What is Text?

A product of cohesive ties (cohesion)

ATHENS, Greece (Ap) A strong earthquake shook the Aegean Sea island of Crete on Sunday but caused no injuries or damage. The quake had a preliminary magnitude of 5.2 and occurred at 5:28 am (0328 GMT) on the sea floor 70 kilometers (44 miles) south of the Cretan port of Chania. The Athens seismological institute said the temblor’s epicenter was located 380 kilometers (238 miles) south of the capital. No injuries or damage were reported.

What is Text?

A product of structural relations (coherence)

Regina Barzilay
{regina}@csail.mit.edu
March 1, 2003

Content-based Structure

- Describe the strength and the impact of an earthquake
- Specify its magnitude
- Specify its location
- ...

Domain-dependent Text Structures

ATHENS, Greece (Ap) A strong earthquake shook the Aegean Sea island of Crete on Sunday but caused no injuries or damage. The quake had a preliminary magnitude of 5.2 and occurred at 5:28 am (0328 GMT) on the sea floor 70 kilometers (44 miles) south of the Cretan port of Chania. The Athens seismological institute said the temblor’s epicenter was located 380 kilometers (238 miles) south of the capital. No injuries or damage were reported.

Domain-dependent Text Structures

What is Text?

A product of structural relations (coherence)

$S_1$: A strong earthquake shook the Aegean Sea island of Crete on Sunday

$S_2$: but caused no injuries or damage.

$S_3$: The quake had a preliminary magnitude of 5.2
**Analogies with Syntax**

Domain-independent Theory of Sentence Structure

- Fixed set of word categories (nouns, verbs, ...)
- Fixed set of relations (subject, object, ...)

P(“A is sentence this weird”)

---

**Motivation**

- Summarization
  Extract a representative subsequence from a set of sentences
- Question-Answering
  Find an answer to a question in natural language
- Text Ordering
  Order a set of information-bearing items into a coherent text
- Machine Translation
  Find the best translation taking context into account

---

**Rhetorical Structure**

- Contrast
- Elaboration

---

**Two Approaches to Text Structure**

- Domain-dependent models (Today)
  - Content-based models
  - Rhetorical models
- Domain-independent models
  - Rhetorical Structure Theory (Next Class)
Today: Domain-Specific Models

- Rhetorical Models:
  - Argumentative Zoning of Scientific Articles (Teufel, 1999)

- Content-based Models:
  - Supervised (Duboue & McKeown, 2001)
  - Unsupervised (Barzilay & Lee, 2004)

Motivation

- Scientific articles exhibit (consistent across domains) similarity in structure
  - BACKGROUND
  - OWN CONTRIBUTION
  - RELATION TO OTHER WORK

- Automatic structure analysis can benefit:
  - Q&A
  - summarization
  - citation analysis

Argumentative Zoning

Many of the recent advances in Question Answering have followed from the insight that systems can benefit from by exploiting the redundancy in large corpora.

Brill et al. (2001) describe using the vast amount of data available on the WWW to achieve impressive performance . . .

The Web, while nearly infinite in content, is not a complete repository of useful information . . .

In order to combat these inadequacies, we propose a strategy in which in information is extracted from . . .
Examples

<table>
<thead>
<tr>
<th>Category</th>
<th>Realization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aim</td>
<td>We have proposed a method of clustering words based on large corpus data</td>
</tr>
<tr>
<td>Textual</td>
<td>Section 2 describes three parsers which are . . .</td>
</tr>
<tr>
<td>Contrast</td>
<td>However, no method for extracting the relationship from superficial linguistic expressions was described in their paper.</td>
</tr>
</tbody>
</table>

Features

- Position
- Verb Tense and Voice
- History
- Lexical Features (“other researchers claim that”)

Approach

- Goal: Rhetorical segmentation with labeling
- Annotation Scheme:
  - Own work: aim, own, textual
  - Background
  - Other Work: contrast, basis, other
- Implementation: Classification

Kappa Statistics

(Siegel & Castellan, 1998; Carletta, 1999)
Kappa controls agreement $P(A)$ for chance agreement $P(E)$

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

Kappa from Argumentative Zoning:

- Stability: 0.83
- Reproducibility: 0.79
Supervised Content Modeling

(Duboue & McKeown, 2001)

- Goal: Find types of semantic information characteristic to a domain and ordering constraints on their presentation
- Approach: find patterns in a set of transcripts manually annotated with semantic units
- Domain: Patients records

Semantic Sequence

age, gender, pmh, pmh, pmh, pmh, med-preop, med-preop, med-preop, drip-preop, med-preop, ekg-preop, echo-preop, hct-preop, procedure, ...

Annotated Transcript

He is 58-year-old male. History is significant for Hodgkin's disease, pmh, treated with ... to his neck, back and chest. Hyperspadias, BPH, pmh, hiatal hernia and proliferative lymph edema in his right arm. No IV's pmh, or blood pressure down in the left arm. Medications — Inderal, Lopid, pmh, Pepcid, nitroglycerine and heparin. EKG has PAC's, ...

Results

- Classification accuracy is above 70%
- Zoning improves classification
**Example of Learned Pattern**

- intraop-problems
- intraop-problems
- operation 11.11%
- drip 33.33%
- intraop-problems 33.33%
- total-meds-anesthetics 22.22%
- drip

---

**Pattern Detection**

Analogous to motif detection

$T_1$: A B C D F A A B F D

$T_2$: F C A B D D F F

- Scanning
- Generalizing
- Filtering

---

**Content Models**

(Barzilay & Lee, 2004)

- Content models represent topics and their ordering in text.

<table>
<thead>
<tr>
<th>Domain: newspaper articles on earthquake</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topics: “strength”, “location”, “casualties”, …</td>
</tr>
<tr>
<td>Order: “casualties” prior to “rescue efforts”</td>
</tr>
</tbody>
</table>

- Assumption: Patterns in content organization are recurrent

---

**Evaluation**

Pattern confidence: 84.62%

Constraint accuracy: 89.45%
**Similarity in Domain Texts**

TOKYO (AP) A moderately strong earthquake with a preliminary magnitude reading of 5.1 rattled northern Japan early Wednesday, the Central Meteorological Agency said. There were no immediate reports of casualties or damage. The quake struck at 6:06 am (2106 GMT) 60 kilometers (36 miles) beneath the Pacific Ocean near the northern tip of the main island of Honshu. . . .

ATHENS, Greece (AP) A strong earthquake shook the Aegean Sea island of Crete on Sunday but caused no injuries or damage. The quake had a preliminary magnitude of 5.2 and occurred at 5:28 am (0328 GMT) on the sea floor 70 kilometers (44 miles) south of the Cretan port of Chania. The Athens seismological institute said the temblor’s epicenter was located 380 kilometers (238 miles) south of the capital. . . .

---

**Computing Content Model**

Implementation: Hidden Markov Model

- States represent topics
- State-transitions represent ordering constraints

![State diagram](image.png)

---

**Similarity in Domain Texts**

TOKYO (AP) A moderately strong earthquake with a preliminary magnitude reading of 5.1 rattled northern Japan early Wednesday, the Central Meteorological Agency said. There were no immediate reports of casualties or damage. The quake struck at 6:06 am (2106 GMT) 60 kilometers (36 miles) beneath the Pacific Ocean near the northern tip of the main island of Honshu. . . .

ATHENS, Greece (AP) A strong earthquake shook the Aegean Sea island of Crete on Sunday but caused no injuries or damage. The quake had a preliminary magnitude of 5.2 and occurred at 5:28 am (0328 GMT) on the sea floor 70 kilometers (44 miles) south of the Cretan port of Chania. The Athens seismological institute said the temblor’s epicenter was located 380 kilometers (238 miles) south of the capital. No injuries or damage were reported.

---

**Narrative Grammars**

- Propp (1928): fairy tales follow a “story grammar”
- Barlett (1932): formulaic text structure facilities reader’s comprehension
- Wray (2002): texts in multiple domains exhibit significant structural similarity
**Initial Topic Induction**

Agglomerative clustering with cosine similarity measure

The Athens seismological institute said the temblor's epicenter was located 380 kilometers (238 miles) south of the capital.

Seismologists in Pakistan’s Northwest Frontier Province said the temblor’s epicenter was about 250 kilometers (155 miles) north of the provincial capital Peshawar.

The temblor was centered 60 kilometers (35 miles) northwest of the provincial capital of Kunming, about 2,200 kilometers (1,300 miles) southwest of Beijing, a bureau seismologist said.

**Estimating Emission Probabilities**

State $s_i$ emission probability:

\[ p_{s_i}(w_0, ..., w_n) = \prod_{j=0}^{n} p_{s_i}(w_j|w_{j-1}) \]

- Estimation for a “normal” state:

\[ p_{s_i}(w'|w) \overset{def}{=} \frac{f_{c_i}(ww')}{f_{c_i}(w)} + \delta_1 |V|, \]

- Estimation for the “insertion” state:

\[ p_{sm}(w'|w) \overset{def}{=} \frac{1 - \max_{i<m} p_{s_i}(w'|w)}{\sum_{u \in V} (1 - \max_{i<m} p_{s_i}(u|w))}. \]

**Model Construction**

- Initial topic induction
- Determining states, emission and transition probabilities
- Viterbi re-estimation

**From Clusters to States**

- Each large cluster constitutes a state
- Agglomerate small clusters into an “insert” state
Estimating Transition Probabilities

\[ p(s_j|s_i) = \frac{g(c_i, c_j) + \delta_2}{g(c_i) + \delta_2m} \]

\( g(c_i, c_j) \) is a number of adjacent sentences \((c_i, c_j)\)
\( g(c_i) \) is a number of sentences in \(c_i\)

Viterbi re-estimation

Goal: incorporate ordering information

- Decode the training data with Viterbi decoding

- Use the new clustering as the input to the parameter estimation procedure

Application: Information Ordering

- Input: set of sentences
- Applications:
  - Text summarization
  - Natural Language Generation
- Goal: Recover most likely sequences “get marry” prior to “give birth” (in some domains)

Information Ordering: Algorithm

Input: set of sentences

- Produce all permutations of the set
- Rank them based on the content model
**Summarization: Algorithm**

Input: source text
Training data: parallel corpus of summaries and source texts (aligned)

- Employ Viterbi on source texts and summaries
- Compute state likelihood to generate summary sentences:
  \[ p(s \in \text{summary}|s \in \text{source}) = \frac{\text{summary\_count}(s)}{\text{source\_count}(s)} \]
- Given a new text, decode it and extract sentences corresponding to “summary” states

**Baselines for Ordering**

- “Straw” baseline: Bigram Language model
- “State-of-the-art” baseline: (Lapata:2003)
  - represent a sentence using lexico-syntactic features
  - compute pairwise ordering preferences
  - find optimally global order

**Application: Summarization**

- Domain-dependent summarization: (Radev&McKeown:1998)
  - specify types of important information (manually)
  - use information extraction to identify this information (automatically)
- Domain-independent summarization: (Kupiec et al:1995)
  - represent a sentence using shallow features
  - use a classifier

**Evaluation: Data**

<table>
<thead>
<tr>
<th>Domain</th>
<th>Average Length</th>
<th>Vocabulary Size</th>
<th>Token/ type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Earthquake</td>
<td>10.4</td>
<td>1182</td>
<td>13.158</td>
</tr>
<tr>
<td>Clashes</td>
<td>14</td>
<td>1302</td>
<td>4.464</td>
</tr>
<tr>
<td>Drugs</td>
<td>10.3</td>
<td>1566</td>
<td>4.098</td>
</tr>
<tr>
<td>Finance</td>
<td>13.7</td>
<td>1378</td>
<td>12.821</td>
</tr>
<tr>
<td>Accidents</td>
<td>11.5</td>
<td>2003</td>
<td>5.556</td>
</tr>
</tbody>
</table>
Baselines for Summarization

- “Straw” baseline: $n$ leading sentences
- “State-of-the-art” Kupiec-style classifier:
  - Sentence representation: lexical features and location
  - Classifier: BoosTexter

Results: Summarization

<table>
<thead>
<tr>
<th>Summarizer</th>
<th>Extraction accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content-based</td>
<td>88%</td>
</tr>
<tr>
<td>Sentence classifier</td>
<td>76%</td>
</tr>
<tr>
<td>(words + location)</td>
<td></td>
</tr>
<tr>
<td>Leading $n$ sentences</td>
<td>69%</td>
</tr>
</tbody>
</table>

Results: Ordering

<table>
<thead>
<tr>
<th>Domain</th>
<th>Algorithm</th>
<th>Prediction Accuracy</th>
<th>Rank</th>
<th>$\tau$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Earthquake</td>
<td>Content</td>
<td>72%</td>
<td>2.67</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>Lapata’03</td>
<td>24% (N/A)</td>
<td>0.48</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bigram</td>
<td>4%</td>
<td>485.16</td>
<td>0.27</td>
</tr>
<tr>
<td>Clashes</td>
<td>Content</td>
<td>48%</td>
<td>3.05</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>Lapata’03</td>
<td>27% (N/A)</td>
<td>0.41</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bigram</td>
<td>12%</td>
<td>635.15</td>
<td>0.25</td>
</tr>
<tr>
<td>Drugs</td>
<td>Content</td>
<td>38%</td>
<td>15.38</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>Lapata’03</td>
<td>27% (N/A)</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bigram</td>
<td>11%</td>
<td>712.03</td>
<td>0.24</td>
</tr>
<tr>
<td>Finance</td>
<td>Content</td>
<td>96%</td>
<td>0.05</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>Lapata’03</td>
<td>17% (N/A)</td>
<td>0.44</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bigram</td>
<td>66%</td>
<td>7.44</td>
<td>0.74</td>
</tr>
<tr>
<td>Accidents</td>
<td>Content</td>
<td>41%</td>
<td>10.96</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>Lapata’03</td>
<td>10% (N/A)</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bigram</td>
<td>2%</td>
<td>973.75</td>
<td>0.19</td>
</tr>
</tbody>
</table>

Ordering: Learning Curve
Summarization: Learning Curve

Summary/source training set size vs. summarization accuracy for different methods:
- hmm-based
- word+loc
- lead