Using lexical chains to build hypertext links in newspaper articles

Stephen Green
Department of Computer Science, University of Toronto
Toronto, Canada, M5S 1A4
sjgreen@cs.toronto.edu

My research interests are in hypertext and browsing systems for information retrieval from newspapers. Marchionini [1989] has argued that browsing places fewer demands on the novice user, and Marchionini *et al.* [1993] report tests performed with search experts and domain experts that showed that browsing was an important component of the search technique of users unfamiliar with an information retrieval system.

The popularity of graphical interfaces to the World Wide Web (WWW) has shown that a hypertext interface can make what was once a daunting task, accessing information across the Internet, considerably easier for the novice user.

Along with — or perhaps because of — the growth of the WWW, many newspapers are beginning to take their first steps into the online world. The problem is that these papers are not making full use of the hypertext capabilities of the WWW. The user might find a particular article from a particular issue using hypertext links, but they must then read the entire article to find the information that interests them. It would be more useful if the full articles were indexed as a hypertext.

Unfortunately, the task of building hypertexts is a difficult and time consuming one. I am currently working on a method using *lexical chains* [Morris and Hirst, 1991] for automatically building hypertexts from newspaper articles .

Lexical chains capture the semantic relations between words that occur throughout a text. Each chain is a set of related words that captures a portion of the cohesive structure of a text. A particular chain will tend to delineate the parts of a text that are about the same thing, by indicating the cohesive ties that exist in the text. These chains can be built using a lexical resource such as WordNet that relates words on the basis of their semantic closeness.

Once the chains have been built, we can use them to decide where to insert hypertext links between the paragraphs of an article.

For each paragraph, we can consider how many words take part in lexical chains and how many different lexical chains appear. Using this information, we can determine which paragraphs have a similar distribution of lexical chains, that is, which paragraphs have a similar set of chains with a significant number of words in these chains. Paragraphs that show a high similarity are then linked together in the hypertext.

We are currently trying to determine exactly what criteria should be used to determine when two paragraphs show a "high similarity". These criteria will be based on example articles that have been linked by human judges.

The next step will be to determine how articles can be related to one another through their lexical chains. This stage will most likely consist of a corpus study, considering related articles and their associated lexical chains. Articles that show a similar distribution of chains and chain words would be linked. It would seem that rules for linking articles in the same newspaper should be easily generalizable to linking articles from different newspapers.

The process is still in its initial stages, but the results so far have been very promising.

References

[Marchionini *et al.*, 1993] Gary Marchionini, Sandra Dwiggins, Andrew Katz, and Xia Lin. Information seeking in full-text end-user-oriented search systems: The roles of domain and search expertise. *Library and Information Science Research*, **15**(1):35–69, 1993.

[Marchionini, 1989] Gary Marchionini. Making the transition from print to electronic encyclopedia: Adaptation of mental models. *International journal of man-machine studies*, **30**(6):591–618, 1989.

[Morris and Hirst, 1991] Jane Morris and Graeme Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, **17**(1):21–48, 1991.