SEMANTIC FOR TEXT PROCESSING

Jean-Guy Meunier
Université du Québec à Montréal

Summary:

Computer text processing is defined formally as a ordered set of sentences on which various interpretative functions operate to produce a transformed text. Semantic is here understood a the set of these functions and their ordering. A common language is suggested both for the sentences and the functions. A case study is presented.

1. Text interpretation

If a computer was to read and understand a sentence such as "Women love bachelors" how would we describe the semantic process going on? A first type of answer would bring us directly in the world denoted by such a sentence: To this sentence corresponds a world situation of which one can give a formal representation for instance in set theoretical terms: there exist a certain state of affair, and a set of human beings, in which there is a subset of women and a subset of man in which the subset bachelor is itself contained and there exist a specific relation between the individuals of the subset of women with the individuals of the subset of bachelors. Hence interpreting the first sentence is thus to know the state of the world in which such a complex relation exist or for which this sentence is true.

In an other vocabulary, one could ask his data base to see if in the world representation or the frames, the scenarios, the templates, the nets, etc. there exist or can be inferred such a relation between these individuals?

In a second line of interpretation, one could stress the fact that it is impossible to set the state of affair in a world, before knowing what "love", "women", and "bachelor" means. Depending on what is contained in these expressions, one can not decide what state of affair is to be chosen. Do women love young postgraduates, students, seals without a mate, or simply young unmarried males?

Other types of interpretation will add that one cannot even decide which reading to give to the sentence if one cannot see what usage of the sentence is made. If the sentence is used in the descriptive-affirmative manner, then the preceding interpretation can be accepted, but if the sentence is use in a more rhetorical manner then the interpretation could be insulting for the feminist user and amusing for a male chauvenist!

Hence interpreting a sentence is not a simple thing to describe. Yet, theories for computer processing of natural language will often stresse only but one aspect of the this semantic process. For instance the recent trend of artificial intelligence, be it the frames paradigm (Winograd 1972, Schank Wilk 1973 or the more fregean-Montague (Schubert 1975, Cercone 1975, Lehmann 1978) insist on the necessity of a world representation for the interpretation of a sentence or a set of sentences. This is a computer variation of the tarskian semantic. In an other tradition that of the lexical semanticists (Katz, Fillmore Miller) or the semantic net theorists (Quillian, Simmons, Woods) it is maintained that a semantic grammar should mainly include a clear relation, not only between an expression of a language and the objects to which they refer in a particular usage, but also between the sense of the expression and their references.

Hence one can see different types of relations can exist in this semantic world: that is, relations between the expressions, the senses, and the objects themselves. As for the use aspect or pragmatics of the problem goes, except for a few odd explorations here and there, one relagates the whole thing for future investigation.

From the point of view of text processing distinguishing the various aspects of the semantic problems is of the highest importance for many recent projects in this field have in one sense reverse the problem. What one encounters is in fact much more conversation in natural language with a formal data base than real text interpretation. That is, given a semi-formal world representation couched in a conceptual dependency or a frames representation one system will try to relate questions to pertinent states of the world (Lehmann 1978, 1979) another system will rewrite the text amplifying it with a set of new sentences said to be presupposed in the understanding of the original

(Schank, 1972). Another will try mainly to disambiguate the original text and produce a set of adequat inferences (Wilks, 1973). In all these systems, a basic postulat is accepted and stressed: understanding can not be realised if there does not exist a minimal frame of reference on which the interpretation of the sentences of a text can rely. But real text processing at its limits although in part, has to accept this postulat must also be seen as a process of reading a text that is in itself a world representation given in natural language. A text not only describes, but also creates a world, of object and events. In other words new and old frames, world representation, data base are in the text itself. But in saying that, one becomes confused in the various world representations that are at work.

To add to the confusion the inference, the desambiguation, the paraphrasing process all rise up in the interpretation of the words and sentences meaning. Each one giving the explanation in a new vocabulary.

It is the aim of the following research to explore a more formal approach to the semantic problem of computer text interpretation. The main hypothesis could be summerize in the following manner. Semantic interpretation of text cannot lean on one unique type of reading be it referential, lexical, syntaxical or pragmatical. Semantic interpretation for texts is the establishement of a complex set of relations between the various aspects of what philosophy, linguistic, logic artificial intelligence has called reference sense, use, lexicon etc. It is also possible to offer for all these aspects a common formal descriptive language.

2. Semantic space

In order to render our explanation more intelligible we shall proceed in two related steps one non formal and the second formal.

For the first step, we shall use a metaphor. Imagine a constellation of planets. Some of these planets cannot be seen by the naked eye. Yet each planet (seen or unseen) depends for its mouvement on the existence or non existence of the other (gravity wise). What actually constitutes the constellation space is not the planets themselves but the gravity relation and movement holding them together. Exploring this metaphor we could underline various interesting properties of this space . A first dimension already stressed is the fact that a space is essentially a set of relations between planets. Without these relations there is no space. Secondly the relation is multiple that is, the path of one planet is the effect of multiple gravity relations among the planets. Thirdly, the effect of the gravity forces on one planet affects each planet itself. There is a resonance relation from one to the other. Information on all the system can be found by analysis of the gravitutional force of one planet. To put in other terms, what affects one planet's gravity affects all others. Forthly there exist a certain relativity of the dependencies, that is: each constellation of planets has its own pattern of dependencies which is different form one constellation to another.

Let us now translate our metaphor into our semantic problem. Imagine that a text is a constellation of sentences, some of which are written down on paper (the material text) others not written down. Each sentence has a set of material properties that is, they are sentences of a language with their syntax and their semantics. Some of these sentences describes the syntactic structure of original sentences of a written text others describe the sense of the sentences, others describe the state of affair, etc. Our semantic space will be filled with different sentences each of which will focus on one or other aspect of a specific sentence to be interpreted. That is each "world" in this semantics space is actually a sentence or a group of sentences of a language each of which having a different role in the overall semantic space. Hence we shall have a syntactic world, a lexical-sense world, a referential world, a natural world. Or to put in a less metaphoric language, each sentences of a language will have a specific relation to its sense its reference, its syntax etc. each of which can be expressed in sentence of a formal language.

It follows from the metaphor that our semantic space is not the sentences themselves but the relations between the sentence and only the sentences. Secondly the relation is multiple. Each sentence has many types of relations with many other sentences. A structural representation sentence can be related to a sense representation sentence and a referential representation sentence etc. Thirdly, each sentence can be modified in its own form by information comming from another sentence. For instance a sentence with variables for ambiguous words could need many type of decision before filling up the variables. Forthly the set of relations is relative to a user or a set of users. Each sub-constellations of sentences can be dependent on a set of possible users. There is a pragmatic relation between these semantic spaces and the users.

From this informal presentation we can see that the semantic space for a text is more than the constituant of the space itself (i.e. the sentences). In other words

a semantic grammar should be understood as set of relations on various information sentences of a language. Therefore we shall define informally this stage of the research a semantic grammar as a set of decision rules (functions) mapping a structured list of symbolic expressions or sentences into another list of symbolic expressions or sentences of a language. For instance a grammar could operate on a sentence such as "John has a dog" and deliver as output "John has ANIMAL" or "John POSSESS dog". Here the grammar has strictly transformed one sentence into another according to a set of decision rules. A semantic grammar is thus, a rule decision process whose domain shall be sets of informations (or sentence of a language) that has been defined in the scientific literature as syntactic structure (vg being a Noun-Phrase), lexical compotents (vg: bachelor: an unmarried man etc.), world representation (vg: to walk: Mouvement : to put one's feet in the front one's body in x y z manner), conceptual dependency (vg to "give" implies the transferring of an object money, a receiver etc), conceptual inference (vg: to sell implies that somebody buys etc.)

Because of the ambiguity of the words transformation and translation. The first operating normally only on structures and the second operating between languages of various alphabets and rules, we have decided to talk of the functions that maps one sentence into another as a transmapping function.

What is here emerging, is the fact that the interpretation of a sentence of a language is the assignment to it a whole set of rules transmapping various sentences from it, that is a sentence of a language can be transmapped into one or many other sentences each of which focuses on different aspects of the original sentence and more important where each transmapping is dependant for its existence on the role and function of the others. In that view of things, semantics would be seen not as a representation of the meaning of a sentence but a space in which differents sentences (from various or the same language) are related among themselves. The semantics of a language will thus be related among themselves. The semantics of a language will thus be related interpretations having the form of sentences of a language. Our semantic representation hence becomes a semantic space.

3. Formal definition of text semantics:

From a set theroretical point of view, a text is nothing more than a set of ordered information units (words) (sentences) that is a text could be defined as a doublet

$$\text{TEXT: } \langle W,R \rangle$$

where W is a set of sentences with an ordering relation R. Any analysis can then be thought as a transmapping function TF whose domain is a text or part of a text (a sentence) T_i and a range T_j. Hence a textual analysis function is defined as

$$TF\ (T_i) = T_j$$

From a logical point of view, a text is hence considered as a language, that is a set of primitives with an ordering relation R. And any textual analysis can be thought of as a type of translation process here called transmapping (for one can stay in the same language) that goes from a language to another language or a sentence in L_i to a sentence in L_j (where i can be=J)

Each transmapping is realized by a set of rules that are sensitive to various contextual features. Each transmap itself becomes the entry for new rules of transmapping also sensitive to various contextual conditions. From a formal point of view, each sentence original or transmapped can hence be understood as the domain of an interpretation function whose range is another set of transmap sentence and so recursively. Semantic is the ordered set of these interpretation functions.

Hence if T is a set of transmapping functions $(TF_1 \ldots TF_n)$ then a Semantic Interpretation SI defined as

$$\text{SI: } \langle T.R \rangle$$

where T is a set of functions and R an ordering relation on these functions.

As each transmap is logically considered a sentence of a language, it is possible to build for each one a specific grammar and vocabulary. But such a way to go about become highly cumbersome and lacks elegance. And in a processing perspective, a set of different formal grammar and vocabulary is not very economical. On theoretical grounds it would not also faithful to the highly recursive but conherent process of language functionning. Hence we shall try to give to all transmapped sentence a common set of primitives and rules such that there existe between each transmap a certain communality. Formally, each transmap will belong to a different sub set a common language and will have a set of common rules and lexems. Differentiation will come by the variation in this common stock.

This common language should apply also to the formulation of the transmapping functions themselves. In a sense, these functions are procedurial, declarative sentences, having specific types of predicats and variables.

Hence they should be formalized in a langage such that sometimes they will be taken at their face value. (i.e. as declarative) sometimes at their reported ("de dicto") value. Hence the transmapping function sentences can be taken as part of the text, meeting in this way the fundamental aspect of natural language recursion. In other words each transmapped sentence and transmapping function will be a sentence of a common language called TML (transmapping language). This language because of its high flexibility will includes an alphabet, a lexicon and rules of formation that allow the description of the various type of predicats, variables and constants that one encounters either in natural language or in the various semi-formal representations (semantic nets — templates — conceptual dependency theory etc). It will be in fact an intensionnal language (Montague 1974, Vanderveken 1980) so that first and second order predicat can be used as much as a formal relation between sense and reference.

As time and place does not permit do explicitate here this language we shall content ourselves with illustration of the semantic process and language.

4. Case study

Let us take the sentence given as example at the beginning of this paper. "Women love bachelors". This simple sentence can explode into a multitude of transmapped sentences S_1 to S_n

S_0: Women love bachelors

This is the rewriting of the original one.

S_1: (S(FNN(N(Women)) (Love) (N(Bachelors)))

The structured sentence S_0 in terms of a categorial grammar.

$S_{2.0}$ (All x All y ((Women x & Bachelor y) (x Love y)) OR

$S_{2.1}$ (All x Ey ((Women x & Bachelor y) (x Love y)) OR

$S_{2.2}$ (Ex Ey ((Women x & Bachelor y) (x love y)))

The quantified transmap of S_1 with the ambiguous structures.

nb. The number are only illustrative and not part of the transmap.

S_3 ((All women love all bachelors) OR (all women love some bachelors) OR (some women love some bachelor)) (some women love some bachelor))

The transmap of S_2 in natural language expression

S_4 ((Women love an (unmarried man) OR an (seal without a mate) OR an (young knight))

Transmap of S_0 & S_1 with non formalized desambiguation of bachelor.

S_5 (Women (x ESSE POSIT (QL) y) & (y ESSE POSIT (QL) x) some bachelors)

Transmap of S_0, S_1 with meaning representation of LOVE

S_6 ((Louise is a woman) &
(Kate is a woman) &
(Jane is a woman) &
(John is a bachelor) &
(John is a man) &
(Peter is a man) &
(Peter is an unmarried man) &
(Andrew is a bachelor) &
(Andrew is a man) &
(Louise loves Andrew) &
(Jane loves Peter) &
(Kate loves John))

This sentence describes the set of properties of all individuals of this small world.

As one can see the simple sentence of the original text has explodes in a multitude of new sentences. One should notice that the S_1 transmap is a purely syntactical representation; S_2 to S_5 are various transmapping for the desambiguation of the various lexical and sentential structures that sentence can have; S_6 is not directly a transmap of the original sentence but a description of the state of affair to which sentence S_0 to S_5 must relate in order to chose the right interpretation.

A better but longer description of the various interpretation would have included also the various inferences and presuppositions illocutionary forces and transmapping function that such a sentence carries. A conceptual dependency model or a semantic net representation would probably be more pedagogicly adequat but still would be logically considered another complex sentence as Schubert (1975) have shown. Also a more homogeneous language than the one here chosen would shorten up the huge proliferation of

repitition. This is one aim the TLM language presents (not illustrated here).

As one can see, the "semantics" of the original sentence is not a simple and unique representation either of its formal lexical or referential structure. All three here are working in the interpretation on the original sentence. Hence for us "semantics" will not be only the representation of the meaning of a sentence in one or the orther language of formal, lexical or referential structure but the set of relations established in among them. To interpret a sentence is here understood as a decision process that establishes specific relations between sentences. It is these semantic relations that the research tries to explore in a systematic way.

Bibliographie

Cercone, N.J. (1975). Representing Natural Language in Extended Semantic Network. Techn. Report. TR75.11. Depart. of Comp. Science University of Alberta, Edmonton, Canada

Fillmore, C.J., (1968). The Cure, for Case in Universal in Linguistic Theory, E. Bach, R.T. Harms (eds) Holt. Rinehart and Winston Inc.

Katz, J.J., (1972). Semantic Theory, N.Y. Harper & Row.

Lehmann, H., (1978). The USL project, its objectives and status. Proceedings of the international Technical Conference, IBM center, Bari, Italy.

Montague, R., (1974). Formal Philosophy, Yale University Press, New Haven.

Quillian, M.R., (1968). Semantic Memory, in Semantic Information processing, M. Minsky (ed.) M.I.T. Press, Cambridge, Masso. p. 227-270.

Schank, R.C., (1968). Conceptual dependency: A theory of natural language understanding. Cognitive Psychology, 3, 552-631.

Simmons, R., (1973). Semantic Network: Their computation and use use for understanding English Sentences. In Computer Models of Tought and Language: R. Schank, and K. Colby (eds) Freeman, San Francisco, California, pp. 66-113.

Vanderveken, D., (1980). Some Philosophical remarks on the theory of types in intensional logic. Forth comming.

Wilks, Y., (1973). Grammar, meaning and machine analysis of natural language. Boston, Routledge and Kegan Paul, 1971.

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

Woods, W.A., (1975). What's in a link: Foundations for semantic Networks in Representation and understanding D. Bobrow, and A. Collins. (eds). Academic Press, N. York, pp. 35-82.