**Domain-Dependent Content Models**

- Capture topics and their distribution
- Based on pattern matching techniques
  - Motifs of semantic units
  - Distributional model
- Useful in generation and summarization

**Domain-Dependent Rhetorical Model**

**Domain:** Scientific Articles

- Human exhibit high agreement on the annotation scheme
- The scheme covers only a small fraction of discourse relations

---

**Domain-independent Models of Text Structure**

**Regina Barzilay**

{regina}@csail.mit.edu

**March 3, 2003**
**Rhetorical Structure Theory**

(Mann&Thompson:1988, Matthessen&Thompson:1988)

- Developed in the framework of natural language generation
- Aims to describe “building blocks” of text structure
  - Nucleus vs Satellites
  - Binary Relations between Discourse Units
- Compositionality principle define how to build a tree from binary relations

**Domain-Independent Rhetorical Model**

- Model elements:
  - Binary Relations
  - Compositionality Principle
- Requirements:
  - Stability and Reproducibility of an Annotation Scheme
  - Expressive Power of a Model

**Binary Relations**

- (JUSTIFICATION, A, B)
- (JUSTIFICATION, D, B)
- (EVIDENCE, C, B)
- (CONCESSION, C, D)
- (RESTATEMENT, D, A)

**Example**

[ No matter how much one wants to stay a non-smoker, A ], [ the truth is that the pressure to smoke in junior high is greater than it will be any other time of one’s life. B ] . [ We know that 3,000 teens start smoking each day, C ] [ although it is a fact that 90% of them once thought that smoking was something that they'll never do. D ]
Compositionality

Whenever two large text spans are connected through a rhetorical relation, that rhetorical relation holds between the most important parts of the constituent spans.

Marcu (1997): used constraint-satisfaction approach to build discourse trees given a set of binary relations.

Ambiguity

John can open the safe. He knows the combination.

Relations

<table>
<thead>
<tr>
<th>Relation</th>
<th>Nucleus</th>
<th>Satellite</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background</td>
<td>text whose understanding is being facilitated</td>
<td>text whose understanding is being facilitated</td>
</tr>
<tr>
<td>Elaboration</td>
<td>basic information</td>
<td>additional information</td>
</tr>
<tr>
<td>Preparation</td>
<td>text to be presented</td>
<td>text which prepares the reader to expect and interpret the text to be presented</td>
</tr>
</tbody>
</table>
**Automatic Computation of RST Relations**

(Marcu, 1997)

- Aggregate discourse relations to a few stable groups: (contrast, elaboration, condition, cause-explanation-evidence)
- Establish deterministic correspondence between cue phrases and discourse relations:
  - { But, However } → Contrast
  - { In addition, Moreover } → Elaboration

**Other Words Also Count!**

(Marcu & Echihabi, 2002)

Surface cues for discourse relations:

I like vegetables, but I hate tomatoes.

**Accuracy**

- Compared against manually constructed trees
- Tested against human-constructed trees
- Automatically constructed trees exhibit high similarity with human-constructed trees
- However, see (Marcu & Echihabi, 2002) CONTRAST vs ELABORATION: only 61 from 238 have a discourse marker (26%)
Method

- Assume that certain markers unambiguously predict discourse relations
- Create Cartesian product of words located on two sides of a discourse marker
- For each pair of words, compute its likelihood to predict a discourse relations

$$\text{argmax}_{r_k} P(r_k|(s_1, s_2)) = \text{argmax}_{r_k} P((s_1, s_2)|r_k) \times P(r_k), \text{ where}$$
$$P((s_1, s_2)|r_k) = \prod_{i,j \in s_1, s_2} P((w_i, w_j)|r_k)$$