A Common Theory of Information Fusion from Multiple Text Sources Step One: Cross-Document Structure

Dragomir R. Radev
550 E. University St.
University of Michigan
Ann Arbor, MI 48109
radev@umich.edu

Abstract

We introduce CST (cross-document structure theory), a paradigm for multi-document analysis. CST takes into account the rhetorical structure of clusters of related textual documents. We present a taxonomy of cross-document relationships. We argue that CST can be the basis for multi-document summarization guided by user preferences for summary length, information provenance, cross-source agreement, and chronological ordering of facts.

1 Introduction

The Topic Detection and Tracking model (TDT) [Allan et al. 98] describes news events as they are reflected in news sources. First, many sources write on the same event and, second, the same source typically produces a number of accounts of the event over a period of time. Sixteen news stories related to the same event from six news sources over a two-hour time period are represented in Figure 1.

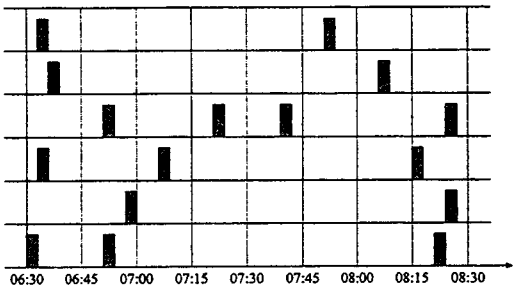

Figure 1 : Time distribution of related documents from multiple sources

A careful analysis of related news articles shows that they exhibit some interesting properties [Radev & McKeown 98]. In some cases, different sources agree with each other, at other times, the information presented in them is contradictory. The same source sometimes adds new information as it becomes available and puts it in the context of what has already been discussed earlier. In other cases, to get a full picture of an event, one has to read stories from multiple sources as neither of them presents all relevant information. All these examples point to the existence of cross-document structure that is waiting to be exploited.

Figure 2 illustrates how the same story can be told in several different ways. The six extracts are from news stories about the same event: the declaration by Bill Clinton at a press conference that millions of dollars will be handed out to low income people affected by recent surges in oil prices.

In this paper we introduce CST, a theory of cross-document structure. CST assigns labels

such as subsumption, update, and elaboration to cross-document conceptual links. We argue that CST is essential for the analysis of contradiction, redundancy, and complementarity in related documents and for multi-document summarization (MDS).

FoxNews: President Clinton announced Wednesday he will release funds to help New Englanders through the heating oil "crisis" and spoke on topics ranging from gun violence to his personal life and foreign policy.
The president said he ordered the release of $125 million from the Low Income Home Energy Assistance Program to help families who must rely on oil to heat their homes. He said that amount was all that is left in the fund this year.

ABCNews: President Clinton today ordered the release of millions of dollars in assistance for Northeastern families struggling with soaring fuel costs, saying Americans have together to help their fellow citizens in times of need.
The release of $120 million from the Low Income Home Energy Assistance Program is to help families who must rely on oil to heat their homes, he said ...

CNN: Citing rising energy costs, President Clinton said Wednesday he is releasing $120 million in funds to help families buy home heating oil in the mid-Atlantic and northeastern states.
The president said he decided the help was needed because steadily rising oil prices have slammed consumers dependent on home heating oil and have boosted the cost of gasoline.

MSNBC: AT A WASHINGTON, D.C. news conference, Clinton announced that he had ordered the release of the remainder of federal heating assistance money to help families in the Northeast hit hard by this year's cold weather and the soaring cost of home heating oil.
Clinton said he ordered the release of $125 million from the Low Income Home Energy Assistance Program to help families who must rely on oil to heat their homes. He said that amount was all that is left in the fund this year.

USA Today: President Clinton, saying too many families are being hurt by the soaring cost of heating their homes, announced Wednesday he will ask Congress for $600 million in emergency assistance to help people meet heating oil costs. He also released an additional $125 million for more immediate help to low-income families.

NY Times: WASHINGTON, Feb. 16 -- President Clinton ordered the release today of another $125 million in government aid to help needy families pay the soaring costs of heating their homes this winter.
The release, announced by Mr. Clinton at the start of a White House news conference this afternoon, comes only six days after the government made $130 million in home-heating aid available.

Figure 2: Six different accounts of the same event

2 Related Work

2.1 Document structure

Rhetorical Structure Theory (RST) [Mann & Thompson 88, Mann 00] is a comprehensive theory of text organization. It is based on "text coherence", or the presence in "carefully written text" of unity that would not appear in random sequences of sentences. RST posits the existence of relations among sentences. Most relations consist of one or more *nuclei* (the central components of a rhetorical relation) and zero or more *satellites* (the supporting components of the relation). An example of an RST relation is *evidence* which is decomposed into a nucleus (a claim) and a satellite (text that supports the claim). RST is intentionally limited to single documents. With CST, we attempt to describe the rhetorical structure of sets of related documents. Unlike RST, CST cannot rely on the deliberateness of writing style. We can however make use of some observations of structure across documents which, while clearly not deliberate in the RST sense, can be quite predictable and useful. In a sense, CST associates a certain behavior to a "collective document author" (that is, the collectivity of all authors of the related documents).

A pioneering study in the typology of links among documents is described in [Trigg 83, Trigg & Weiser 87]. Trigg introduces a taxonomy of link types across scientific papers. The 80 suggested link types such as *citation, refutation, revision, equivalence, and comparison* are grouped in two categories: *Normal* (inter-document links) and *Commentary* (deliberate cross-document links). While the taxonomy is quite exhaustive, it is by no means appropriate or intended for general domain texts (that is, other than scientific articles).

A large deal of research in the automatic induction of document and hyperdocument structure is due to Salton's group at Cornell [Salton et al. 91]. [Allan 96] presents a graph simplification technique for "hyperlink typing", that is, assigning link types from Trigg's list to links between sentences or paragraphs of a pair of documents. Allan tested his techniques on sets of very distinct articles (e.g. "John F. Kennedy" and "United States of America" from the Funk and Wagnalls encyclopedia). As the author himself admits, the evaluation in [Allan 96] is very weak and doesn't indicate to any extent whether the techniques actually achieve anything useful.

More recently, [Salton et al. 97] introduced a technique for document structuring based on *semantic hyperlinks* (among pairs of paragraphs which are related by a lexical similarity significantly higher than random). The authors represent single documents from the Funk and Wagnalls encyclopedia on topics such as Abortion or Nuclear Weapons in the form of *text relationship maps*. These maps exploit the *bushiness* (or number of connecting edges) of a paragraph to decide whether to include it in a summary of the entire article. The assumption underlying their technique is that *bushy paths* (or paths connecting highly connected paragraphs) are more likely to contain information central to the topic of the article. The summarization techniques described in Salton et al.'s research are limited to single documents.

One of the goals of CST is to extend the techniques set forth in Trigg, Salton, and Allan's work to cover sets of related documents in arbitrary domains.

2.2 Multi-document summarization

SUMMONS [Radev & McKeown 98] is a knowledge-based multi-document summarization system, which produces summaries of a small number of news articles within the domain of terrorism. SUMMONS uses as input a set of semantic templates extracted by a message understanding system [Fisher et al. 96] and identifies some patterns in them such as change of perspective, contradiction, refinement, agreement, and elaboration. The techniques used in SUMMONS involved a large amount of knowledge engineering even for a relatively small domain of text (such as accounts of terrorist events) and is not directly suitable for domain-independent text analysis. The planning operators used in it present, however, the ideal first step towards CST.

[Mani & Bloedorn 99] use similarities and differences among related news articles for MDS. They measure the effectiveness of their method in two scenarios: paragraph alignment across two articles and query-based information retrieval. None of these scenarios evaluates the generation of query-independent summaries of multiple articles in open domains.

The Stimulate projects at Columbia University [Barzilay & al. 99], [McKeown & al. 99] have been using natural language generation to produce multi-document summaries. Their technique is called *theme intersection*: paragraph alignment across news stories with the help of a semantic network to identify phrases which convey the same meaning and then generate new sentences from each theme and order them chronologically to produce a summary.

We should note here that RST has been used to produce single-document summaries [Marcu 97]. For multi-document summaries, CST can present a reasonable equivalent to RST.

2.3 Time-dependent documents

Time-dependent documents are related to the observation that perception of an event changes over time and include (a) *evolving summaries* (summaries of new documents related to an ongoing event that are presented to the user assuming that he or she has read earlier summaries of related documents) [Radev 99] and (b) *chronological briefings* [Radev & McKeown 98]. [Carbonell et al. 98] discuss the motivation behind the use of time-dependent documents and [Berger & Miller 98] describe a language model for time-dependent corpora.

3 Representing cross-document structure

We will introduce two complementary data structures to represent multi-document clusters: the *multi-document cube* (Section 0) and the *multi-document graph* (Section 0).

3.1 Multi-document cubes

Definition A *multi-document cube C* (see Figure 3 (a)) is a three dimensional structure that represents related documents. The three dimensions are t (time), s (source) and p (position within the document).

Definition A *document unit U* is a tuple (t,s,p) – see Figure 3 (b). Document units can be defined at different levels of granularity, e.g., paragraphs, sentences, or words.

Definition A *document D* is a sequence of document units $U_1U_2...U_n$ which corresponds to a one-dimensional projection of a multi-document cube along the source and time dimensions.

Some additional concepts can be defined based on the above definitions.

Definition A *snapshot* is a slice of the multi-document cube over a period of time Δt – see Figure 3 (c).

Definition An *evolving document* is a slice of the multi-document cube in which the source is fixed and time and position may vary.

Definition An *extractive summary S* of a cube *C* is a set of document units, $S \subset C$, see Figure 3 (d).

Definition A *summarization operator* transforms a cube *C* into a summary *S*.

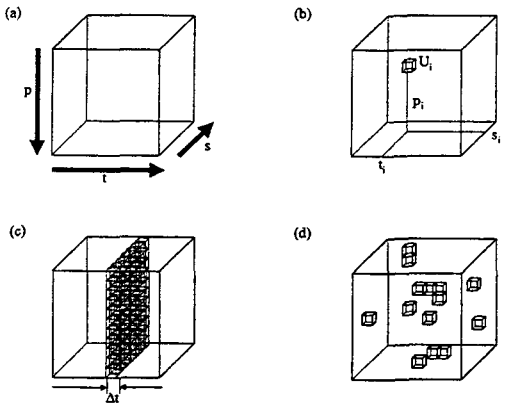

Figure 3: (a) A multi-document cube, (b) A document unit, (c) A cube slice, (d) An extracted summary

3.2 Multi-document graphs

While multi-document cubes are a useful abstraction, they cannot easily represent text simultaneously at different levels of granularity (words, phrases, sentences, paragraphs, and documents). The second formalism that we introduce is the *multi-document graph*. Each graph consists of smaller subgraphs for each individual document (Figure 4). We use two types of links. The first type represents inheritance relationships among elements within a single document. These links are drawn using thicker lines. The second type represents semantic relationships among textual units. The example illustrates sample links among documents, phrases, sentences, and phrases.

4 A taxonomy of cross-document relationships

Figure 5 presents a proposed taxonomy of cross-document relationships. The Level column indicates whether the relation applies to words (W), phrases (P), sentences or paragraphs (S), or entire documents (D). The examples are from our MDS corpus (built from TDT and Web-based sources).

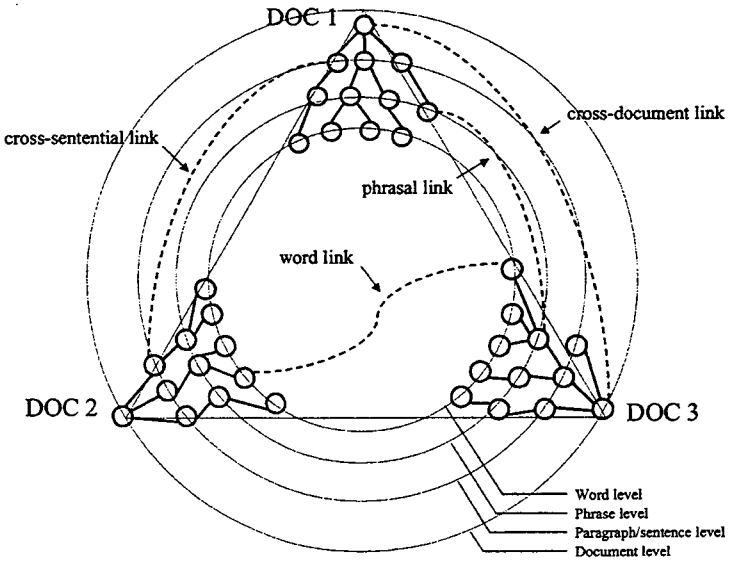

Figure 4: Sample multi-document graph

#	Relationship type	Level	Description
1	Identity	Any	The same text appears in more than one location
2	Equivalence (paraphrasing)	S, D	Two text spans have the same information content
3	Translation	P, S	Same information content in different languages
4	Subsumption	S, D	One sentence contains more information than another
5	Contradiction	S, D	Conflicting information
6	Historical background	S	Information that puts current information in context
7	Cross-reference	P	The same entity is mentioned
8	Citation	S, D	One sentence cites another document
9	Modality	S	Qualified version of a sentence
10	Attribution	S	One sentence repeats the information of another while adding an attribution
11	Summary	S, D	Similar to Summary in RST: one textual unit summarizes another
12	Follow-up	S	Additional information which reflects facts that have happened since the last account

13	Elaboration	S	Additional information that wasn't included in the last account
14	Indirect speech	S	Shift from direct to indirect speech or vice-versa
15	Refinement	S	Additional information that is more specific that the one previously included
16	Agreement	S	One source expresses agreement with another
17	Judgment	S	A qualified account of a fact
18	Fulfilment	S	A prediction turned true
19	Description	S	Insertion of a description
20	Reader profile	S	Style and background-specific change
21	Contrast	S	Contrasting two accounts or facts
22	Parallel	S	Comparing two accounts of facts
23	Generalization	S	Generalization
24	Change of perspective	S, D	The same source presents a fact in a different light

Figure 5: Sample types of edges (relationships between textual spans)

One example of a cross-document relationship is the *cross-sentence informational subsumption* (CSIS, or subsumption), which reflects that certain sentences repeat some of the information present in other sentences and may, under certain circumstances, be omitted during summarization. In the following example, sentence (2) subsumes (1) because the crucial information in (1) is also included in (2) which presents additional content: "the court", "last August", and "sentenced him to life".

(1) John Doe was found guilty of the murder.
(2) The court found John Doe guilty of the murder of Jane Doe last August and sentenced him to life.

Paraphrase

(3) Ford's program will be launched in the United States in April and globally within 12 months.
(4) Ford plans to introduce the program first for its employees in the United States, then expand it for workers abroad.

Modality

New stories are often written in a way that makes misattributions of information difficult, e.g., by referring to a person arrested at a crime scene as an "alleged" or "suspected" perpetrator.

(5) Adams **reportedly** called for an emergency meeting with Trimble to try to salvage the assembly.
(6) Sinn Fein leader Gerry Adams appealed for an urgent meeting with Trimble.

(7) The GIA **is the most hardline** of the Islamic militant groups which have fought the Algerian authorities since 1992.
(8) The GIA **is seen as most hardline** of the Islamic militant groups which have fought the Algerian government during the past seven years.

Attribution

(9) In the strongest sign yet that Russia's era of space glory is coming to an end, **space officials announced** today that cosmonauts will leave the Mir space station in August and it will remain unmanned.
(10) The crew aboard the Mir space station will leave in August, and the craft will orbit the Earth unmanned until early next year.

Indirect Speech

(11) An anonymous caller told the Interfax news agency that the Moscow explosion and a Saturday night bomb blast in southern Russia were **in response to Russia's military campaign against Islamic rebels in the southern territory of Dagestan.**

(12) An anonymous caller to Interfax said the blast and a car-bomb earlier this week at a military apartment building in Dagestan were **"our response to the bombing of villages in Chechnya and Dagestan."**

Followup

(13) Denmark's largest industrial unions have rejected a wage proposal, setting the stage for a nationwide general strike, officials announced **Friday**.

(14) A national strike entered its second week **Monday**, paralyzing Denmark's main airport and leaving most gasoline stations out of fuel and groceries short of frozen and canned foods.

Judgment

(15) **Hardline militants** of Algeria's Armed Islamic Group (GIA) **threatened** Sunday **to create a "bloodbath"** in Belgium if the authorities there do not release several of its leaders jailed last month.

(16) The GIA **is demanding** that Belgium release several of its leaders jailed in Belgium last month.

Fulfillment

(17) WASHINGTON, **May 31** The Federal Bureau of Investigation **plans to put** suspected terrorist Osama bin Laden, sought in connection with the bombings of the US embassy bombings in Africa, on its "Ten Most Wanted" list, CNN reported Saturday.

(18) WASHINGTON, **June 7** The Federal Bureau of Investigation **added** Saudi fugitive Osama Bin Laden, sought for his part in the 1998 bombings of US embassies in Africa, to its "Ten Most Wanted List" Monday.

Elaboration

(19) Fugitive Saudi national bin Laden is believed to be the mastermind behind last year's bloody attacks against US embassies in Kenya and Tanzania.

(20) Bin Laden, **41**, is believed to be the mastermind behind last year's bloody attacks against US embassies in Kenya and Tanzania.

Update

(21) The confirmed death toll has already reached **49**, while over **50** people are still unaccounted for, many presumed dead and buried in the ruins.

(22) The confirmed death toll has already reached **60**, and another **40** people are still unaccounted for, most presumed dead and buried in the ruins.

Definition

(23) Yeltsin said the security forces must unite to fight terrorists, adding that he had appointed **Interior Minister Vladimir Rushailo** to head a special tea m coordinating anti-terrorist activities.

(24) Yeltsin said the security forces must unite to fight terrorists, adding that he had named **Rushailo** to head a special team coordinating anti-terrorist activities.

Contrast

(25) Agriculture Minister Loyola de Palacio estimated the loss at dlrs 10 million.

(26) Agriculture Minister Loyola de Palacio has estimated losses from ruined produce at 1.5 billion pesetas (dlrs 10 million), although farmers groups earlier claimed total damages of nearly eight times that amount.

Historical background

(27) Elian's mother and 10 others died when their boat sank as they tried to reach the United States from Cuba.

5 Using CST for information fusion

In this section we describe how CST can be used to generate personalized multi-document summaries from clusters of related articles in four steps: clustering, document structure analysis, link analysis, and personalized graph-based summarization (Figure 6).

The first stage, clustering, can be either query-independent (e.g., based on pure document

similarity [Allan et al. 98]) or based on a user query (in which case clusters will be the sets of documents returned by a search engine).
The second stage, document analysis, includes the generation of document trees representing the sentential and phrasal structure of the document [Hearst 94, Kan et al. 98].

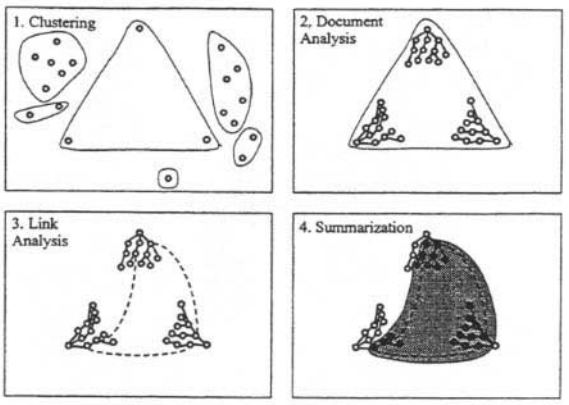

Figure 6: Processing stages

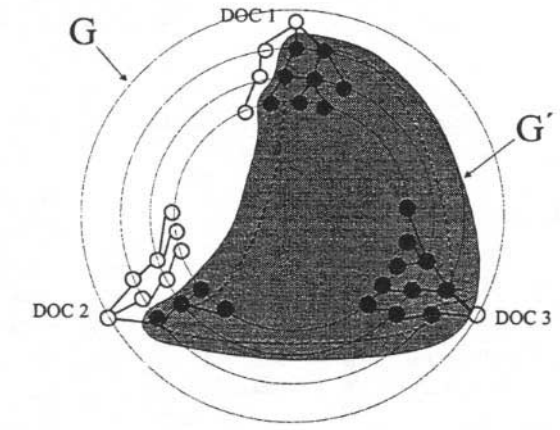

Figure 7: Summarization using graph cover operators

The third stage is the automatic creation and typing of links among textual spans across documents. Four techniques for identifying related textual units across documents can be used: *lexical distance*, *lexical chains*, *information extraction*, and *linguistic template matching*. Lexical distance (see e.g., [Allan 96]) uses cosine similarity across pairs of sentences. Lexical chains [Barzilay & Elhadad 97] are more robust than lexical matching as they take into account linguistic phenomena such as synonymy and hypernymy. The third technique, information extraction [Radev & McKeown 98] identifies salient semantic roles in text (e.g., the place, perpetrator, and effect of a terrorist event) and converts them to semantic templates. Two textual units are considered related whenever their semantic templates are related. Finally, a technique that will be used to identify some relationships such as citation, contradiction, and attribution is *template matching* which takes into account transformational grammar (e.g., relative clause insertion). For link *type analysis*, machine learning using lexical metrics and cue words is most appropriate (see [Kupiec et al. 95], [Cohen & Singer 96]).

The final step is summary extraction, based on the user-specified constraints on the summarizer. A graph-based operator defines a transformation on a multi-document graph (MDG) ***G*** which preserves some of its properties while reducing the number of nodes. An example of such an operator is the *link-preserving graph cover operator* (Figure 7). Its effect is to preserve only these nodes from the source MDG that are associated with the preferred cross-document links. In the example, the shaded area represents the *summary subgraph* ***G'*** of ***G*** that contains all four cross-document links and only these nodes and edges of G which are necessary to preserve the textual structure of ***G'***.

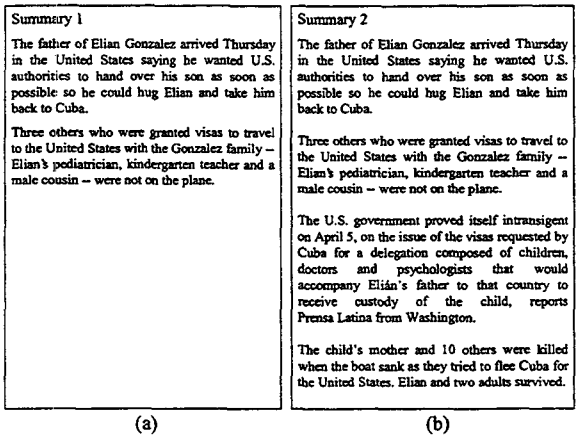

Figure 8: Two summaries from the same set of input documents

5.1 Example

The example in Figure 8 shows two summaries based on different user preferences. Summary (b) is based on "longer extract", "report background information", and "include all sources". Summary (a) is generated from two CNN articles, while (b) is generated from two CNN articles plus one from the Granma of Havana, and one from ABC News.

6 Ongoing work and conclusion

6.1 Ongoing work

We are in the process of performing a user study to collect interagreement data among judges who are asked to label cross-document rhetorical relations.

We are also currently building a system for automatic identification of relationships in document clusters as well as a library of summarization operators. User preferences are used to constrain the summarizers. For example, a user may prefer that in the event of contradiction, both sources of information should be represented in the summary. Another user may have preferences for a given source over all others and choose an operator which will only reflect his preferred source.

We will facilitate the user's navigation in the space of all possible summarizers. By specifying their preferences, users will build their own summarizers and test them on a collection of documents and then refine them to fit their needs.

6.2 Conclusion

We introduced a theory of cross-document structure based on inter-document relationships such as paraphrase, citation, attribution, modality, and development. We presented a taxonomy of cross-document links. We argued that a CST-based analysis of related documents can facilitate multi-document summarization.

References

James Allan. "Automatic hypertext link typing". *Hypertext'96, The Seventh ACM Conference on Hypertext*, pages 42—52.

James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang, "Topic detection and tracking pilot study: final report". *Proceedings of the Broadcast News Understanding and Transcription Workshop*, 1998.

Regina Barzilay and Michael Elhadad. "Using Lexical Chains for Text Summarization". *Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization.* Madrid, Spain, July 1997, Pages 10—17.

Regina Barzilay, Kathleen McKeown, and Michael Elhadad. "Information Fusion in the Context of Multi-Document Summarization". *ACL'99.* College Park, Maryland, June 1999.

Adam Berger and Robert Miller. "Just in Time Language Modelling". IEEE Conference on Acoustic, Speech and Signal Processing. Seattle, WA.

Jaime Carbonell and Jade Goldstein. "The use of MMR, diversity-based reranking for reordering documents and producing summaries". *Proceedings of ACM-SIGIR'98.* Melbourne, Australia, August 1998.

Jaime Carbonell, Mark Craven, Steve Fienberg, Tom Mitchell, and Yiming Yang. "Report on the CONALD Workshop on Learning from Text and the Web", Pittsburgh, PA, June 1998.

William Cohen and Yoram Singer. "Context-sensitive learning methods for text categorization". *Proceedings, 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Zurich, Switzerland, August 1996. Pages 307—315.

David Fisher, Stephen Soderland, Joseph McCarthy, Fangfang Feng, and Wendy Lehnert. "Description of the UMass System As Used for MUC-6". *Proceedings of the Sixth Message Understanding Conference (MUC-6).* 1995. Pages 221—236.

Marti Hearst. "Multi-Paragraph Segmentation of Expository Text". *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics.* Las Cruces, NM, June 1994.

Min-Yen Kan, Judith L. Klavans, and Kathleen McKeown. "Linear segmentation and segment relevance". *Proceedings of 6th International Workshop of Very Large Corpora (WVLC-6),* pages 197—205, Montreal, Quebec, Canada, August 1998.

Julian Kupiec, Jan Pedersen, and Francine Chen. "A Trainable Document Summarizer". *Proceedings, 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* Seattle, WA, July 1995.

Inderjeet Mani and Eric Bloedorn, "Summarizing Similarities and Differences Among Related Documents", Information Retrieval 1 (1-2), pages 35—67, June 1999.

William Mann and Sandra Thompson. "Rhetorical Structure Theory: Toward a functional theory of text organization". Text, 8(3). 243-281.

William Mann. Rhetorical Structure Theory Web Site. http://www.sil.org/linguistics/RST/

Daniel Marcu. "From Discourse Structures to Text Summaries". *Proceedings of the ACL/EACL Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain, July 1997.

Kathleen McKeown, Judith Klavans, Vasileios Hatzivassiloglou, Regina Barzilay, and Eleazar Eskin, "Towards Multidocument Summarization by Reformulation: Progress and Prospects", *Proceedings of AAAI'99*, Orlando, FL, July 1999.

Dragomir R. Radev and Kathleen McKeown. "Generating natural language summaries from multiple on-line sources". Computational Linguistics, 24 (3), pages 469—500, September 1998.

Dragomir R. Radev. "Topic Shift Detection - finding new information in threaded news". Technical Report CUCS-026-99, Columbia University Department of Computer Science. January 1999.

Gerard Salton, Chris Buckley and James Allan. "Automatic structuring of text files". Technical Report TR 91-1241, Computer Science Department, Cornell University, Ithaca, NY, 1991.

Gerard Salton, Amit Singhal, Mandar Mitra, Chris Buckley. "Automatic Text Structuring and Summarization". Information Processing and Management 33 (2), pages 193—207, 1997.

Randall Trigg. "A Network-Based Approach to Text Handling for the Online Scientific Community". Ph.D. Thesis. Department of Computer Science, University of Maryland. November 1983.

Randall Trigg and Mark Weiser. "TEXTNET: A network-based approach to text handling". ACM Transactions on Office Information Systems, 4 (1), pages 1—23, January 1987.