

SEMANTIC RELEVANCE AND ASPECT DEPENDENCY IN A GIVEN SUBJECT DOMAIN
Contents-driven algorithmic processing of fuzzy wordmeanings
to form dynamic stereotype representations

Burghard B. Rieger

Arbeitsgruppe für mathematisch-empirische Systemforschung (MESY)
German Department, Technical University of Aachen,
Aachen, West Germany

ABSTRACT

Cognitive principles underlying the (re-)construction of word meaning and/or world knowledge structures are poorly understood yet. In a rather sharp departure from more orthodox lines of introspective acquisition of structural data on meaning and knowledge representation in cognitive science, an empirical approach is explored that analyses natural language data statistically, represents its numerical findings fuzzy-set theoretically, and interprets its intermediate constructs (stereotype meaning points) topologically as elements of semantic space. As connotative meaning representations, these elements allow an aspect-controlled, contents-driven algorithm to operate which reorganizes them dynamically in dispositional dependency structures (DDS-trees) which constitute a procedurally defined meaning representation format.

0. Introduction

Modelling system structures of word meanings and/or world knowledge is to face the problem of their mutual and complex relatedness. As the cognitive principles underlying these structures are poorly understood yet, the work of psychologists, AI-researchers, and linguists active in that field appears to be determined by the respective discipline's general line of approach rather than by consequences drawn from these approaches' intersecting results in their common field of interest. In linguistic semantics, cognitive psychology, and knowledge representation most of the necessary data concerning lexical, semantic and/or external world information is still provided introspectively. Researchers are exploring (or make test-persons explore) their own linguistic/cognitive capacities and memory structures to depict their findings (or let hypotheses about them be tested) in various representational formats (lists, arrays, trees, nets, active networks, etc.). It is widely accepted that these modelstructures do have a more or less ad hoc character and tend to be confined to their limited theoretical or operational performances within a specified approach, subject domain or implemented system. Basically interpretative approaches like these, however, lack the most salient characteristics of more constructive modelstructures that can be developed along the lines of an entity-relationship approach (CHEN 1980). Their properties of flexibility and dynamics are needed for automatic meaning representation from input texts to build up and/or modify the realm and scope of their own knowledge, however baseline and vague that may appear compared to human understanding.

In a rather sharp departure from those more orthodox lines of introspective data acquisition in meaning and knowledge representation research, the present approach (1) has been based on the algorithmic analysis of discourse that real speakers/writers produce in actual situations of performed or intended communication on a certain subject domain, and (2) the approach makes essential use of the word-usage/entity-relationship paradigm in combination with procedural means to map fuzzy word meanings and their connotative interrelations in a format of stereotypes. Their dynamic dependencies (3) constitute semantic dispositions that render only those conceptual interrelations accessible to automatic processing which can - under differing aspects differently - be considered relevant. Such dispositional dependency structures (DDS) would seem to be an operational prerequisite to and a promising candidate for the simulation of contents-driven (*logically-associative*), instead of formal (*logically-deductive*) inferences in semantic processing.

1. The approach

The empirical analysis of discourse and the formal representation of vague word meanings in natural language texts as a system of interrelated concepts (RIEGER 1980) is based on a WITTGENSTEINian assumption according to which a great number of texts analysed for any of the employed terms' usage *regularities* will reveal essential parts of the concepts and hence the meanings conveyed.

It has been shown elsewhere (RIEGER 1980), that in a sufficiently large sample of pragmatically homogeneous texts, called *corpus*, only a restricted vocabulary, i.e. a limited number of lexical items will be used by the interlocutors however comprehensive their personal vocabularies in general might be. Consequently, the lexical items employed to convey information on a certain subject domain under consideration in the discourse concerned will be distributed according to their conventionalized communicative properties, constituting *semantic regularities* which may be detected empirically from the texts.

For the quantitative analysis not of propositional strings but of their elements, namely words in natural language texts, rather simple statistics serve the basically descriptive purpose. Developed from and centred around a correlational measure to specify intensities of co-occurring lexical items used in natural language discourse, these analysing

algorithms allow for the systematic modelling of a fragment of the lexical structure constituted by the vocabulary employed in the texts as part of the concomitantly conveyed world knowledge.

A correlation coefficient appropriately modified for the purpose has been used as a mapping function (RIEGER 1981a). It allows to compute the relational interdependency of any two lexical items from their textual frequencies. Those items which co-occur frequently in a number of texts will positively be correlated and hence called *affined*, those of which only one (and not the other) frequently occurs in a number of texts will negatively be correlated and hence called *repugnant*. Different degrees of *word-repugnancy* and *word-affinity* may thus be ascertained without recurring to an investigator's or his test-persons' word and/or world knowledge (*semantic competence*), but can instead solely be based upon the usage regularities of lexical items observed in a corpus of pragmatically homogeneous texts, spoken or written by real speakers/hearers in actual or intended acts of communication (*communicative performance*).

2. The semantic space structure

Following a system-theoretic approach and taking each word employed as a potential descriptor to characterize any other word's virtual meaning, the modified correlation coefficient can be used to map each lexical item into fuzzy subsets (ZADEH 1981) of the vocabulary according to its numerically specified usage regularities. Measuring the differences of any one's lexical item's usages, represented as fuzzy subsets of the vocabulary, against those of all others allows for a consecutive mapping of items onto another abstract entity of the theoretical construct. These new operationally defined entities - called an item's *meanings* - may verbally be characterized as a function of all the differences of all regularities any one item is used with compared to any other item in the same corpus of discourse.

UNTERNEHM/enterpr	0.000		
SYSTEM/system	2.035	LEIT/guide	2.113
ELEKTR/electron	2.195	COMPUTER	2.208
DIPLOM/diploma	2.288	VERBAND/assoc	2.279
INDUSTR/industry	2.538	STELLE/position	2.620
SUCHE/search	2.772	SCHREIB/write	2.791
SCHUL/school	2.922	AUFTRAG/order	3.058
FOLGE/consequ	3.135	BERUF/professn	3.477
ERFAHR/experienc	3.485	UNTERR/instruct	3.586
ORGANISAT/organis	3.846	VERWALT/administ	3.952
GEBIET/area	4.055	WUNSCH/wish/desir	4.081
...		...	

Table 1: Topological environment E<UNTERNEHM>

The resulting system of sets of fuzzy subsets constitutes the *semantic space*. As a distance-relational datastructure of stereotypically formatted meaning representations it may be interpreted topologically as a hyperspace with a natural metric. Its linguistically labelled elements represent *meaning points*, and their mutual distances represent *meaning differences*.

The position of a meaning point may be described by its semantic environment. Tab.1 shows the *topologi-*

cal environment E<UNTERNEHM>, i.e. those adjacent points being situated within the hypersphere of a certain diameter around its center meaning point UNTERNEHM/enterprise as computed from a corpus of German newspaper texts comprising some 8000 tokens of 360 types in 175 texts from the 1964 editions of the daily DIE WELT.

Having checked a great number of environments, it was ascertained that they do in fact assemble meaning points of a certain semantic affinity. Further investigation revealed (RIEGER 1983) that there are regions of higher point density in the semantic space, forming clouds and clusters. These were detected by multivariate and cluster-analyzing methods which showed, however, that the both, paradigmatically and syntagmatically, related items formed what may be named *connotative clouds* rather than what is known to be called *semantic fields*. Although its internal relations appeared to be un-specifiable in terms of any logically deductive or concept hierarchical system, their elements' positions showed high degree of stable structures which suggested a regular form of contents-dependant associative connectedness (RIEGER 1981b).

3. The dispositional dependency

Following a more semiotic understanding of meaning constitution, the present semantic space model may become part of a word meaning/world knowledge representation system which separates the format of a basic (stereotype) meaning representation from its latent (dependency) relational organization. Whereas the former is a rather static, topologically structured (associative) memory representing the data that text analysing algorithms provide, the latter can be characterized as a collection of dynamic and flexible structuring processes to re-organize these data under various principles (RIEGER 1981b). Other than declarative knowledge that can be represented in pre-defined semantic network structures, meaning relations of lexical relevance and semantic dispositions which are heavily dependent on context and domain of knowledge concerned will more adequately be defined procedurally. i.e. by generative algorithms that induce them on changing data only and whenever necessary. This is achieved by a recursively defined procedure that produces hierarchies of meaning points, structured under given aspects according to and in dependence of their meanings' relevancy (RIEGER 1984b).

Corroborating ideas expressed within the theories *spreading activation* and the process of *priming* studied in cognitive psychology (LORCH 1982), a new algorithm has been developed which operates on the semantic space data and generates - other than in RIEGER (1982) - dispositional dependency structures (DDS) in the format of n-ary trees. Given one meaning point's position as a start, the algorithm of least distances (LD) will first list all its neighbouring points and stack them by increasing distances, second prime the starting point as head node or root of the DDS-tree to be generated before, third, the algorithm's generic procedure takes over. It will take the first entry from the stack, generate a list of its neighbours, determine from it the least distant one that has already been primed, and identify it as the ancestor-node to

which the new point is linked as descendant-node to be primed next. Repeated successively for each of the meaning points stacked and in turn primed in accordance with this procedure, the algorithm will select a particular fragment of the relational structure, latently inherent in the semantic space data and depending on the aspect, i.e. the initially primed meaning point the algorithm is started with. Working its way through and consuming all labeled points in the space structure - unless stopped under conditions of given target nodes, number of nodes to be processed, or threshold of maximum distance - the algorithm transforms prevailing similarities of meanings as represented by adjacent points to establish a binary, non-symmetric, and transitive relation of semantic relevance between them. This relation allows for the hierarchical re-organization of meaning points as nodes under a primed head in an n-ary DDS-tree (RIEGER 1984a).

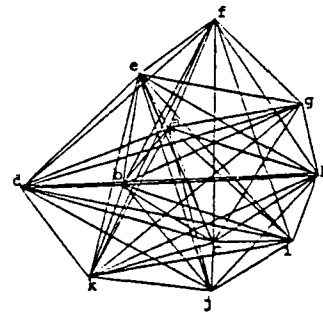


Fig. 1.1

Without introducing the algorithms formally, some of their operative characteristics can well be illustrated in the sequel by a few simplified examples. Beginning with the schema of a distance-like data structure as shown in the two-dimensional configuration of 11 points, labeled a to k (Fig. 1.1) the stimulation of e.g. points a or c will start the procedure and produce two specific selections of distances activated among these 11 points (Fig. 1.2). The order of how these particular distances are selected can be represented either by step-lists (Fig. 1.3), or n-ary tree-structures (Fig. 1.4), or their binary transformations (Fig. 1.5). It is apparent that stimulation of other points within the same configuration of basic data points will result in similar but nevertheless differing trees, depending on the aspect under which the structure is accessed, i.e. the point initially stimulated to start the algorithm with.

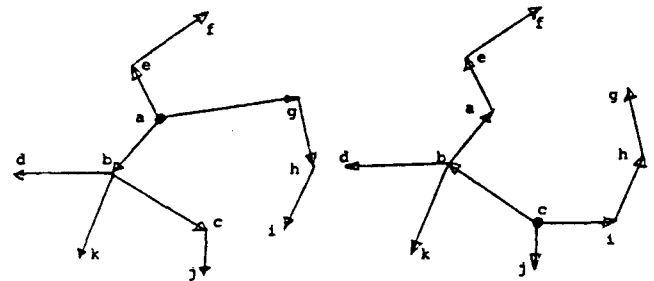


Fig. 1.2

Step	z_d	z_a
0		a → a
1		e → a
2		b → a
3		c → b
4		f → e
5		g → a
6		d → b
7		h → g
8		i → h
9		k → b
10		j → c

Step	z_d	z_a
0		c → c
1		j → c
2		i → c
3		b → c
4		h → i
5		k → b
6		a → b
7		g → h
8		d → b
9		e → a
10		f → e

Fig. 1.3

Applied to the semantic space data of 360 defined meaning points calculated from the textcorpus of the 1964 editions of the German newspaper *DIE WELT*, the *Dispositional Dependency Structure (DDS)* of UNTERNEHM/enterprise is given in Fig. 2 as generated by the procedure described.



Fig. 1.4

Beside giving distances between nodes in the DDS-tree, a numerical measure has been devised which describes any node's degree of relevance according to that tree structure. As a numerical measure, a node's *criticality* is to be calculated with respect to its root or aspect and has been defined as a function of both, its distance values and its level in the tree concerned. For a wide range of purposes in processing DDS-trees, different criticalities of nodes can be used to estimate which paths are more likely being taken against others being followed less likely under priming of certain meaning points. Source-oriented, contents-driven search and retrieval procedures may thus be performed effectively on the semantic space structure, allowing for the activation of dependency paths. These are to trace those intermediate nodes which determine the associative transitions of any target node under any specifiable aspect.

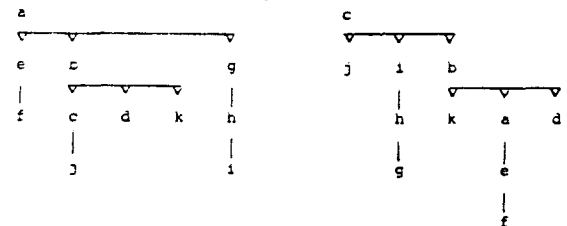


Fig. 1.5

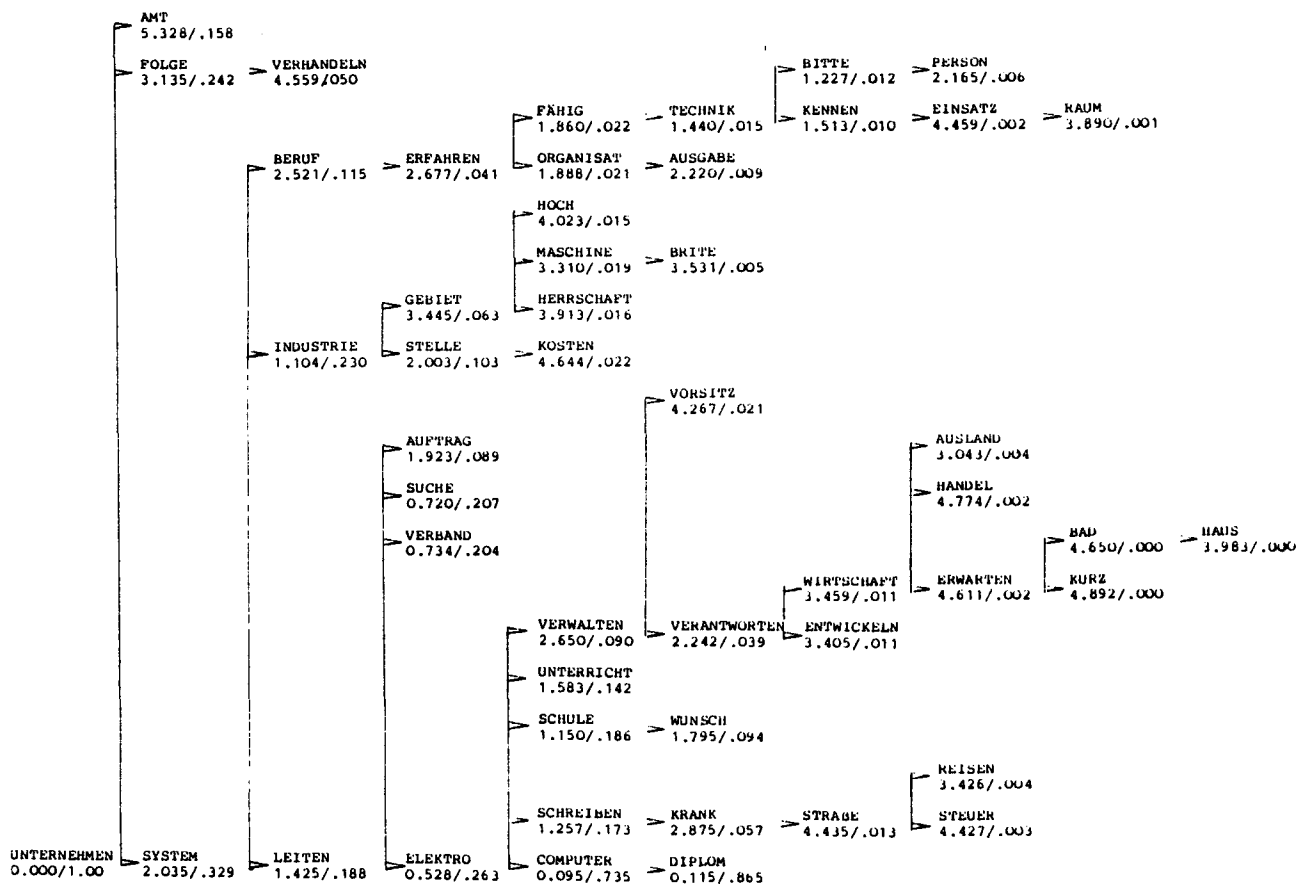


Fig. 2

Using these tracing capabilities within DDS-trees proved particularly promising in an analogical, contents-driven form of automatic inferencing which - as opposed to logical deduction - has operationally be described in RIEGER (1984c) and simulated by way of parallel processing of two (or more) dependency-trees.

REFERENCES

Chen, P.P. (1980) (Ed.): Proceedings of the 1st Intern. Conference on Entity-Relationship Approach to Systems Analysis and Design (UCLA), Amsterdam/New York (North Holland) 1980

Lorch, R.F. (1982): Priming and Search Processes in Semantic Memory: A Test of Three Models of Spreading Activation. *Journ. of Verbal Learning and Verbal Behavior* 21 (1982) 468-492

Rieger, B. (1980): Fuzzy Word Meaning Analysis and Representation, Proceedings of COLING 80, Tokyo 1980, 76-84

Rieger, B. (1981a): Feasible Fuzzy Semantics. Eikmeyer/Rieser (Eds.): Words, Worlds, and Contexts. New Approaches to Word Semantics, Berlin/ New York (deGruyter) 1981, 193-209

Rieger, B. (1981b): Connotative Dependency Structures in Semantic Space. in: Rieger (Ed.): Empirical Semantics II, Bochum (Brockmeyer) 1981, 622-711

Rieger, B. (1982): Procedural Meaning Representation. in: Horecky (Ed.): COLING 82. Proceedings of the 9th Intern. Conference on Computational Linguistics, Amsterdam/New York (North Holland) 1982, 319-324

Rieger, B. (1983): Clusters in Semantic Space. in: Delatte (Ed.): Actes du Congrès International Informatique et Sciences Humaines, Université de Liège (LASLA), 1983, 805-814

Rieger, B. (1984a): Semantische Dispositionen. Prozedurale Wissensstrukturen mit stereotypisch repräsentierten Wortbedeutungen. in: Rieger (Ed.): Dynamik in der Bedeutungskonstitution, Hamburg (Buske) 1983 (in print)

Rieger, B. (1984b): Inducing a Relevance Relation in a Distance-like Data Structure of Fuzzy Word Meaning Representation. in: Allen, R.F. (Ed.): Data Bases in the Humanities and Social Sciences (ICDBHSS/83), Rutgers University, N.J. Amsterdam/New York (North Holland) 1984 (in print)

Rieger, B. (1984c): Lexikalische Relevanz und Semantische Disposition. in: Hoppenbrouwers/Seuren/Weijters (Eds.): Meaning and the Lexicon. Nijmegen University (M.I.S. Press) 1984 (in print)

Zadeh, L.A. (1981): Test-Score Semantics for Natural Languages and Meaning Representation via PRUF. in: Rieger (Ed.): Empirical Semantics I, Bochum (Brockmeyer) 1981, 281-349