HOW TO SUMMARIZE THICK TEXT
(And Represent It Too)

Richard Alterman
Lawrence Bookman
Computer Science Department
Brandeis University

TEXT & SUMMARIZATION

1. Introduction

Consider the two pieces of text shown in figures 1 and 2. In one case we have a piece of text that is almost a caricature of a story, in the other case we have a piece of *thick text* taken from a book of folktales ("The clever peasant and the czars general:" Protter, 1961). The second story is thick because of the richness and complexity of the relationships amongst the events described; it is thick because it shows in detail the events of the story, rather than telling us their skeletal structure. One might argue, for example, that stories like the one described in figure 1 are unnatural because they are *pointless* (see Wilensky 1980,1982), or because their language is overly simplified. But, in contrast to "The Clever Peasant and the Czar's General," the reason for their unnaturalness becomes apparent: The first one is already a summary. In this paper we will describe a summarizer called SSs (**Summarization** Summarization Summarization...). that takes a thick piece of text, like the one shown in figure 2, and produces a skeletal piece of text, like the one shown in figure 1. SSs bases its summary on an event concept coherence analysis (ECC) (Alterman 1985,1988) of the text, as produced by a program called **NEXUS**.

Figure 3 depicts at the system level the relation between **NEXUS** and SSs. **NEXUS** produces an ECC representation by matching case encoded text against an associative network of event/state concepts. The ECC representation essentially sorts out and identifies the event concepts invoked by the text. SSs takes as input an ECC representation of a piece of text. It has basically two summarization strategies. The first technique works by first delineating boundaries of cohering event concepts and then extracting their core concept(s) - a shift in granularity The second technique works by identifying the major narrative thread of the story - a kind of outline. In this paper we present the first of these techniques. (See Alterman & Bookman, for details on the second of these techniques).

John was sailing his boat. Suddenly a gust of wind caught the sails. The boat capsized. John was very upset.

FIGURE 1. THE SAILING STORY

"Well, little farmer, what would you like?" the czar asked (1) when the peasant was brought (2) before him. "Nothing, father czar. I have come (3) only to bring (4) you a gift." And he opened (5) the lid of the metal chest. "And what would you like in return (6) for this gift of gold?" the czar inquired (7). "Father czar, just give (8) me a hundred lashes of the whip." "What?" the astonished czar exclaimed (9). "You ask (10) for a hundred lashes? You have brought (11) me gold and you want (12) a hundred lashes? No, no, no; you certainly deserve better (13)." "Please, father czar," the peasant insisted (14). "I don't want (15) any other reward. Just give (16) me a hundred lashes." So the czar reluctantly summoned (18) one of his

guards to fetch the whip. "Are you ready?" demanded (19) the puzzled ruler. "No, we must wait (20). I have (21) a partner who should share (22) the reward." A partner?" echoed (23) the bewildered czar. "Yes," the peasant answered (24). "When I came (25) to your door, the general would not let (26) me through until I vowed (27) that he would get (28) half my reward. So go ahead and start with him. Give (29) him the first fifty lashes."

FIGURE 2. A PORTION OF THE CLEVER PEASANT AND THE CZAR'S GENERAL

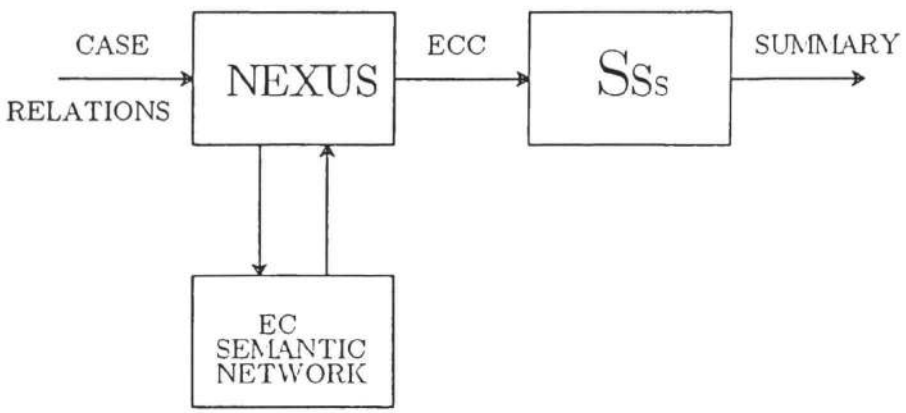

FIGURE 3. SYSTEM ARCHITECTURE

2. Event Concept Coherence

The ECC representation of the event descriptions of the text is a reflection of the relationships of the event/state concepts invoked by the text: the form of the knowledge and form of the text representation mirror one another. Its construction makes the following assumption: two event descriptions in a piece of text are *event concept coherent* if the positions of the concepts they invoke are proximal to one another in the underlying conceptual network. For NEXUS, knowledge about the relationships between event/state concepts are represented by an associative network, and an ECC interpretation captures the connectivity of a piece of text by essentially copying out and instantiating the relevant portion of the network. The representation was constructed so as to characterize the dependencies amongst the events in a causally neutral, but causally relevant form (Alterman 1988).

NEXUS uses intersection search as a basis for computing the representation (Alterman 1985; c.f. Quillian 1968; Charniak 1983; Norvig 1987). The input to NEXUS is the text in case notation form. NEXUS' program works in two stages. In stage one, it does a bi-directional breadth-first search to find a path between the event concepts previously introduced in the text (either explicitly or implicitly) and the new input event concept. In stage two, if a path is found, NEXUS propagates the case constraints along the path, simultaneously checking path consistency and performing some reference resolution. During the second stage default values are introduced to aid this process. The case constraints are also used by NEXUS to track the location of characters and to insure the correct sequencing of events in time.

Even though the representation of the text and the underlying conceptual network share the same structure, throughout this paper when we talk about the representation of the text, we will

emphasize the fact that the representation produced by NEXUS - because the representation it produces corresponds to the narrative structure of the text - is composed of multiple interconnected event concept trees (see Figure 4). A *concept tree* is recursively composed of all its subclass, part, and before/after descendants of a given event concept. Concept trees delineate the boundaries of individual event concepts and are temporally linked to one another forming what can be described as a *narrative stream*. The ECC representation characterizes the complex relationships amongst the events in a thick piece of text by grouping conceptually related events into tree-like structures for individual concepts, and temporal sequences of events for the narration as a whole.

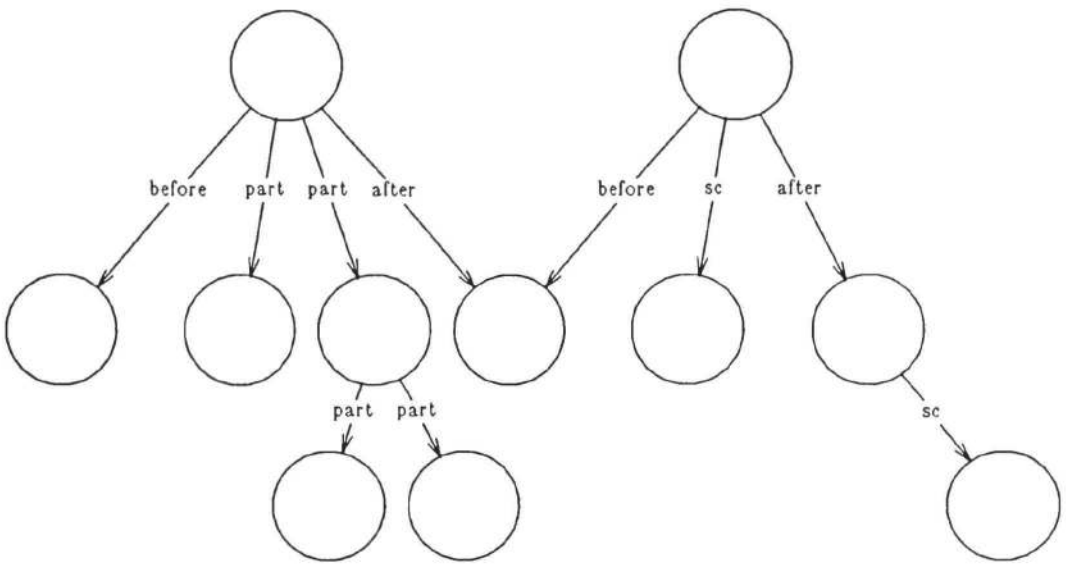

FIGURE 4. TWO INTERCONNECTED CONCEPT TREES.[1]

3. Thinning Techniques

Thick text is composed of a number of embellished text descriptions that are enmeshed together. A change in textual granularity from thick to thin text is achieved by systematically going through the text to identify and bundle together little pieces of text that conceptually form a larger event and then describing these concepts by their core concept. The shift in granularity is accomplished by a process of *delineation* and *extraction*. Delineation untangles the mesh of concepts by delineating the boundaries of the larger events and from these larger events the core concept is extracted.

Delineation occurs as a by-product of the ECC representation, as each concept tree in the ECC representation implicitly represents a delineated concept. Because of the hierarchical organization of the concept tree, extracting the central concept is relatively straightforward.

[1] There are actually seven coherence relations, since the *part*, *before*, and *after* relations used are further discriminated into (*subseq* & *coord*), (*ante* & *prec*), (*seq* & *consq*), respectively.

With one exception single concept trees can be summarized by extracting the top node in the concept tree. The basic idea of delineation and extraction is to consider each of the events in the representation produced by NEXUS and determine if it is the top of some concept tree. If it is a top, it is included in the summary, else it is discarded. Internally SSs represents each of the relations of the concept tree with a 3-tuple of the following form:

(Coherence_Relation Event/State1 Event/State2)

A *top* can be defined as follows:

For all events x,y and coherence relations c,
TOP(x) iff there exists no tuple such that (c,y,x) is true.

The exception occurs because NEXUS uses property inheritance to represent knowledge in the network. In a class/subclass relationship the subclass concept is preferred because it is more informative - the reader can predict all its relationships as well as the relationships it inherits from its ancestors.

4. An Example

Consider the portion of text from "The Clever Peasant and the Czar's General" shown in figure 2. When NEXUS and SSs were applied to this text using the delineation and extraction techniques it produced the following summary:

> The peasant had a audience with the czar. The peasant exchanged the chest of gold with the czar for a whipping. The peasant had a deal with a partner about sharing half the reward with him. The guards was to give the general a whipping. The czar requested his guards to fetch the whip.[2]

Figure 6 shows the analysis produced by NEXUS in which it identifies five event concept trees: *audience*, *exchange-gifts*, *request-of-underling*, *deal*, and *whipping*. Below is shown a description of some of the details of that analysis along with the correct sequencing of the concept trees.[3]

1. **The peasant has an audience with the czar**: beginning with event (2) the text introduces the fact that the peasant was brought before the czar. There is a conversation between the czar and the peasant (1,7,10,14,24).
2. **There is an exchange of gifts**: the peasant opens (5) a metal chest and the czar inquires what would he like in return (6). He tells the czar to give (8) him a hundred lashes.
3. **A request of underling**: the czar summoned (17) one of his guards to fetch (18) the whip.
4. **The peasant has made a deal with the general**: before he is to receive his reward he tells the general I have (21) a partner who should share (22) the reward.
5. **Administering the punishment**: the general was to be given (29) the first fifty lashes.

---

[2] Although we have some heuristics to handle the sequencing of events, these heuristics are not foolproof, as the event *request-of-underling* is not in the correct order

[3] Note the numbers in parenthesis indicate the events in the order in which they appear in the text in figure 2.

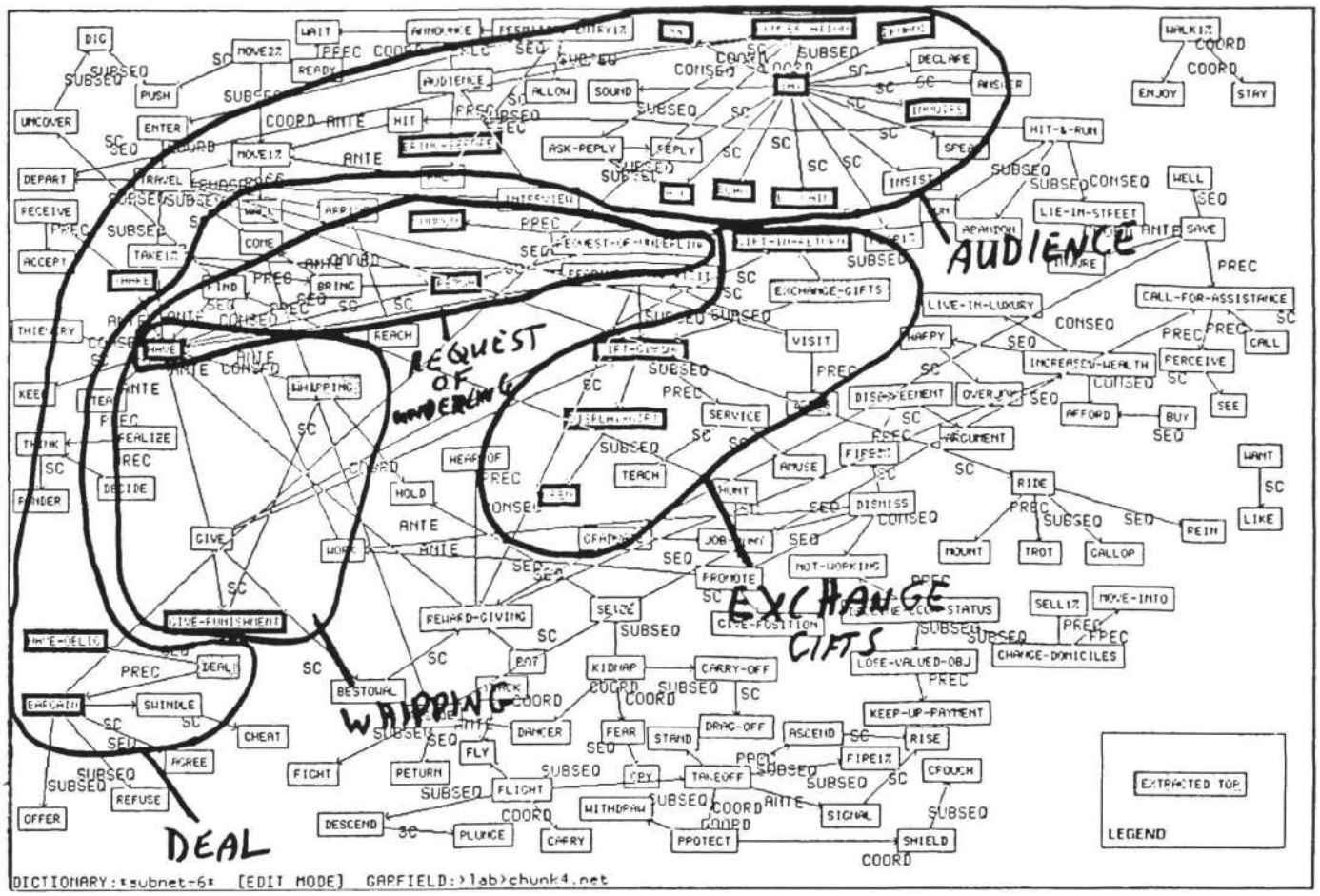

FIGURE 6. SCHEMATA OF "THE CLEVER PEASANT AND THE CZAR'S GENERAL"[4]

In terms of the narration of the text as a whole, the text is composed of three *chunks* of concept trees. (By chunks we mean interconnected concept trees.) In the first chunk, there is a connecting path between the concept trees *audience* and *deal*: Because the general and the peasant made a deal, the peasant has an obligation to the general, and this obligation is reported to the czar in the course of their conversation during the peasant's audience. In the second chunk there is a connecting path between the concept trees *request-of-underling* and *whipping*: The czar requests his guards to fetch the whip. As a result they have the whip in which to punish the peasant's partner, the general. The third chunk is the *exchange-gifts* concept tree which is not connected to either of the other chunks. SSs takes this underlying event structure and thins the text by extracting the tops of each of the concept tree within each of the interconnected chunks.

5. The Computational Literature on Summarization

Lehnert & Loiselle (1988) developed a scheme for summarizing text based on plot units (Lehnert, 1981). Plot units represent affect-state patterns. Lehnert identifies a number of primitive plot units (e.g. motivation, success, perseverance) which can be combined into more complex

[4] Note the shaded rectangles represent the ECC interpretation produced by NEXUS For further details of this analysis, see Alterman & Bookman (forthcoming) The graphics for this picture was created using AINET-2 (Chun 1986)

plot units (e.g. fortuitous problem resolution, fleeting success, giving up). Narrative text is represented by interconnected plot units and summaries are based on the identification of pivotal plot units, i.e., the plot units which are maximally connected.

Wilensky introduced a theory of summarization based on the identification of *story points* (Wilensky, 1982, 1980). Story points roughly correspond to the essential tension points of a story, i.e., what the story is about. The idea is that points represent what is interesting in a story and therefore likely to be included in a summary. Wilensky suggests some rules for recognizing points. The rules are based on his theory of goal interaction (Wilensky, 1983). If a character plans to go outside to get the newspaper and discovers it is raining outdoors, a *goal conflict* occurs between the goal to get a newspaper and the goal to stay dry. Wilensky argues that situations where goal interactions occur are potentially dramatic and consequently likely candidates as story points.

Both Wilensky and Lehnert are describing summarization techniques that attempt to identify critical points of interest. Both their techniques are based on analyses that impose high-level concepts in a top-down manner on the text. The difficulty is that there is a gap between the initial representation of the text and the actual representation used for computing both the story points and plot units respectively. An ECC representation for thick text, however, acts as a way station that sorts out some of the underlying event structure. It provides a schemata from which, perhaps these other summarizers can further reduce the volume of text by deciding what is of interest.

Research on story trees characterize the text by a set of meta-descriptions, e.g., episode or setting. Rumelhart (1975; Simmons & Correira, 1980; and Correira, 1980) summarized text based on a hierarchical organization of the text. Again we have a gap between a high-level structure and the events depicted in the text, which can be reduced by the analysis we are proposing. The work of Van Dijk (Macro-structures, 1976) suggests techniques for thinning out text, but organizes it in terms of a single hierarchical structure. The work described in this paper differs from the work on macro-structures with regards to the problems of thick text in two ways. First it emphasizes the role of events in the thinning process. Second the representation it produces does not take the form of a single hierarchical structure, but rather it is structured by an ECC analysis into a narrative stream.

FRUMP (DeJong 1979) produces representations of text by applying in a top-down fashion sketchy scripts (e.g. accidents and terrorist acts). It extracts from wire-service newspaper stories just enough facts to fill in the arguments of a sketchy script. Because the stories that FRUMP works with are so stereotyped, it could summarize text by using a set of fill-in-the-blank type summarization statements attached to each sketchy script. These techniques are by definition limited to stereotypical situations. Thick text includes many relationships between events which cannot be accounted for by a sketchy script, hence a sketchy script analysis can only account for some of the event relationships in a thick piece of text.

6. Summary and Conclusions

Thick text is composed of a number of embellished text descriptions that are enmeshed together. A piece of text is thick when it shows the reader, in detail, the events of the story, rather than telling the reader only of its skeletal structure. This paper has described a

summarizer called SSs that takes a thick piece of text and produces a skeletal structure for the events described in that text. SSs bases its summary of the story on an ECC analysis of the text. An ECC analysis of the text sorts through the events of the story, grouping together events based on their relative positions in an underlying conceptual network of events.

NEXUS and SSs have been applied to several examples of text, including a page and half folktale taken from a book of folktales ("The Clever Peasant and the Czar's General," Protter 1961), "The Xenon Story" (Wilensky 1980), and "The Czar's Three Daughters" (Lehnert & Loiselle, 1988). (See Alterman & Bookman for details). Experiments show NEXUS and SSs reducing the volume of text between 60% and 80%. The importance of this work, with regards to the problem of thick text, is that it reduces the gap between top-down theories of text concerned with issues of interest and salience (e.g. plot units or prototype points) and the complexity of the event relationships as they are depicted in the text.

References

Alterman, R. (1985). A Dictionary Based on Concept Coherence. *Artificial Intelligence*, Vol. 25, pp 153-186, North Holland.

Alterman, R. (1988). Event Concept Coherence. In D. Waltz (Ed.), Advances in Natural Language Understanding. Hillsdale, NJ: Lawrence Earlbaum (forthcoming).

Alterman, R. & Bookman, L. Some Computational Experiments in Summarization (in preparation).

Charniak, E. (1983). Passing Markers: A Theory of Contextual Influences in Language Comprehension. *Cognitive Science*, Vol. 7, pp 171-190.

Chun, H.W. AINET-2 User's Manual. Computer Science Department, Brandeis University, CS-86-126, Waltham MA, 1986.

Correira, A. (1980). Computing story trees. *American Journal of Computational Linguistics*, 6, pp 135-149.

DeJong, G. (1979). Prediction and Substantiation: A New Approach to Natural Language Processing. *Cognitive Science*, Vol 3, pp 251-273.

Lehnert, W. (1981). Plot Units and Narrative Summarization. *Cognitive Science*, Vol 5, no. 4, pp 293-331.

Lehnert, W. & Loiselle, C. (1988). An Introduction to Plot Units. In D. Waltz (Ed.), Advances in Natural Language Understanding. Hillsdale, NJ: Lawrence Earlbaum (forthcoming).

Norvig, P. (1987). Inference Processes and Knowledge Representation for Text Understanding. Phd Thesis, University of California at Berkeley, UCB/CSD 87/339.

Protter, E. (1961). A Children's Treasury of Folk and Fairy Tales. Channel Press.

Quillian, R. (1968). Semantic Memory. In M. Minsky (Ed.), Semantic Information Processing, MIT Press.

Rumelhart, D. E. (1975). Notes on A Schema for Stories. In D. G. Bobrow and A. Collins, (Eds.), Representation and Understanding, Academic Press.

Simmons, R. F. & Correira, A. (1980). Rule Forms for Verse, Sentences, and Story Trees. In N. Findler (Ed.), Associative networks: The Representation and Use of Knowledge in Computers. NY: Academic Press.

Van Dijk, T. A. (1976). Macro-structures and Cognition. In *Twelfth Annual Carnegie Symposium on Cognition*. Carnegie Foundation.

Wilensky, R. (1980). What's the point? In *Proceedings of the Third National Conference of the Canadian Society for the Computational Studies of Intelligence*.

Wilensky, R. (1982). Points: A theory of the structure of stories in memory. In W Lehnert & M. Ringle (Eds.), Strategies for natural language processing. Hillsdale, NJ: Lawrence Earlbaum.

Wilensky, R. (1983). Planning and understanding. Reading, MA: Addison-Wesley.