____ "Linguistics and information science," (To appear in the Journal of the American Society for Information Science, May-June 1972.) 101 p.

3. Ramo Wooldridge Division, Thompson Ramo Wooldridge, Inc. Word correlation and automatic indexing. Phase I Final Report to the Council on Library Resources: Vol. XIII, Experimental data. 1960.

4. _____ Word correlation and automatic indexing. Phase IIA Final Report to the Council on Library Resources: Appendix A: Thesaurus, and Appendix C: Automatic conversion of 88 questions to search terms. January 1962.

5. Kuhns, J. L., "Logical aspects of question-answering by computer," in Software Engineering, vol. 2 (J. Tou, ed), pp. 89-104, Academic Press, Inc., New York, 1971.

6. Watt, W. C., "Habitability," American Documentation (Journal of the American Society for Information Science), Vol. 19.3: 338-351.

7. Simmons, R. F., "Natural language question-answering systems: 1969," Communications of the ACM, 13.1 (January 1970): 15-30.

8. Montague, R., The proper treatment of quantification in ordinary English," in Hintikka, J., Moravcsik, J., and Suppes, P., (eds), Approaches to natural language, Humanities Press, Dordrecht: Reidel (in press).