Building a Large Knowledge Base for a Natural Language System

Jerry R. Hobbs

Artificial Intelligence Center
SRI International
and
Center for the Study of Language and Information
Stanford University

Abstract

A sophisticated natural language system requires a large knowledge base. A methodology is described for constructing one in a principled way. Facts are selected for the knowledge base by determining what facts are linguistically presupposed by a text in the domain of interest. The facts are sorted into clusters, and within each cluster they are organized according to their logical dependencies. Finally, the facts are encoded as predicate calculus axioms.

1. The Problem[1]

It is well-known that the interpretation of natural language discourse can require arbitrarily detailed world knowledge and that a sophisticated natural language system must have a large knowledge base. But heretofore, the knowledge bases in natural language systems have either encoded only a few kinds of knowledge - e.g., sort hierarchies - or facts in only very narrow domains. The aim of this paper is to present a methodology for constructing an intermediate-size knowledge base for a natural language system, which constitutes a manageable and principled midway point between these simple knowledge bases and the impossibly detailed knowledge bases that people seem to use.

The work described in this paper has been carried out as part of a project to build a system for natural language access to a computerized medical textbook on hepatitis. The user asks a question in English, and rather than attempting to answer it, the system returns the passages in the text relevant to the question. The English query is translated into a logical form by a syntactic and semantic translation component [Grosz et al., 1982]. The textbook is represented by a "text structure", consisting, among other things, of summaries of the contents of individual passages, expressed in a logical language. Inference procedures, making use of a knowledge base, seek to match the logical form of the query with some part of the text structure. In addition, they attempt to attempt to solve various pragmatic problems posed by the query, including the resolution of coreference, metonymy, and the implicit predicates in compound nominals. The inference procedures are discussed elsewhere [Walker and Hobbs, 1981]. In this paper a brief example will have to suffice.

Suppose the user asks the question, "Can a patient with mild hepatitis engage in strenuous exercise?" The relevant passage in the textbook is labelled, "Management of the Patient: Requirements for Bed Rest". The inference procedures must show that this heading is relevant to this question by drawing the appropriate inferences from the knowledge base. Thus the knowledge base must contain the facts that rest is an activity that consumes little energy, that exercise is an activity, and that if something is strenuous it consumes much energy, and axioms that relate the concepts "can" and "require" via the concept of possibility.

One way to build the knowledge base would have been to analyze the queries in some target dialogs we collected to determine what facts they seem to require, and to put just these facts into our knowledge base. However, we are interested in discovering general principles of selection and structuring of such intermediate-sized knowledge bases, principles that would give us reason to believe our knowledge base would be useful for unanticipated queries.

Thus we have developed a three-stage methodology:

1. Select the facts that should be in the knowledge base by determining what facts are linguistically presupposed by the medical textbook. This gives us a very good indication of what knowledge of the domain the user is expected to bring to the textbook and would bring to the system.

2. Organize the facts into clusters and organize the facts within each cluster according to the logical dependencies among the concepts they involve.

3. Encode the facts as predicate calculus axioms, regularizing the concepts, or predicates, as necessary.

These stages are discussed in the next three sections.

2. Selecting the Facts

To be useful, a natural language system must have a

[1] I am indebted to Bob Amsler and Don Walker for discussions concerning this work. This research was supported by NIH Grant LM03611 from the National Library of Medicine, by Grant IST-8209346 from the National Science Foundation, and by a gift from the Systems Development Foundation.

large vocabulary. Moreover, when one sets out to axiomatize a domain, unless one has a rich set of predicates and facts to be responsible for, a sense of coherence in the axiomatization is hard to achieve. One's efforts seem ***ad hoc***. So the first step in building the knowledge base is to make up an extensive list of words, or predicates, or concepts (the three terms will be used interchangeably here), and an extensive list of relevant facts about these predicates. We chose about 350 words from our target dialogs and headings in the textbook and encoded the relevant facts involving these concepts. Because there are dozens of facts one could state involving any one of these predicates, we were faced with the problem of determining those facts that would be most pertinent for natural language understanding in this domain.

Our principal tool at this stage was a full-sentence concordance of the textbook, displaying the contexts in which the words were used. Our method was to examine these contexts and to ask what facts about each concept were required to justify each of these uses, what did their uses linguistically presuppose.

The three principal linguistic phenomena we looked at were predicate-argument relations, compound nominals, and conjoined phrases. As an example of the first, consider two uses of the word "data". The phrase "extensive data on histocompatibility antigens" points to the fact about data that it is a *set* (justifying "extensive") of particular facts *about some subject* (justifying the "on" argument). The phrase "the data do not consistently show ..." points to the fact that data is assembled to support some conclusion. To arrive at the facts, we ask questions like "What is data that it can be extensive or that it can show something?" For compound nominals we ask, "What general facts about the two nouns underlie the implicit relation?" So for "casual contact circumstances" we posit that contact is a concomitant of activities, and the phrase "contact mode of transmission" leads us to the fact that contact possibly leads to transmission of an agent. Conjoined noun phrases indicate the existence of a superordinate in a sort hierarchy covering all the conjoined concepts. Thus, the phrase "epidemiology, clinical aspects, pathology, diagnosis, and management" tells us to encode the facts that all of these are aspects of a disease.

As an illustration of the method, let us examine various uses of the word "disease" to see what facts it suggests:

- "destructive liver disease": A disease has a harmful effect on one or more body parts.
- "hepatitis A virus plays a role in chronic liver disease": A disease may be caused by an agent.
- "the clinical manifestations of a disease": A disease is detectable by signs and symptoms.
- "the course of a disease": A disease goes through several stages in time.
- "infectious disease": A disease can be transmitted.
- "a notifiable disease": A disease has patterns in the population that can be traced by the medical community.

We emphasize that this is not a mechanical procedure but a method of discovery that relies on our informed intuitions. Since it is largely background knowledge we are after, we can not expect to get it directly by interviewing experts. Our method is a way of extracting it from the presuppositions behind linguistic use.

The first thing our method gives us is a great deal of selectivity in the facts we encode. Consider the word "animal". There are hundreds of facts that we know about animals. However, in this domain there are only two facts we need. Animals are used in experiments, as seen in the compound nominal "laboratory animal", and animals can have a disease, and thus transmit it, as seen in the phrase "animals implicated in hepatitis". Similarly, the only relevant fact about "water" is that it may be a medium for the transmission of disease.

Secondly, the method points us toward generalizations we might otherwise miss, when we see a number of uses that seem to fall within the same class. For example, the uses of the word "laboratory" seem to be of two kinds:

1. "laboratory animals", "laboratory spores", "laboratory contamination", "laboratory methods".
2. "a study by a research laboratory", "laboratory testing", "laboratory abnormalities", "laboratory characteristics of hepatitis A", "laboratory picture".

The first of these rests on the fact that experiments involving certain events and entities take place in laboratories. The second rests on the fact that information is acquired there.

A classical issue in lexical semantics that arises at this stage is the problem of polysemy. Should we consider a word, or predicate, as ambiguous, or should we try to find a very general characterization of its meaning that abstracts away from its use in various contexts? The concordance method suggests a solution. The rule of thumb we have followed is this: if the uses fall into two or three distinct, large classes, the word is treated as having separate senses, whereas if the uses seem to be spread all over the map, we try to find a general characterization that covers them all. The word "derive" is an example of the first case. A derivation is either of information from an investigative activity, as in "epidemiologic patterns derived from historical studies", or of chemicals from body parts, as in "enzymes derived from intestinal mucosa". By contrast, the word "produce" (and the word "product") can be used in a variety of ways: a disease can produce a condition, a virus can produce a disease or a viral particle, something can produce a virus ("the amount of virus produced in the carrier state"), intestinal flora can produce compounds, and something can produce chemicals from blood ("blood products"). All of this suggests that we want to encode only the fact that if x produces y, then x causes y to come into existence.

At this stage in our method, we aimed at only informal, English statements of the facts. We ended up with approximately 1000 facts for the knowledge base.

3. Organizing the Knowledge Base

The next step is to sort the facts into natural "clusters" (cf. [Hayes, 1984]). For example, the fact "If x produces y, then x causes y to exist" is a fact about causality. The fact "The replication of a virus requires components of a cell of an organism" is a fact about viruses. The fact "A household is an environment with a high rate of intimate contact, thus a high risk of transmission" is in the cluster of facts about people and their activities. The fact "If bilirubin is not secreted by the liver, it may indicate injury to the liver tissues" is in the medical practice cluster.

It is useful to distinguish between clusters of "core knowledge" that is common to most domains and "domain-specific knowledge". Among the clusters of core knowledge are space, time, belief, and goal-directed behavior. The domain-specific knowledge includes clusters of facts about viruses, immunology, physiology, disease, and medical practice. The cluster of facts about people and their activities lies somewhere in between these two.

We are taking a rather novel approach to the axiomatization of core knowledge. Much of our knowledge and language seems to be based on an underlying "topology", which is then instantiated in many other areas, like space, time, belief, social organizations, and so on. We have begun by axiomatizing this fundamental topology. At its base is set theory, axiomatized along traditional lines. Next is a theory of granularity, in which the key concept is "x is indistinguishable from y with respect to grain g". A theory of scalar concepts combines granularity and partial orders. The concept of change of state and the interactions of containment and causality are given (perhaps overly simple) axiomatizations. Finally there is a cluster centered around the notion of a "system", which is defined as a set of entities and a set of relations among them. In the "system" cluster we provide an interrelated set of predicates enabling one to characterize the "structure" of a system, producer-consumer relations among the components, the "function" of a component of a system as a relation between the component's behavior and the behavior of the system as a whole, notions of normality, and distributions of properties among the elements of a system. The applicability of the notion of "system" is very wide; among the entities that can be viewed as systems are viruses, organs, activities, populations, and scientific disciplines.

Other general commonsense knowledge is built on top of this naive topology. The domain of time is seen as a particular kind of scale defined by change of state, and the axiomatization builds toward such predicates as "regular" and "persist". The domain of belief has three principal subclusters in this application: learning, which includes such predicates as "find", "test" and "manifest"; reasoning, explicating predicates such as "leads-to" and "consistent"; and classifying, with such predicates as "distinguish", "differentiate" and "identify". The domain of modalities explicates such concepts as necessity, possibility, and likelihood. Finally, in the domain of goal-directed behavior, we characterize such predicates as "help", "care" and "risk".

In the lowest-level domain-specific clusters – viruses, immunology, physiology, and people and their activities – we begin by specifying their *ontology* (the different sorts of entities and classes of entities in the cluster), the *inclusion relations* among the classes, the behaviors of entities in the clusters and their interactions with other entities. The "Disease" cluster is axiomatized primarily in terms of a temporal schema of the progress of an infection. The cluster of "Medical Practice", or medical intervention in the natural course of the disease, can be axiomatized as a plan, in the AI sense, for maintaining or achieving a state of health in the patient, where different branches of the plan correspond to where in the temporal schema for disease the physician intervenes and to the mode of intervention.

Most of the content of the domain-specific clusters is specific to medicine, but the general principles along which it was constructed are relevant to many applications. Frequently the best way to proceed is first to identify the entities and classification schemes in several clusters, state the relationships among the entities, and encode axioms articulating clusters with higher- and lower-level clusters. Often one then wants to specify temporal schemas involving interactions of entities from several domains and goal-directed intervention in the natural course of these schemas.

The concordance method of the second stage is quite useful in ferreting out the relevant facts, but it leaves some lacunae, or gaps, that become apparent when we look at the knowledge base as a whole. The gaps are especially frequent in commonsense knowledge. The general principle we follow in encoding this lowest level of the knowledge base is to aim for a vocabulary of predicates that is minimally adequate for expressing the higher-level, medical facts and to encode the obvious connections among them. One heuristic has proved useful: If the axioms in higher-level domains are especially complicated to express, this indicates that some underlying domain has not been sufficiently explicated and axiomatized. For example, this consideration has led to a fuller elaboration of the "systems" domain. Another example concerns the predicates "parenteral", "needle" and "bite", appearing in the domain of "disease transmission". Initial attempts to axiomatize them indicated the need for axioms, in the "naive topology" domain, about membranes and the penetration of membranes allowing substances to move from one side of the membrane to the other.

Within each cluster, concepts and facts seem to fall into small groups that need to be defined together. For example, the predicates "clean" and "contaminate" need to be defined in tandem. There is a larger example in the "Disease Transmission" cluster. The predicate "transmit" is fundamental, and once it has been characterized as the motion of an infectious agent from a person or animal to a person via some medium, the predicates "source", "route", "mechanism", "mode", "vehicle" and "expose" can be defined in terms of its schema. In addition, relevant facts about body fluids, food, water, contamination, needles, bites, propagation, and epidemiology rest on an understanding of "transmit". In each domain there tends to be a core of central predicates whose nature must be explicated with some care. A large number of other predicates can then be characterized fairly easily in terms of these.

4. Encoding the Facts in Predicate Calculus

Encoding world knowledge in a logical language is often taken to be a very hard problem. It is my belief that the difficulties result from attempts to devise representations that lend themselves in obvious ways to efficient deduction algorithms and that adhere to stringent ontological scruples. I have abandoned the latter constraint altogether (see [Hobbs, 1984], for arguments) and believe the former concern should be postponed until we have a better idea of precisely what sort of deductions need to be optimized. Under these ground rules, translating individual facts into predicate calculus is usually fairly straightforward.

There are still considerable difficulties in making the axioms mesh well together. A predicate should not be used in some higher-level cluster unless it has been elucidated in that or some lower-level cluster. This necessarily restricts one's vocabulary. For example, the predicate "in" does a lot of work. There are facts about viruses *in* tissues, chemicals *in* body fluids, infections *in* patient's bodies, and so on, and a direct translation of some of these axioms back into English is somewhat awkward. One has the feeling that subtle shades of meaning have been lost. But this is inevitable in a knowledge base whose size is intended to be intermediate rather than exhaustive.

5. Summary

Much of this paper has been written almost as a case study. It would be useful for me to highlight the new and general principles and results that come out of this project. The method of using linguistic presuppositions as a "forcing function" for the underlying knowledge is fairly generally applicable in any domain for which there is a large body of text to exploit. It has been used in ethnography and discourse analysis, but to my knowledge it has not been previously used in the construction of an AI knowledge base. The core knowledge has been encoded in ways that are independent of domain and hence should be useful for any natural language application. Of particular interest here is the identification and axiomatization of the topological substructure of language The domain-specific knowledge will not of course carry over to other applications, but, as mentioned above, certain general principles of axiomatizing complex domains have emerged.

References

Grosz, B., N. Haas, G. Hendrix, J. Hobbs, P. Martin, R. Moore, J. Robinson, and S. Rosenschein, 1982. DIALOGIC: A core natural language processing system. *Proceedings of the Ninth International Conference on Computational Linguistics*. 95-100. Prague, Czechoslovakia.

Hayes, P., 1984. The second naive physics manifesto. In Hobbs, J. and R. Moore (Eds.), *Formal Theories of the Commonsense World*. Ablex Publishing Company, Norwood, New Jersey.

Hobbs, J. 1984. Ontological promiscuity. Manuscript.

Walker, D. and J. Hobbs, 1981. Natural language access to medical text. SRI International Technical Note 240. March 1981.