Knowledgeable Schema

Alka Irani and **P Sadanandan**
National Centre for Software Technology,
Gulmohar cross road No. 9,
Juhu, Bombay 400049, India
e-mail:alka@ncst.ernet.in and ps@ncb.ernet.in

Abstract.
Discovering appropriate information from a huge and complex information base is a non-trivial task. In order to make a database a good description of the world it is intended to model, meaning of data in the application environment should be captured more precisely and formally. Researchers are turning to biological systems to learn from them. 'How the brain works' is still an open question and various theories have been proposed in literature. We feel the principles underlying cognition and perception in man and the tools developed by him to manage information like natural languages and discourse strategies should be studied and used rather than mimic the 'hardware' - neuron network, to make the mechanical system flexible and powerful.

Introduction

There are many ways known of systematically representing different kinds of knowledge in a sufficiently precise notation that it can be used in, or by, a computer program to solve problems. However, when the knowledge is vast and complex, the computer systems using this knowledge are likely to collapse under their own weight. Though the advances in computer hardware have made it possible to have tremendous processing power and very large capacity to store the information, the basic issue of how to model the complex data has remained a challenge.

In his 1971 Turing award lecture [7], McCarthy emphasized generality as an essential characteristic of computer systems. Generality implies representing knowledge of various sorts and retrieving it in a way that can be useful to people. It is not easy to make a system general-purpose. The construction of large, knowledge-based applications is a complex task that comprises a number of activities and involves various participants. Not only should the components - knowledge acquisition, knowledge representation and knowledge retrieval - work well individually but they should work in co-operation with one another and also with the humans who are the creators, mentors and beneficiaries of these systems. To meet this requirement, it is mandatory that both the system and people understand each other's language. In other word, if queries are to be answered at what we have been calling the computational level, some principled way of structuring knowledge must be found as the structure of knowledge has a bearing on what can be answered.

What is Knowledge Representation?

Knowledge representation is a medium of human expression, in which we primarily describe various aspects of the world. Any representation is fundamentally a surrogate [2]. It is not a thing in itself. Thus a knowledge representation formalism can be judged for its adequacy by finding how it maps the two worlds (external world) and (internal world or mental world) of human beings. The convention it uses for this purpose must be universal and stable.

Simultaneously producing good candidates for each of the three ingredients - the representation language, the inference regime and the particular domain knowledge - is as David Israel [6] characterized the crux of the representation problem.

Where to look for help?

We all know that symbol-processing systems of today are far from satisfactory. Their performance doesn't match their potential.

The nearest and more or less perfect model for the knowledge representation that can be aimed at, is the human brain. However, the nature of information we carry in our head is still not clear to us.

Researchers in various disciplines have done experimentation and have come up with various theories about how the mind works and about the functioning of brain.

Our work is based on the fundamental assumption that *the human information storage is basically semantic. The various kinds of structures we, human beings have invented like lists, tables, trees, networks, frames, etc for information management are just compression techniques and optimization techniques (syntactic devices) and they can be interpreted using a natural language. Rules, procedures, scripts, prototypes and episodes are, on the other hand, semantic devices; they serve different purposes.*

Biological Information Systems(BIS)

The tool devised by humans(BIS) to facilitate the knowledge management viz. natural language and reproduction of the knowledge in the textual form give some insight into the characteristics of their representation scheme. Language plays a unique role in the progress of BIS. It is claimed that one of the two tools that was responsible for the revolutionary progress of mankind is *language*; (the other tool is use of hands!) Noam Chomsky [1] called language "The mirror of mind". Language characterizes the input and output behaviour of the human information system. Over the years, natural languages evolved as means of communication, but according to us, *it is the role* *of language as a representation medium that has made the progress of mankind possible.*

Language is a symbol system that works within human capabilities. Language provides us with the following:

- A set of "meaningful" symbols
- A set of ways of forming "meaningful" compositions using these symbols.

The difference between Natural Languages and Existing Computer Languages

A large knowledge base contains (representation of) knowledge about an application domain and knowledge of how to perform a task relevant to the application domain. To deserve their name, the knowledge bases have to be endowed with semantics - i.e. with an account of what their contents say about the application domain, as well as with appropriate inference mechanisms compatible with this account.

Representation schemes: internal language of mental representations, spoken languages, written languages are built incrementally. Each one having its own *plane*, having what we call a "concept base". Each succeeding plane, provides a mapping for *concepts* from the earlier plane; in addition to that it has its own "vernacular" concepts.

It follows from this that the computer plane, if it has to be truly representational, should have the mappings of *concepts* from all the three planes in addition to its own *concepts*.

Thus if we want a formalism to represent 'knowledge' on computers, it need not and should not start from scratch, inventing its own vocabulary, but instead (1) provide a mapping for the concepts that are there in the previous planes, (2) provide a mechanism to form descriptions

Why do we say that existing computer languages have limited semantics?

Semantics is concerned with the relation between the representation and the world being modelled. The representation should not be limited only to the *concepts* in its own plane. We feel that is what database languages do! They operate in the world of computers which is isolated with its own vocabulary. We can say they have limited 'semantics'.

What needs to be done?

We strongly feel, that natural language should be used not only at the interface level, but at the representation level as well.

Concepts from natural languages should be the building blocks for computer-based systems as well. Concepts are not isolated units but inter-related. We should provide enough explicit knowledge about concepts to make them unique. Thus there is a need to build a concept-base for computer-based systems.

How to build the Concept Base

Language builds our conceptual frame of mind.

The conceptualizations provided by a language essentially provide a framework for the language-using community for information representation, acquisition and retrieval.

Most of the words in a language essentially provide 'names' for the *concepts*. Words in the vocabulary of a language are more or less expressions standing for individual concepts.

Words of a language are relatively stable, universal, meaningful and atemporal units. They represent *concepts* for the community that uses them. They make the social interaction possible among people. However just a word is not enough to represent a concept in a computer. A word in general can have many meanings. Thus we restrict its sense by pinning it down by providing extra dimensions.

A word in the language, in particular category, in particular domain, in particular plane and having a primitive word associated with it, denotes a concept.

concept = word, category, domain, plane, primitive, basic

Where word is a symbol from natural language that stands for this concept.

Category corresponds to the grammatical category of English (noun, adjective, verb, adverb etc.) corresponding to this concept.

Domain is the subject in which this concept is defined. Examples of domains are mathematics, biology,..

Plane is the level at which it is defined, "physical" or "mental" or "discourse" or "computer".

Primitive is one of the partitions the concept is put into. We use Aristotelian primitives (with a few extra primitives) as a partitioning mechanism for concepts. These partitions are Act, Person, Object, entity, State, Happening, Theme, Information, Time, Space, Property, Quality. The primitives we have chosen are hierarchically ordered under one of these partitions.

Basic concepts are the concepts a normal adult is familiar with. Basic concepts are the concepts universally known and generally have surface words representing them in a natural language. Our cross-linguistic study has helped us identify many characteristics of natural languages. We have taken as basic concepts, concepts associated with words which have equivalent words in all 14 Indian languages and English.

Composition of Descriptions

Use of words for concepts makes it possible to give a social meaning to concepts. Communicating through language is an optimization technique. With the help of the stable units(words), a mechanism to compose structures dynamically using an agreed upon convention(grammar), and associating agreed upon meanings(roles) to the compositions, human beings are able to generate infinite descriptions using finite means.

We must provide this compositionality for descriptions on computer plane.

Problems with a Natural Language

Natural languages have several devices for putting words into meaningful combinations. The three most important ones are word order, function words, and infections. Words fall into different *categories*.

Using natural language in computer systems presents many problems. The problems can be characterized as problems due to ambiguity, problems due to fuzziness of symbols, problems with context sensitivity and problems with idiosyncrasies of the language (too many usages).

Problems in Parsing English

We have selected English as a speciman language for understanding natural languages.

In English, there is no clear correspondence between words, their grammatical categories, their syntactic roles(place) in a phrase or in a sentence and their functions within a phrase or within a clause.

Words in general have many meanings. Some words can belong to different grammatical categories. We have already mentioned that a sentence can be considered as consisting of foreground description and background where foreground has roles (S,V,O,A or C). The syntactic roles taken by elements in the foreground can basically be found by the word order, since we know they follow the order : S, V, Oi, Od, (A or C). However, the positions of these roles are not absolute. Each of these the roles, in turn, can be a multi-word phrase with one or more clause associated with it. In the absence of syntactic markers for the roles and separators and linkers for parts of speech, processing English mostly depends upon human beings ability to make 'sense' out of the construction.

The problem of getting roles of the constituent phrases is difficult, because in a sentence, the structure is flattened. There is a mixing of boundaries. A participant in a sentence can be a head word, which has a part to play as one of the roles in a sentence, or it is a modifier to one of the head words. Most often, it is possible to get the roles of the participants correctly if the modifying symbols belong only to one grammatical category(adjectives in case of nouns and adverbs in case of verbs) and therefore can be recognized syntactically.

Necessity for Streamlining English

In spite of these problems with natural languages, we feel that in general, a natural language is a relatively efficient and accurate encoding of the information it conveys. What makes it difficult to accept as a semantic theory is "ambiguity". However, ambiguity is not a feature of a language; rather it is a side-effect. Whenever possible, language makes an effort to differentiate between different meanings.

We hypothesize that

The primary function of language syntax is to help in conveying the meaning of the sentence. Thus many of the so-called peculiarities, can be traced down to efforts at disambiguating the meanings.

Some of the peculiarities can be considered as high level(multiple word) patterns, which become stable units in the language just as words have become and they should be treated like words. For examples, phrasal verbs and phrases like 'in spite of'.

A few peculiarities are due to historic reasons. We have no explanation for them, and we need not stick to them. Irregular forms for past tense and past participles of the verbs again can be looked upon as techniques, to keep the word syntactically close to its base form, by keeping its consonants more or less same, and by changing its vowels. By doing so the length of a word is kept the same by avoiding the use of suffix '-ed' or '-en'.

Thus, natural language has mechanisms to make a sentence unambiguous for *human beings.* Human beings tend to choose the meaning that makes *sense* by considering, along with the syntax, the overall pattern, meanings of participating words, their categories, context etc. However, if we have to use the language for humans as well as machines, we feel that many of the sentences which are unambiguous for human beings may appear ambiguous for computers.

We postulate here that

Streamlining and disciplining natural language can make it a good semantic language. The requirement of compositionality can be met if the syntax of a Natural language can be used for semantic compositions in the streamlined language.

Instead of devising an altogether new language, which people have to learn from scratch, we select an existing natural language to start with and streamline it to suit our purpose.

Streamlined English (Singlish)

We *streamline* English by providing syntactic markers for grouping, separating, linking and highlighting various elements. We also allow alternate verb forms for words that have irregular verb forms.

Punctuations and Markers

The punctuations and markers a used for streamlining are as follows:

- Clauses are separated by backslash.
- Relative clauses are enclosed between a pair of double-backslashes.
- A noun phrase always has a determiner. (We provide semantically empty determiner '@' whose only function is to separate a noun phrase.)
- In constructs using 'infinity to' , *to* is to be connected to the verb by tilde.
- In case of ambiguity, the head of a noun phrase can be marked by following it by an up-arrow.
- Connections between head and modifiers can be explicitly shown by putting 'tilde' between them.
- When a whole clause takes part in a sentence as one of the parts-of-speech, it is enclosed in back quotes.
- A part of speech can be enclosed in a square brackets.
- A verb can be marked by a 'star' in front of it.

Examples of Singlish Text

We have taken ambiguous sentences from a collection of papers from Semantic Interpretations and the Resolution of Ambiguity [3] and verified that most of the problems get eliminated. We will now rewrite some sample sentences in Singlish.

```
Examples:

put the block~[in the box] on the table.
 (The prepositional phrase
 'in the box' is linked to the block)

which years do you have cost~figures for?
 (Cost~figures is treated as a compound noun)

the old *man the boats.
  (The word 'man' is highlighted
     to show that it is a verb)

'that deer ate everything in -
     my garden' surprised me.
 (The whole clause enclosed in backquotes is
 taken as the subject in the sentence.)

the falling~block needs painting.
   (A compound noun is identified.)
```

From Singlish to Descriptions

We have written a parser to convert Singlish sentences into *descriptions* [4]. *Description* is a unit of discourse. Typically a *description* corresponds to a sentence in a natural language. The difference between a *description* and a sentence is that a description represents discourse entities in a *surface structure independent form*. In other words, the program unflattens a sentence (give it a tree structure), tries to identify a unique concept behind every word, and explicitly identifies syntactic roles of various constituents of the sentence. These syntactic structures provide the basis for analyzing meaning.

Interpret is the interpreter for Singlish that converts Singlish sentences into *descriptions* by making sentential structure explicit, by attaching roles to parts-of-speeches and by linking various clauses and other units. In *Interpret*, one particularly interesting method we has used is the *reduction method* for parts of speech ambiguation. The idea is to determine the roles of concepts and identifiers (open class words) in a sentence using closed-class words like conjuncts, determiners, prepositions, pronouns, question marks, auxiliary verbs and relative clause words. We take into account the absolute as well as relative positions of words. We also take into account the type of the word (noun, adjective, adverb or verb). The problem arises when the same word is for a noun as well as adjective or noun as well as verb and so on. We consider each such combination, and look for the cues in the surrounding words to get its unique meaning. This goes in parallel, and resolution of ambiguity in one pair helps resolution of ambiguity in the others. For example, if it is a sentence with a single clause, and the main verb is identified, then some other word which can be either a noun or a verb will be assigned the category *noun*.

It is beyond the scope of this paper to describe other aspects of the system. For details refer to [5].

Knowledgeable Schema

When any general purpose query is made using Singlish, it is converted into a *description*. This *description* can find corresponding information using the schema if

- A language for lexical phrases to be used to represent the database and the query language both have the *same* interpretational base.
- the intensions of various data structures are explicitly specified.

Example of a Schema using Singlish

Let us assume here a simple information retrieval scenario where an answer to a question exists as one of the descriptions in the system. The problem is to find which description corresponds to the question asked. The input description(query) has to be mapped onto an internal description. The simple-minded solution will be syntactically matching words from the input description to the words in the internal description. However, this is not enough as every word in the question cannot be a *cue*. In general, the matching words in the query provide the *referent* of the query: the topic or subject of the query. Thus they will only give a partial match. The only other thing that can provide *cue* to further matching is what is called *intension* of the query.

In order to have a general-purpose retrieval system, intensions of queries should be matched with intensions of internal data descriptors. Thus it is necessary to in a knowledge-based system to capture intensions of internal descriptors at various levels. Therefore, intensions of data structures should be formed out of the common (generic or public) knowledge to make a query answerable.

Here we give some example of a *knowledgeable schema*. (the items in braces correspond to primitives which are associated with the concepts.)

Given an explicit data model, the system is able to get for each phrase the *concepts* corresponding to the *head* words of the phrases.

A typical noun phrase has a head word and a few modifiers. In absence of an exact match semantic distances among words can be used to guide the queries. Semantic distances among phrases can be composed out of semantic distances among the words of the phrases.

```
[library management] is the THEME
[reader] is  an OBJECT
ROLE of [reader] is [Borrower]
ATTRIBUTES are
  [name of the reader] : (NAME PERSON )
  [address of the reader] : (PLACE PERSON)
  [city in which the reader lives] : (PLACE PERSON)
  [priority of the reader] : (SCALE PERSON)
  [money deposited by the reader] : (MONEY PERSON)
  [category of the reader] : (SCALE PERSON)
-------------------------------------------
[book] is an OBJECT
ATTRIBUTES are
  [name of the book] : (NAME BOOK)
  [edition of the book] : (SCALE BOOK)
  [name of the author of the book] : (NAME PERSON)
  [accession number of the book]: (IDENTIFICATION BOOK)
  [price of the book] : (MONEY BOOK)
  [category of the book] : (SCALE BOOK)
  [name of the publisher of the book] :(NAME COMPANY BOOK)
  [agent for buying the book] : (NAME COMPANY BOOK)
  [number of pages in the book]: (NUMBER BOOK)
---------------------------------------------
[borrow] is an ACT
FORM of [borrow] is [reader borrows a book]
OBJECTs involved are (reader book)
ATTRIBUTEs are
  [name of the reader] : (NAME PERSON)
  [name of the book] : (NAME BOOK)
  [accession number of the book] : (IDENTIFICATION BOOK)
  [date of borrowing the book] : (DATE BOOK)
-----------------------------------------------
[recommend] is an ACT
FORM of [recommend] is [reader recommends a book]
OBJECTs involved are (reader book)
ATTRIBUTES  are
  [name of the reader] : (NAME PERSON)
  [name of the book] : (NAME BOOK)
----------------------------------------
```

Let us now see how the system will try to guess the 'most appropriate data element' for the following query.

Who has written the book 'Algorithms and Complexity' ?

The query contains the name of a book and a person is to be searched *in connection* with the book. The act mentioned in the question is *write*.

From the decriptions in data model we gather the following:

- There are four descriptions in the data model, two describing OBJECTs and two describing ACTs.
- The word 'write' or its synonym is nowhere in the head of any of the data descriptions.
- A structure (name person) appears in all the four descriptions.
- A structure (name book) appears in second, third and fourth description.
- Thus second, third and fourth descriptions have (name person)as well as (name book) specified in them.

The problem is to get the person associated with the *book*. The modifiers or ACTs associated with (name person) should be such that they match the word *write* as closely as possible.

Thus further analysis of the phrases is required.

In second relation, the person is (author of the book). In third relation, the person is involved in the ACT of borrowing. In the fourth relation, the person is involved in the ACT of recommending. If we compare the sematic distances between the words (write and author), (write and borrow) and (write and recommend) we will find that (write and author) are the closest. Thus (author of the book) in the second data description is the preferred data element for the query.

Conclusion

We strongly believe that the language for data descriptions as well as for query should be based on the conceptual basis of natural languages to make it a standard language(*interlingua*). If idiosyncrasies of surface structures of a language are removed and then it is used for knowledge representation, many other interfaces are possible for the same knowledge reservoire. Knowledge representation will be simpler to understand. Browsing through the discourse written in the natural language like descriptions will be natural. Streamlining essentially adds a layer of extra markers.

We feel that people, not trained in its usage, will not have much difficulty in adopting Singlish.

As it is, learning English becomes a burden, as with every new word, one doesn't have to just know its meaning, but also its spelling and its pronunciation. Phonetic English can be a step in the right direction for computer systems of future. Phonetic English can be based on the roman alphabets; spellings and pronunciations having a one-to-one mapping as in the case for Indian languages.

One important factor that should be taken into account, while describing data descriptors is their semantic content, which can be used to guide queries. In any knowledge representation formalism we observe that data is divided into two kinds - data description and data value. The intension of the data - data descriptions - should provide *handles* to lift the data.

Some problems with natural languages seem to be universal like multiple meanings for the words, multiple categories for the same word, optimization of usage by taking *short-cuts*, omission of the details that can be derived from the situational context and flattening of the sentence structure. We have attempted to remove the surface language barrier from the knowledge-driven systems. When idiosyncrasies of surface structures of the language English are removed and it is used for representation, various other *gateways* like Sgerman, Shungerian can be made available to make the concept-base available to a larger community. Canonical and unambiguous representation of knowledge with flexible input-output gateways is crucial for the world that hosts as many as 2500 languages.

Acknowledgements

The authors wish to thank NCST for the facilities to carry out this work. This work also forms part of the first authors Ph.D. work at Birla Institute of Technology and Science, Pilani, India.

References

[1] Noam Chomsky, *Logical Structure of Linguistic Theory*, University of Chicago Press., 1975.

[2] Shrobe Howard Davis Randall and Szolovits Peter, 'What is a knowledge representation', *AI Magazine*, (Spring 1993. 1993).

[3] G. Hirst, *Semantic interpretation and the resolution of ambiguity Studies in Natural language processing*, Cambridge Univ. Press., 1987.

[4] Alka Irani, 'Documentation on ciba interpreter'. An internal Memo, NCST, 1995., 1995,.

[5] Alka Irani, *A Unified Model for Concept Structuring*, Ph.D. dissertation, Birla Institute of Technology and Science, Pilani, Rajasthan, India, 1995,.

[6] Israel D. J., 'The role of logic in knowledge representation', *IEEE Computer*, **16**(10), (1983).

[7] John in McCarthy, *ACM Turing award lectures:First Twenty years,1966-85*, Association for computing machinery., 1971.