Diagrams of Social Cohesion

Warren SACK
MIT Media Laboratory
20 Ames Street, E15-020b
Cambridge MA 02139
wsack@media.mit.edu

Abstract

A procedure for computing and then combining social networks and lexical cohesion chains between newsgroup messages is described. This combination is called *social cohesion*. These relations of social cohesion are diagrammed in a graphical interface. An example of the diagrams can be found here: www.media.mit.edu/~wsack/acl99-demo.html

Introduction

Recent work in cohesion analysis (Halliday and Hasan, 1976) has produced new computational methods for summarizing and cross-indexing documents. New work in cohesion analysis includes, for example, lexical cohesion analysis of single documents (e.g., Hirst and St-Onge, 1998; Stairmand, 1997); discourse segmentation of single documents (e.g., Marcu, 1998); and, coreference analysis of multiple documents (e.g., Bagga and Baldwin, 1998). Interesting new work shows how cohesion structure can be used in the inference of discourse coherence relations (e.g., Harabagiu, 1998). However, the goal of work in computational cohesion analysis has largely been to automatically find intra- and inter-textual (Halliday, 1994) relations between documents. The work described in this paper is an attempt to analyze inter-textual relations between documents in conjunction with inter-personal (Halliday, 1994) relations. In other words, the goal is to analyze both the textual and the social structure of a discussion involving a large number of people.

This analysis is carried out by overlaying a lexical cohesion analysis of multiple documents on top of a computed social network. The concept of "social network" is imported from sociology (e.g., Wasserman and Galaskiewicz, 1994).

The domain of analysis is archives of Usenet newsgroup postings. The implemented system produces two outputs and overlays them one on top of the other: (1) a diagram of the social network of newsgroup participants detailing who is responding to and/or citing (by quotation) whose messages in the group discussions; and, (2) a list of conversational "themes" (more precisely, a list of lexical chains) extracted from the archives by examining which words and synonyms of words are picked up in the discussion threads by posters in their quotations from and replies to previous messages. The combination of the two is here referred to as an analysis of *social cohesion*.

Examination of a diagrammatic display of the social cohesion analysis gives a rough approximation of which participants are exchanging messages about specific themes. Hearst (Hearst, 1997) and others have convincingly argued for the use of graphical interfaces to display the results of complex information retrieval and computational linguistics algorithms. By drawing the results of the social cohesion analysis as a diagram and so, thereby, overlaying the output of the lexical cohesion analysis on top of the derived social network, a workable overview of one thousand or more messages can be plotted on a single page or screen.

The remainder of this paper will sketch out the social cohesion analysis procedure and briefly describe how the system graphically diagrams the results of an analysis of an archive of Usenet newsgroup messages. More detailed explanations of the analysis and graphic display routines will be the subjects of forthcoming papers.

1 Analysis Procedure

The analysis procedure performs the following steps on an archive of Usenet newsgroup messages in order to compute a social network, an analysis of inter-message, lexical cohesion, and the interrelationships between the two:

(a) Messages are "threaded."

(b) Quotations in the messages are identified and their sources (in other messages) are found.
(c) An index of posters (i.e., newsgroup participants) to messages is built.
(d) For every poster, the set of all other posters who replied to the poster -- or quoted from messages authored by the poster -- is recorded. Posters who reply to and/or quote from one another are linked together in the social network. Reciprocity is therefore highlighted in the computed social network.
(e) The words in the messages are divided into sentences, tagged with part-of-speech information, and their roots are identified.
(f) An analysis of lexical cohesion is performed on every pair of messages where a pair consists of one message of a "thread" and another message that either immediately follows the first message in the thread (i.e., is a reply to the first message) and/or follows the first message in the thread and contains a quotation from the first message. This analysis produces a series of lexical chains between messages that can be understood as a crude approximation to the theme of the conversation in a sequence of messages. The lexical database WordNet is used in the lexical cohesion procedure.
(g) By using the index created in step (c) with the results of step (f) a set of lexical chains are computed for every pair of posters who have replied to and/or quoted from one another over the course of time represented by the Usenet newsgroup archive under analysis. These aggregated lexical chains are layered on top of the social network computed in step (d).

2 Graphical Interface

The graphical display of the system is one screen divided into four parts.
(1) The upper left-hand side of the screen displays the social network. The nodes of the social network are labeled with the names of reciprocating posters. The edges between the posters are labeled with lexical chains.
(2) These chains (or, roughly "themes") are represented as synonym sets (i.e., synsets from WordNet) and listed in a menu displayed to the right of the social network. If a synonym set is selected from this menu, one or more links in the social network are highlighted allowing one to see which posters have exchanged messages concerning the "theme" selected.
(3) The "semantic networks" panel to the right of the themes menu displays output for a part of the system not discussed in the present paper.
(4) At the bottom of the screen is a graphical display of the threads of the discussion.

Conclusion

The simplest description of the system described is that it is a Usenet newsgroup browser that allows one to very quickly see who has been "talking" to whom about what in a newsgroup and then "zoom into" the archive to read the messages authored by particular people or *apropos* of particular conversational themes.

References

Amit Bagga and Breck Baldwin "Entity-Based Cross-Document Coreferencing Using the Vector Space Model" In *Proceedings of ACL-COLING'98*, June 1998, Montreal, Canada.

Michael A. K. Halliday. *An Introduction to Functional Grammar, Second Edition* (London: Edward Arnold, 1994).

Michael A. K. Halliday and Ruqaiya Hasan. *Cohesion in English* (New York: Longman, 1976).

Sanda Harabagiu. "WordNet-Based Inference of Textual Cohesion and Coherence" in the *Proceedings of FLAIRS-9*8, May 1998, Sanibel Island, FL.

Marti Hearst. "Interfaces for Searching the Web" *Scientific American*, March 1997..

Graeme Hirst and David St-Onge. "Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms" in Christiane Fellbaum (editor) *WordNet: An Electronic Lexical Database* (Cambridge, MA: MIT Press, 1998).

Daniel Marcu. "Improving summarization through rhetorical parsing tuning" in the *Sixth Workshop on Very Large Corpora*, Montreal, Canada, August 1998.

Mark A. Stairmand. "Textual context analysis for information retrieval" in the *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, August 1997.

Stanley Wasserman and Joseph Galaskiewicz (editors) *Advances in Social Network Analysis: Research in the Social and Behavioral Sciences* (Thousand Oaks, CA: Sage Publications, 1994).