Abstract
We introduce CST (cross-document structuretheory), 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 how events they are reflected in new sources. First, many sources write about the same event. Second, the same source typically produces a number of accounts of the event over time. Sixteen stories related to the same event from six news sources over a two-hour period are represented in Figure 1.

Figure 1: Time distribution of related documents from multiple sources

A careful analysis of related 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. Sometimes a source adds new information as it becomes available and puts it in the context of what has already been discussed. In other cases, get full picture of an event from stories from multiple sources either present all relevant information. An example point 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 extracts are from news stories about the same event: the declaration by Bill Clinton at a press conference that millions of 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 concept links. Warguth argued that CST is essential for the analysis of contradiction, redundancy, and complementarity.

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. Clinton 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.

**Figure 2: Six different accounts of the same event**

### 2 Related Work

#### 2.1 Document structure

Rhetorical Structure Theory (RST) [Mann & Thompson 88, Manto 00] is a comprehensive theory of organization based on text coherence and organization of text. It has been extensively used in text analysis. In RST, the rhetorical structure of a document is described using a set of relations among sentences. These relations consist of a nucleus (the central component) and one or more satellites (the supporting components of the relation). An example of RST relations includes evidence, subsumption, and update. In this relation, the nucleus (evidence) is the claim that satisfies the requirement, and the satellites (subsumption and update) are the supporting components of the relation.

### 2.2 Cross-document analysis

CST is essential for the analysis of contradiction, redundancy, and complementarity. Cross-document links such as subsumption, update, and elaboration to cross-document concept links. Warguth argued that CST is essential for the analysis of contradiction, redundancy, and complementarity.

### 2.3 Citation, refutation, revision, equivalence, and comparison

Citations, refutations, revisions, equivalences, and comparisons are grouped into two categories: Normal (inter-document links) and Commentary (deliberate cross-document links). While the taxonomy is exhaustive, it means appropriate intended general domain texts (that is, 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 Trigg's list of link types from Trigg's list to links between sentences or paragraphs 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 technique actually achieve anything useful.

More recently [Salton et al. 97] introduced a technique for document structuring based on semantic hyperlinks (among pair of paragraphs which are related by a lexical similarity significantly higher than random). The authors represent 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 businesse of connecting edges of paragraphs to decide whether to include in a summary. The summarization technique described in Salton et al.'s research is limited to single documents.

One of the goals of CST is to extend the techniques described in 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 large number of documents. SUMMONS uses a set of semantic templates extracted by a natural language understanding system [Fishman 97] to identify 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 for a relatively small domain of text such as account of terrorism and is not directly suitable for domain-independent text analysis. The planning operator used in SUMMONS present, however, the idea first step towards CST.

[Mani & Bloedorn 99] uses similarities and differences among related news articles for MDST. They measure the effectiveness of their method in two scenarios; paragraph alignment across article and query-based information retrieval. None of these scenarios evaluate the generation of query-independent summaries of multiple articles in more domains.

Th 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 the "theme intersection" : paragraph alignment across stories with the help of semantic network to identify phrases which convey the same meaning. The author generates new sentences from each theme and orders them chronologically to produce a summary.

We should also note 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 includes (a) evolving summaries (summaries of new documents related to an ongoing event that are presented to the user assuming that he has read earlier summaries of related documents) [Radev 99] and (b) chronological briefings [Radev & McKeown 98]. Carbonell et al. discussed the motivation behind the use of time-dependent documents in the language model of 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 $U$ (see Figure 3(b)) is a tuple $(t, s, p)$ defined at different levels of granularity, e.g., paragraphs, sentences, words.

**Definition** A document $D$ is a sequence of document units $U_1 U_2 \ldots U_n$ which correspond to a one-dimensional projection of a multi-document cube along the source and time dimensions.

Some additional concepts are defined based on the above definitions.

**Definition** A snapshot is a slice of the multi-document cube over a period of time $\Delta t$ (see Figure 3(c)).

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

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

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

![Figure 3](image)

- (a) multi-document cube
- (b) document unit
- (c) cube slice
- (d) 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 subgraphs for each individual document (Figure 4). Two types of links exist. The first type represents inheritance relationships among entities 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 words.
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).

| # | 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                | S, P     | Similar information content in different languages                         |
| 4 | Subsumption                | S        | One sentence contains more information than another                         |
| 5 | Contradiction              | S, D     | Conflict in information                                                    |
| 6 | Historical background      | S        | Information about another document                                          |
| 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 information on another while adding attribution        |
| 11| Summary                    | S, D     | Similar to RST: another document summarizes                                  |
| 12| Follow-up                  | S        | Additional information which reflects or updates the account               |

Figure 6: Sample multi-document graph
|   | Type                  | Action                           | Example                                                                 |
|---|----------------------|----------------------------------|-------------------------------------------------------------------------|
| 13| Elaboration          | Additional information that wasn't included in the last account. |
| 14| Indirect Speech      | Shift from direct to indirect speech. Versa versa. |
| 15| Refinement           | Additional information that is more specific than previously included. |
| 16| Agreement            | One source expresses agreement with another. |
| 17| Judgment             | A qualified account conflict. |
| 18| Fulfilment           | A prediction turns true. |
| 19| Description          | Insertion description. |
| 20| Reader Profile       | Style and background-specific change. |
| 21| Contrast             | Comparing two accounts. |
| 22| Parallel             | Comparing two accounts. |
| 23| Generalization       | Generalization. |
| 24| Change of perspective| There is a perspective change. |

**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 information presented in one sentence may 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 murder.
(2) The court found John Doe guilty of murder last August and sentenced him to life.

**Paraphrase**

(3) Ford’s program will be launched in the United States in spring globally within 12 months.
(4) Ford announced that it will launch a new program for its employees in the United States and then expand the program abroad.

**Modality**

New stories are often written in a way that makes misattributions of information difficult. e.g., referring to persons arrested for crimes, the actual incident may not be 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 that Russia’s space program is recovering, space officials announced today that Roscosmos will launch a new satellite in August and will remain manned. (10) The Discovery spacecraft will leave Earth and return by August 2018.

**Indirect Speech**
An anonymous caller told 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.

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."

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

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.

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.

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

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

The Federal Bureau of Investigation plans to put suspected terrorist Osama bin Laden, sought in connection with the bombings of the US embassies in Africa, on its "Ten Most Wanted" list, CNN reported Saturday.

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 Most Wanted List Monday.

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

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

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

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 query-independent (e.g., based on pure document similarity [Allan et al. 98]).
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 are 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 matchings, 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 such as perpetrator and effect of 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 based on identifying some relationships such as citation, contradiction, and attribution is template matching, which takes into account transformation grammars (e.g., relative clause insertion). For the link type analysis, machine learning in lexical and cues 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) $G$ that 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 which are necessary to preserve the textual structure of $G'$.

### 5.1 Example

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

### 6 Ongoing Work and Conclusion

#### 6.1 Ongoing Work

We are in the process of performing user study to collect agreement data among human judges who asked to cross-document rhetorical relations. We also are currently building a system for automatic identification of relationships in documents clusters as well as a library of summarization operators. User preferences are used to constrain the summarizers. For example, if a user prefers that the event of contradiction, both sources of information should be represented in the summary. Another user may prefer that only one source ever be chosen, which is reflected by the preference of the user.

We will facilitate user 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 refine them to fit their needs.

#### 6.2 Conclusion

We introduced the 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 and used CST-based analysis to related documents to facilitate multi-document summarization.
References

James Allan. “Automatichyperlinktyping”. Hypertext’96 the Seventh ACM Conference on Hypertext, pages 2—52.

James Allan,Carbonel,GeorgeDoddington, JonathanYamron,andYimingYang,”Topic detectionandracktrackingpilotstudyinreport”. 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 9 Workshop on Intelligent Scalable Text Summarization, Madrid, Spain1999, pages 10—17.

Regina Barzilay, Kathleen McKeown, and Michael Elhadad. “Information Fusion in the Context of Multi-Document Summarization”. ACL ’99, College Park, Maryland, June1999.

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

James Carbonell and Jade Goldstein. “Thesos MMR: Diversity-based ranking for ordering documents and producing summaries”. Proceedings of ACM-SIGIR ‘98. Melbourne, Australia, August 1998.

James Carbonell, Mark Craven, Steve Fienberg, and Randall Trigg. “Learning from Text and the Web”. Technical Report University of Arizona, June 1998.

David Fisher, Stephen Soderland, Joseph McCarthy, Fenng, and Vendi Chmert. “DescriptionalMasSystemAlUsefor MUC-6”. Proceedings of the Sixth Message Understanding Conference (MUC-6), 1995, pages 281—236.

Marie Hearst. “Multi-Paragraph Segmentation of Expository Text”. Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Las Cruces, NM, June 1994.

Min-Yen Kan, Judith Klawans, and Kathleen McKeown. “Line segmentation and segmentation relevance”. Proceedings of the Second International Workshop Very Large Corpora (WLC-6), pages 197—205. Montreal, Quebec, Canada, August 1998.

Julia Kupiec, Peter Pedersen, and Franci Chen. “A Trainable Document Summarizer”. Proceedings, 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, WA, June 1995.

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

William Mann and Andrew Thompson. “Rhetorical Structure Theory Toward a Function Theory of Text Organization”. Text 13 (3), pages 243-281.

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

Dan Metcalf. “From Discourse Structures to Text Summaries”. Proceedings of the ACL/EACL Workshop on Intelligent Scalable Text Summarization, Madrid, Spain, July 1997.

Kathleen McKeown, Judith Klawans, Vasileios Hatzivassiloglou, Regina Barzilay, Eleazar Eskin, “Towards Multidocument Summarization by Reformulation: Progress and Prospects”. Proceedings of AAAI ’99, Orlando, FL, July 1999.

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

Dragomir 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. “Automatizing the Structural Extraction of Expository Text”. Technical Report TR 91-1241, Computer Science Department, Cornell University, Ithaca, NY, 1991.

Gerard Salton, Atri Singh, Abhijit Mitra, and Chris Buckley. “Automatic Text Structuring and Summarization”. Information Processing and Management, 28(2), pages 493—207, 1992.

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, 1993.

Randall Trigg and Mark Weiser. “TEXTNET: A Network-based Approach to Text Handling”. ACM Transactions on Office Information Systems, 1(1), 1—23, January 1987.