Recent Model-Based and Model-Related Studies of a Large Scale Lexical Resource [Roget's Thesaurus]

Sally Yeates Sedelow
Professor, Computer Science,
University of Arkansas (UA)/Little Rock;
Adjunct Professor, Electronics and Instrumentation,
UA/Graduate Institute of Technology
USA

Walter A. Sedelow, Jr.
Professor, Computer Science, University of
Arkansas (UA)/Little Rock;
Adjunct Professor, Electronics and Instrumentation,
UA/Graduate Institute of Technology; Adjunct
Professor, UA/College of Medicine
USA

*(This research is supported by Grant *IRI-9114068 from The National Science Foundation.)*

In an era when knowledge, if not king, is certainly an equal partner with the methodologies directed toward such eminently desirable goals as computer-based commonsense reasoning and understanding-in-general, large-scale resources such as Roget's Thesaurus self-evidently are necessary to advanced knowledge-based computational systems. In contrast to such efforts as Lenat's to recreate encyclopedic resources to fit currently popular cognitive and computational models, our research emphasis has been upon models and programs which neatly finesse re-creation by making explicit and accessible such resources as people already process effectively and use effectively. Other major research sites and groups now share this orientation, and we expect that our research team's recent work with various aspects of our model will be of particular interest to them. Specifically, we wish to report here on three different interpretations of a component of our topological model (Bryan, 1973, 1974), applied to its instantiation, Roget's International Thesaurus, 3rd edition (1962).

The model, itself, has been described extensively elsewhere (e.g., Sedelow and Sedelow, 1986, 1987). For this discussion, we need definitions of an Entry, Word, and Category, as well as of a Type 10 Chain. In the model, a thesaurus, T, is a triple <E,W,C> where

i) E is a non-null, finite set;

ii) W and C are non-null collections of subsets of E;

iii) distinct elements of W are disjoint, and distinct elements of C are disjoint;

iv) given any $e \varepsilon E$, $e \varepsilon w$ for some $w \varepsilon W$ and e εC for some $c \varepsilon C$;

v) given $w \varepsilon W$ and $c \varepsilon C$, w^$c \leq 1$.

Elements of E are called entries, elements of W are called words, and elements of C are called categories. (Bryan, 1973)

'Navigation' within the thesaurus can take two basic routes: 1. it can depend upon the explicit hierarchy, comprising seven to nine levels (depending upon how fine-grained the distinctions are); 2. it can move cross-hierarchically from one category to others. Within the second type of navigation, those cross-hierarchical 'hops' of greatest interest to us are enabled by the multiple occurrence (multilocality property) of given "words" (strings with identical spelling, which do not necessarily have the same meaning).

The model defines the cross-hierarchical form of navigation in terms of Chains, as well as, within Chain types, Stars and Neighborhoods. Chains range from Type 1, the least restricted, to Type 10, the most restricted. They consist of entries, each of which represents the intersection of a Word and a Category. As might be expected, a Type 1 Chain consists of any group of entries. At the other end of the restriction scale is the Type 10 Chain, which must be both word-strong and category-strong. Categories are said to be strongly connected if they have at least two words in common, and words are strongly connected if they have at least two categories in common. Intuitively, one sees that the convergence of words and categories in the Thesaurus represents a selection of an appropriate semantic sub-space (meaning) within a larger semantic space representing multiple meanings (ambiguities). We have spoken and written elsewhere about the application of this model to a number of natural-language computational tasks, all of which confirmed our (originally weak) belief that the Thesaurus is a quite good model of 'normal' word association patterns in English (Brady, 1988, 1991; Patrick, 1985; S. Sedelow, 1985; Sedelow and Sedelow, In Press; W. Sedelow, 1988; Warfel, 1972). It is, nonetheless, desirable to study the impact of various interpretations of the model upon the model's representational strengths (and weaknesses); that is the focus of this presentation.

Victor Jacuzzi, our graduate student, has just completed a comparison of two approaches to Type 10 Chain semantic decomposition of the Thesaurus. Both approaches isolate quartets of words (string quartets) which represent strong connections between categories and words; (in these interpretations of the Bryan model, categories are taken to be the groupings of words in the Thesaurus bounded by semicolons, the lowest level of grouping in the explicit hierarchy). For example, in the T-Graph in Figure 1, Entries 1, 2, 6, and 4 form a quartet in which Categories 2 and 5 have two Words, W_1 and W2, in common and Words W_1 and W_2 have Categories C_2 and C_5 in common.

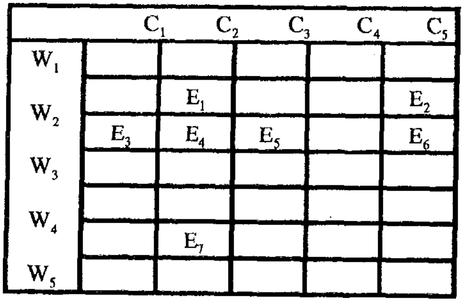

Figure 1.

In the earlier study (Talburt and Mooney, 1989, now validated by Jacuzzi, 1991), if any entry forming the quartet functioned as an entry in a second (or third, etc.) quartet, then the second quartet became a part of the Type 10 component identified by the first quartet, etc. Much of the time, this approach adequately discriminates among homographs, as well as discriminating among word senses (and parts of speech within senses). For example, examination of Jacuzzi's recent validation of the Talburt-Mooney results (Jacuzzi's validation utilized an independently developed algorithm) shows the following apropos "nosy" and related words: one component consists of "nosy," "prying," and "snoopy," all adjectives at the intersection of the meanings Intrusion (#237 in the explicit hierarchy) and Curiosity (#526); another component consists of the words "nosy," "odorous," "smelling," "smellsome," "smellful," "smelly," and "whiffy" all adjectives at the intersection of the meanings Odor (#434) and Malodor (#436). Clearly, the homograph "nosy" is separated out into distinct meanings by the algorithm. Now, to explore the second semantic subspace a little further, we find that the noun, "odorousness" (see the adjective "odorous" above) is grouped with the noun "smelliness" (again, see above) in another intersection of Odor (#434) and Malodor (#436). But yet another grouping, using Odor (#434) as a departure point, links the adjective "odoriferous" with the adjective "redolent," supplying an intersection with Fragrance (#435). Hence, the contrasting, more pleasant sense related to "odor" is also singled out by the algorithm.

Many analogous groupings could be cited as exemplifications of the utility of this approach for appropriate word sense identification. Nonetheless, when looking for the output of the validation program, it is impossible to ignore a reason for our desire to have such a validation: one enormous component comprising 22,431 entries. Although a tracing of the links among the quartets pulling all these entries together would doubtless show an associatively plausible link between each component, the sum total of these components ranges unacceptably across too many domains; obviously, discrimination of any useful sort, not to mention fine granularity discrimination, is hardly the apposite term for a group of this size.

Faced with this anomaly, Jacuzzi then proposed a restriction on the Talburt-Mooney quartet approach: henceforth, at least two words or at least two categories in the original quartet must appear in the second quartet in order for the second to be included in a component with the first. The implementation of this algorithm produced markedly different results. In both cases, as would be expected, the number of individual quartets was the same: 59,541. From this number, the original algorithm yielded 5,960 components, whereas the Jacuzzi algorithm produced 10,341 components. In the original algorithm, the seven largest components were, in ascending order: 120 entries, 134, 143, 200, 210, 229, and 22,431. By comparison, the Jacuzzi algorithm produced the following: 282, 388, 427, 469, 491, 705, and 1490.

Inspection shows that the set of the largest Jacuzzi components (all of those just listed plus others) represent 'breakouts' from the 22,431 entry component produced by the original algorithm. Jacuzzi's largest component (1490 entries) has as its largest group words encapsulating intersections of hostility, irritation, disasters, turmoil (including noise), and physical competition (as in "bout"). Smaller sets including terms having to do with, for example, direction (aim, ambit, circle, etc.) seem puzzling at first; but in this case, for example, the word "course" ties to "flood" which intersects with the disaster terms. Another small set including words having to do with "manner" and "mode" ties to words intersecting with the sense "irritation." Hence this largest Jacuzzi component is clearly explicable, although a further restriction, either on the algorithm, or on the component produced, might seem desirable for adequate selection of certain semantic subspaces. (It should be noted here that although the Thesaurus has performed remarkably well on a range of tasks and data types, we certainly don't claim that it is 'perfect.' Investigations such as this point the way to possible modifications; but, given the quality of much of the output based solely on the model and algorithms interpreting it, we strongly feel that modifications should be made with caution. Even as it stands, the Thesaurus provides a very good foundation on which to build.)

To take another example from the 'break-out' of the 22,431 entry component, the Jacuzzi output gave the following group: Geist, bosom, breast, bottom of the heart, cockles of the heart, heart, heart's core, inmost heart, heart of hearts, inmost soul, mind, secret recesses of the heart, soul, spirit. This grouping seems internally consistent, a result typical of the smaller groups as well as of some of the largest in this restricted algorithm's output; (for example, the Jacuzzi component with 388 entries was consistently concerned with the sense carried by words such as "abhorrent," "abominable," "atrocious," etc.)

Having looked at the high end of the scale, what about groupings with small numbers of entries? First, we should note the comparative numbers: for four-entry components, the Jacuzzi algorithm produced 6584 components, compared with 3372 in his validation of the other algorithm; for six-entry components, the comparison is 1789 to 925; for eight-entry components, 700 to 342; for nine-entry components, 47 to 163; for ten-entry components, 350 to 171, and for eleven-entry components, 35 to 92. Our primary concern here is whether the further restriction hurts us all in the sense that the semantic subspaces so identified are too small to be useful for information retrieval, concept extraction, etc. Although a final answer awaits renewed efforts at applications, preliminary inspection suggests that although a four-entry component won't lead us beyond two closely-related terms (remember that repeated "words" within repeated "categories" [semi-colon groups] form the strong ties giving us the Type 10 definition), at least we certainly won't be led astray. Some examples: abreast-alongside; abrade-rub off; Gaucho-vaquero; Fritz-Jerry; Zero hour-H-hour; heaven-providence; Hephaestus-Vulcan; abandon-abandonment (intersection of Freedom and Vice); abandonment-renunciation (intersection of Submission and Relinquishment); abandon-quit (intersection of Departure, Abandonment, and Insufficiency).

Referring back to Figure 1, it can be observed that entry E7 (W_4, C_2) does not form part of a quartet, and thus would not be picked up by either algorithm. But given the interpretation of C_2 as a semicolon group and given the fact that the semicolon group level provides in the explicit hierarchy the most closely related grouping of words semantically, it may well be desirable to include entry E7 in the component. Bryan's model provides for such inclusion at the Type 9 Chain level (connections must be either word-strong or category-strong) and we plan to investigate the decomposition of the Thesaurus using that point of departure. We also have begun work with lattice representations, in cooperation with Professor Dr. Rudolph Wille and his colleagues at the Technische Hochschule, Darmstadt, but that exploration is too preliminary to report on here.

Another of our graduate students, John Old, has used the concept of Type 10 chains in a way of examining, among other properties, the cross-referencing system in the Thesaurus. That is, first using output produced by the earlier of the two different 'quartet' approaches for the word "lead," he has then turned to cross-referencing to provide semantic maps showing connectedness (and lack of connectedness) among various senses of the word. His comparison of Type 10 output, cross-referencing information, and the index in the printed Thesaurus with reference to meanings of "lead" is documented in Old (1991a).

More recently, Old (1991b) has compared "over" (i) as defined associationally in the Thesaurus, (ii) as defined through "definitions" in the Oxford English Dictionary, and (iii) through "Cognitive Topology and Lexical Networks" by Brugman and Lakoff (1988). His approach to the analysis of "over" in the Thesaurus was first to identify all semicolon groups in which "over" occurs. This process resulted in twenty-two senses (nodes in the network he constructed). Links between the nodes were words repeated in two or more of the semicolon groups containing "over". Hence, in his example, the word "on" in the groups "over, on, on top of," and "over, on, upon" would form the link between the groups (Old, 1991b). When he turned to the OED definitions, the number of definitions sharing at least two words resulted in more than a thousand links; for the purposes of graphic representation he restricted the algorithm, requiring that three or more words be shared for links among nodes to occur. As to Brugman and Lakoff, he worked with the networks as provided in their report (1988).

Old's determination of the central senses in the two lexical treatments of "over" is in fact much more complicated than indicated by this brief sketch. The results, though, were reassuring in that they showed significant correlations among the three works while, at the same time, there were significant distinctions. Brugman and Lakoff identify the central sense of "over" as the combination of the "elements above and across" (1988). Interestingly enough, Old's data extraction method for the OED resulted in a central sense of "from side to side; across to." Old notes that the OED's "across to" "closely matches Brugman and Lakoff's choice of a central sense of over and is also the sense of the "across" containing semicolon group in the Thesaurus" (1991b). As that observation implies, the Thesaurus network includes the "across" and "above" interpretations; but, contrastively, the central sense in the Thesaurus is "additionality," closely followed by "excess-related."

For some applications, it may not much matter which senses of a given word are 'central'; rather, it is important to be able to place a word in an appropriate semantic space or subspace and then perhaps to see what specific ties it has to other subspaces. It is important, though, to see how interpretations of a model differentially partition semantic space -- important so as to heighten the realization that disappointments with a large-scale resource are not necessarily due to shortcomings of the resource but rather of the model or of the interpretations/implementations of the model. Too many glib assertions were made earlier about the inadequacy of the Thesaurus as well as about other large-scale resources. The experience of human users certainly would lead one to suppose that such culturally-validated large-scale resources must "be doing something right." Our own computational research experience with the Thesaurus, as well as the computational experience of others with dictionaries of various sorts, leads us to believe that we are finding ways to model and then refine our models of such resources so as to make them of far greater utility to knowledge-based computer systems.

This emphasis we are bespeaking on a 'differential diagnosis' as to alternative algorithms is in keeping with the generalization of the methodology so successfully employed in AI vision research by the late David Marr: clearly establishing the basic transfer function and then comparing algorithms for accomplishing it, before any programming is undertaken. That methodology also comports well with the widely employed approach utilized and advocated by Wayne Wymore (1977) for interdisciplinary efforts directed at solving large systems-analytic problems.

References

Brady, John. 1988. "ICSS (Interlingual Communication Support System) and a Wittgensteinian Language Game," Proceedings, European Studies Conference, University of Nebraska/Omaha, pp 20.27.

Brady, John. 1991. "Towards Automatic Categorization of Concordances Using Roget's International Thesaurus," Proceedings, Third Annual Midwest Artificial Intelligence and Cognitive Science Society Conference, ed. Gamble and Ball, Washington University, St. Louis, pp. 93-97.

Brugman, Claudia, and George Lakoff. 1988. "Cognitive Topology and Lexical Networks," in Small, Cottrell, and Tanenhous, eds., Lexical Ambiguity Resolution, Palo Alto: Morgan Kaufmann, pp. 477-508.

Bryan, Robert. 1973. "Abstract Thesauri and Graph Theory Applications in Thesaurus Research," in S. Sedelow, et al., Automated Language Analysis, 1972-1973, pp. 45-89. Lawrence: University of Kansas Departments of Computer Science and of Linguistics.

Bryan, Robert. 1974. "Modelling in Thesaurus Research," in S. Sedelow, et al., Automated Language Analysis, 1973-1974, pp. 44-59. Lawrence: University of Kansas Departments of Computer Science and of Linguistics.

Jacuzzi, Victor A. 1991. "Modeling Semantic Association Using the Hierarchical Structure of Roget's International Thesaurus," Oral Presentation, Dictionary Society of North America Biennial Meeting, University of Missouri, Columbia, Missouri, August, 1991.

Marr, David. 1982. Vision, W. H. Freeman.

Old, John. 1991a. "Analysis of Polysemy and Homography of the Word "lead" in Roget's International Thesaurus," Proceedings, Third Midwest Artificial Intelligence and Cognitive Science Society Conference, ed. Gamble and Ball, Washington University, pp. 98-102.

Old, John. 1991b. "Image Schemas and Lexicons: A Comparison between Two Lexical Networks," Oral Presentation, Dictionary Society of North America Biennial Meeting, University of Missouri, Columbia, Missouri, August, 1991.

Patrick, Archibald. 1985. An Exploration of an Abstract Thesaurus Instantiation. M.S. Thesis, Computer Science Department, University of Kansas/Lawrence.

Roget's International Thesauru, 3rd ed. 1962. New York: Thomas Y. Crowell.

Sedelow, Sally Yeates. 1985. "Computational Literary Thematic Analysis: The Possibility of a General Solution," Proceedings, 48th Annual Meeting of the American Society for Information Science, Vol. 22, pp. 359-362.

Sedelow, Sally Yeates and Walter A. Sedelow, Jr. 1986. "Thesaural Knowledge Representation," Proceedings, Advances in Lexicology, UW Centre for the New Oxford English Dictionary: Waterloo, Canada, pp. 31-43.

Sedelow, Sally Yeates and Walter A. Sedelow, Jr. (In Press) "A Topologic Model of the English Semantic Code and Its Role in Automatic Disambiguation for Discourse Analysis," in Hockey and Ide, eds., Proceedings, 10th International Conference on Computers and the Humanities, Oxford University Press, Oxford, England.

Sedelow, Walter A., Jr. and Sally Yeates Sedelow. 1987. "Semantic Space," Computers and Translation, 2, pp. 231-242.

Sedelow, Walter A., Jr. 1988. "Knowledge Retrieval from Domain-Transcendent Expert Systems: I. Some Concepts from Cognitive Robotics," Proceedings, 51st Annual Meeting of the American Society for Information Science, 1988, Vol. 25, pp. 205-208.

Talburt, John R. and Donna M. Mooney. 1989. "The Decomposition of Roget's International Thesaurus into Type-10 Semantically Strong Components," Proceedings, 1989 ACM South Regional Conference, Tulsa, Oklahoma, pp/. 78-83.

Warfel, Sam. 1972. "The Value of a Thesaurus for Prefix Identification," in S. Sedelow, et al., Automated Language Analysis, 1971-1972, pp. 31-49. University of Kansas Departments of Computer Science and of Linguistics.

Wymore, A. Wayne. 1977. Mathematical Theory of Systems Engineering: The Elements. Melbourne, Fla.: Krieger.