TOWARDS A COMPUTABLE MODEL OF MEANING-TEXT RELATIONS WITHIN A NATURAL SUBLANGUAGE

Richard Kittredge and Igor Mel'Cuk

Department of Linguistics
University < of Montreal
Montreal H3C 3J7, Canada

ABSTRACT

A computable linguistic model is proposed for relating texts to their meanings within a natural sublanguage of English (stock market reports). Oriented networks are used to represent meanings which are first established by a linguistic analysis of the paraphrase sets found in the sublanguage. Several types of correspondence rules map fragments of the semantic network onto portions of deep syntactic dependency trees in a recursive process which does not "consume" the network, but rather uses it as a kind of blueprint. Additional representation levels (not illustrated) are required to relate these trees to final texts through surface syntactic and morphological stages. Important features of this model are (1) its capacity to represent the full paraphrastic power of language within an interesting natural sublanguage, and (2) Its bidirectionality, allowing the modelling of both analysis and synthesis of texts. Implementation is planned first as a device for synthesizing stock market reports (SMRAD system), but later possible applications include translation or paraphrasing of texts from this natural domain.

I GENERAL FRAMEWORK

The mastery of a natural language is considered to amount to the ability of speakers to express a given semantic content (meaning) by a set of synonomous utterances (texts), and conversely, to extract from a given text its possible meanings. An important goal of linguistic research is therefore to develop formal models to represent the two-way passage between meaning representations and text representations. The notorious complexity of the meaning-text relation has led many researchers to introduce intermediate levels of representation in order to simplify the rule types and actual rules for this passage. Out of these considerations a particular linguistic approach has emerged, the Meaning<->Text Model Theory (Mel'fuk, f ll), which will serve as the general theoretical framework in what follows.

A full-fledged model of the meaning-text relation is required in the general solution of several practical problems in the area of language processing such as automatic translation, question answering, text synthesis and man-machine communication. This is particularly the case when our goal is to deal with natural texts as opposed to artificially constrained texts, The full paraphrastic possibilities of natural language are so rich that even in a narrow technical domain, only a powerful linguistic model can cope with the astronomically large variety of semantically equivalent forms.

II CHOICE OF NATURAL SUBLANGUAGE

Our work here has two goals. First, we are constructing an experimental device to test the general Meaning<=> Text Theory with special emphasis on verifying its consistency, filling in missing details in the rule mechanisms, etc. Second, we are orienting our model towards practical applications, intending to make it operational within a reasonable time. In order to reconcile these two goals, we have chosen to develop our model in a fragment of natural language which is simultaneously small enough to me manageable, yet rich enough to represent the language as a whole. This forces us to choose a natural sublanguage used in a very restricted semantic domain, One such sublanguage, already described linguistically, is that of English stock market reports (Kittredge, [2]), Salient features of this sublanguage are: (1) limited vocabulary, (2) restricted syntax, (3) very narrow and well-defined semantic domain, yet (4)nearly the full paraphrasing power of general English, Natural constraints on the domain are such that we can envisage making a semantic calculus of all admissable semantic patterns found in the texts under study,

III PERFORMANCE GOALS OF THE SYSTEM

The system (called SMRAD -- Stock Market Report Automation Device) is being designed to carry out the following two operations: (1) given a formally correct semantic representation of a message possible within the stock market report domain, SMRAD produces for it (ideally) all synonymous text segments likely to be found in actual stock market reports; (2) conversely, given a text segment from a natural stock market report, it produces for it all the appropriate semantic representations. However, it should be emphasized that the severe restrictions on our domain eliminate almost all cases of ambiguity which complicate the task of recognizing meanings in texts from less restricted domains.

In the short term, we plan to run our model as a device for generating paraphrastic sets from meaning representations, a completely deterministic

process. In the longer terra, we plan to invert our model, using it for text analysis as well in this sublanguage. When bi-directional, SMRAD can be used as the basis for such applied tasks as automatic translation, abstracting, paraphrasing, creative writing, etc. Experimental implementation of SMRAD as a synthesizer for short segments of English stock market reports is planned for late summer 1983, using PROLOG. In contrast to other recent research on text synthesis for this limited sublanguage [3 1 , we are not concerned with modelling the process by which semantic representations of messages" are extracted from raw numerical data, since this is a non-linguistic problem. Instead, we expect to provide a much more powerful and transportable nodel than heretofore of the paraphrase mechanism required for synthesizing stylistically interesting texts.

IV ORGANIZATION OF THE RESEARCH

In accordance with the above, our research is divided into three tasks: (1) developing a formal language for semantic representations, (2) developing formal languages for intermediate levels, (3) writing the system's rules (which function to carry out the mappings between the above representations) .

A. Semantic Representation

We illustrate- our approach with a few representative examples. Consider the following five sentences, which are typical for natural stock market reports:

(1) Bow Valley jumped 2 1/2\ to 25.
(2) Rio Algom eased J to close at 39 1/2.
(3) Abitibi was up sharply, gaining 5 to 49 1/2,
(4) IBM chalked up a 5-point gain, closing at 64.
(5) Bank of Montreal moved up smartly, adding *2 1/2* to 34.

To describe the meanings of these sentences (and all similar sentences) we propose an oriented semantic network in which nodes are labelled with the primitive semantic elements of our universe and arcs are labelled with numbers indicating the argument slots. Semantic elements are of two basic types? (a) predicates (in a broad sense) which in our sublanguage have from one to four arguments; (b) (names of) entities such as stocks, companies, sums of money, etc.: these may include variables for entities. Our network representation is basically equivalent to a formula of predicate logic, but has the advantage of perspicuity primarily because our network is free of linear order. Since sentences (1) - (5) all concern changes in the price of stocks, we introduce a 4-place predicate named CHANGE(W,X,Y,Z) whose arguments are limited in our domain to the following types:

W: the price of a stock or index manifested in the text by the name of the corresponding company
X: an initial value in dollars (for North American stocks)
Y: a final value in dollars
Z: a signed value equal to the difference Y-X,

Although the X-value ±s rarely expressed in stock market reports, It has been retained here (at the risk of introducing redundancy) in order to show how very general semantic primitives may be used in particular domains, The Z-value indicates the direction of change explicitly, although this directional component may be incorporated in the lexical meaning of a verb expressing change (e.g.,jump and move up mean 0<Z; plunge,sag mean Z<0), Verbs may also be lexically marked for degree of change (i,e. .subjective size of Z), For example, jump, plunge signify large changes; creep up, edge up indicate small changes. Figures 1 and 2 give preliminary representations for sentences (1) and (2) respectively,

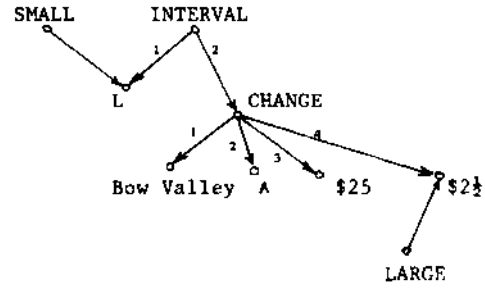

Fig.1: Provisional semantic network for sentence (1).

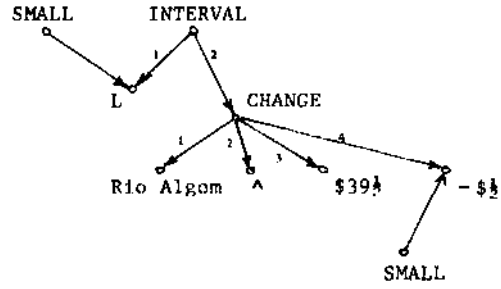

Fig.2: Provisional semantic network for sentence (2).

Two comments may be in order. First, we wee that in both Figs 1 and 2 the Z-values of CHANGE have subjective qualifications of amount (e.g., LARGE, SMALL). Moreover, the time INTERVAL over which the change takes place (I.e., the unknown quantity L) is considered SMALL. No value is given to the second argument of CHANGE (indicated by "-'') since no initial value is expressed linguistically in either example. Second, in stock market reports, company names show a systematic 4-way ambiguity. Bow Valley can refer to a company, or its spokesman, or its stock, or the price of the stock.

Our semantic representations are postulated on the basis of a systematic comparison of observable paraphrases within this sublanguage of English. We select as most "semantically transparant" that paraphrase which satisfies two conditions: (a) all relevant meanings are expressed analytically, i.e. by separate words; (b) its words have the freest occurrence in the sublanguage, i.e., they are the

least idiomatic. This semantically transparant paraphrase is converted into a semantic representation by replacing the English surface syntactic means by the labelled arcs of our network. The resulting representation must meet a number of general formal requirements: (1) the formal representation language must be rich enough to allow assigning two different representations to two sentences which are felt intuitively to have different meanings; (2) representations must reflect the additivity and compositionality of meanings, etc. Our use of paraphrase as an approach to deriving semantic representations is based on work by Kittredge [A].

B. Intermediate Levels of Representation

We restrict ourselves here to the deep syntactic level, which we will not attempt to justify, Surface syntactic and morphological levels are even less debatable and farther from the semantic problems which are our main concern. Figures 3 and A give two possible deep syntactic realizations for sentences whose content is represented by the network of Figure 1.

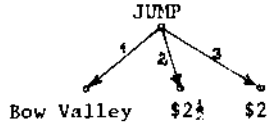

Fig.3

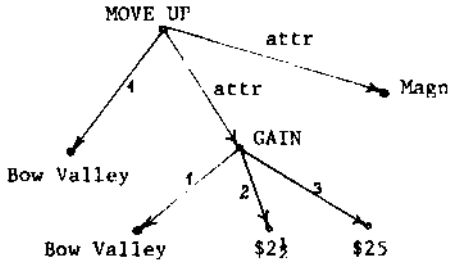

Fig.4

We use unordered dependency trees in which arcs are labelled with syntactic relations of two different types. First, we use a predicative relation between a term and its subject and objects or complements. Second, we use an attributive relation, which covers all kinds of attributes, modifiers and adverbials. This labelling is used to constrain the ordering and inflecting of lexical items at the superficial syntactic level (not shown here). Nodes of trees are marked with English lexical expressions or lexical functions such as Magn in Fig.4 (see Mel'fcuk [5]) .

C. Correspondence Rules

The SMRAD system uses a variety of rules for relating semantic networks to deep syntactic dependency trees. These are organized into two main components: (1) lexical rules, and (2) syntactic rules.

Lexical rules are of two types: (I) a lexical correspondence rule associates an English lexeme, along with its syntactic government pattern and some semantic labels, with a subnetwork of a semantic network representing the sentence meaning. Fig. 5 represents the lexical correspondence rule for jump.

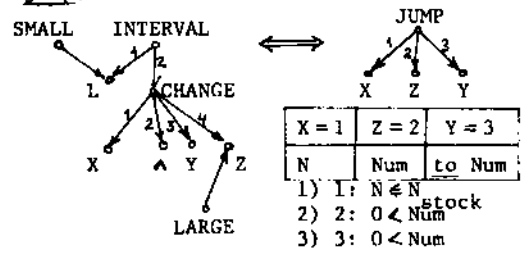

Fig.5: Lexical correspondence rule for JUMP, determining the relation between semantic subgraph and deep syntactic tree.

Within the boxes on the right we give any restrictions on the syntactic form of the arguments in the syntactic dependency tree. Any semantic constraint (e.g., reference to semantic word classes) is given under the boxes: (II) a lexical function rule introduces a lexical function, such as the intensifier function, as the syntactic correlate of a semantic subnetwork.

The syntactic rules specify how to put the whole syntactic structure of the sentence together. They are also of two types: (I) general rules give global principles for encoding meanings in the deep syntactic structure (e.g., "if one verb is used to express an argument of a second verb, the first becomes a syntactic attribute of the second'). (H)meta-rules give general principles for constructing a syntactic tree for the English sentence on the basis of its semantic network. It is important to state here that the network is not "consumed" in the process. Instead, it is used as a "blueprint" for the rules which construct the tree. Items already 'used" in the network are checked off in a recursive process until all parts of the network have been encoded in the syntactic tree. Logically speaking, the rules which do this are not ordered.

REFERENCES

[1] Mel'uk.I, "Meaning-Text Models" Ann.Rev.Anthropoid 10 (1981) pp. 27-62.

|2] Kittredge.R. "Variation and Homogeneity of Sublanguages'' Sublanguage (Kittredge & Lehrberger, eds.) de Gruyter, 1982a.

[3] Kukich,K. "Design of a Knowledge-Based Report Generator" Proceedings of the 21st ACL meeting, 1983.

[4] Kittredge.R. "Semantic Processing of Texts in Restricted Sublanguages" Computers and Mathemat. with Applications 8, 1982b.

[5] Mel'cuk,I. "Lexical Functions in Lexicographic Description" Proceedings of the 8th Annual Meeting of the Berkeley Linguistic Society,1982.