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).

![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 $U_1; U_2; ...; U_n$ 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 $\Delta t$ – see Figure 3 (c).

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

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).

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

Figure 4: Sample multi-document graph
<table>
<thead>
<tr>
<th></th>
<th>Elaboration</th>
<th>S</th>
<th>Additional information that wasn’t included in the last account</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>Indirect speech</td>
<td>S</td>
<td>Shift from direct to indirect speech or vice-versa</td>
</tr>
<tr>
<td>14</td>
<td>Refinement</td>
<td>S</td>
<td>Additional information that is more specific that the one previously included</td>
</tr>
<tr>
<td>15</td>
<td>Agreement</td>
<td>S</td>
<td>One source expresses agreement with another</td>
</tr>
<tr>
<td>16</td>
<td>Judgment</td>
<td>S</td>
<td>A qualified account of a fact</td>
</tr>
<tr>
<td>17</td>
<td>Fulfilment</td>
<td>S</td>
<td>A prediction turned true</td>
</tr>
<tr>
<td>18</td>
<td>Description</td>
<td>S</td>
<td>Insertion of a description</td>
</tr>
<tr>
<td>19</td>
<td>Reader profile</td>
<td>S</td>
<td>Style and background-specific change</td>
</tr>
<tr>
<td>20</td>
<td>Contrast</td>
<td>S</td>
<td>Contrasting two accounts or facts</td>
</tr>
<tr>
<td>21</td>
<td>Parallel</td>
<td>S</td>
<td>Comparing two accounts of facts</td>
</tr>
<tr>
<td>22</td>
<td>Generalization</td>
<td>S</td>
<td>Generalization</td>
</tr>
<tr>
<td>23</td>
<td>Change of perspective</td>
<td>S, D</td>
<td>The same source presents a fact in a different light</td>
</tr>
</tbody>
</table>

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

**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 cosmonaunts 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."

(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.

(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.

(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.


(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.

(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.

(23) 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.

(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.

(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.

(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](image)

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]).

![Figure 7: Summarization using graph cover operators](image)
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](image)

**Summary 1**
The father of Elián González arrived Thursday in the United States saying he wanted U.S. authorities to hand over his son as soon as possible so he could hug Elián and take him back to Cuba. Three others who were granted visas to travel to the United States with the González family — Elián's pediatrician, kindergarten teacher and a male cousin — were not on the plane.

**Summary 2**
The father of Elián González arrived Thursday in the United States saying he wanted U.S. authorities to hand over his son as soon as possible so he could hug Elián and take him back to Cuba. Three others who were granted visas to travel to the United States with the González family — Elián's pediatrician, kindergarten teacher and a male cousin — were not on the plane.

The U.S. government proved itself intragovernmental on April 5, on the issue of the visas requested by Cuba for a delegation composed of children, doctors and psychologists that would accompany Elián's father to that country to receive custody of the child, reports Prensa Latina from Washington.

The child's mother and 10 others were killed when the boat sank as they tried to flee Cuba for the United States. Elián and two adults survived.

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


