Domain-dependent Text Structures

Regina Barzilay

{regina}@csail.mit.edu

March 1, 2003

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)

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

Content-based Structure

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

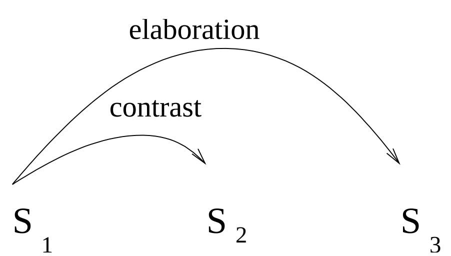
Analogy with Syntax

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”)

Rhetorical Structure



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

Two Approaches to Text Structure

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

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 . . .

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

Today: Domain-Specific Models

- Rhetorical Models:
  - Argumentative Zoning of Scientific Articles (Teufel, 1999)
- Content-based Models:
  - Supervised (Duboue&McKeown, 2001)
  - Unsupervised (Barzilay&Lee, 2004)

Argumentative Zoning

BACKGROUND
Many of the recent advances in Question Answering have followed from the insight that systems can benefit from by exploiting the redundancy . . .
OTHER WORK
Brill et al. (2001) describe using the vast amount of data available on the WWW to achieve impressive performance . . .
WEAKNESS
The Web, while nearly infinite in content, is not a complete repository of useful information . . .
OWN CONTRIBUTION
In order to combat these inadequacies, we propose a strategy in which in information is extracted from . . .

Examples

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

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

(Siegal&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, . . .

Results

- Classification accuracy is above 70%
- Zoning improves classification

Annotated Transcript

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

Example of Learned Pattern

intraop-problems
intraop-problems

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

drip

Content Models

(Barzilay&Lee, 2004)

- Content models represent topics and their ordering in text.

> Domain: newspaper articles on earthquake
> Topics: “strength”, “location”, “casualties”, . . .
> Order: “casualties” prior to “rescue efforts”

- Assumption: Patterns in content organization are recurrent

Pattern Detection

Analogous to motif detection

T_1: $\underline{\text{A B C D}}$ F A $\underline{\text{A B F D}}$

T_2: F C $\underline{\text{A B D D}}$ F F

- Scanning
- Generalizing
- Filtering

Evaluation

Pattern confidence: 84.62%

Constraint accuracy: 89.45%

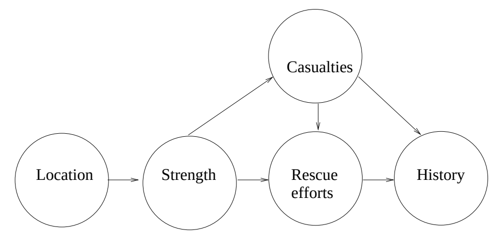
Similarity in Domain Texts

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 k ilometers (238 miles) south of the capital. . . .

Computing Content Model

Implementation: Hidden Markov Model

- States represent topics
- State-transitions represent ordering constraints



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 k ilometers (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

(Iyer&Ostendorf:1996,Florian&Yarowsky:1999, Barzilay&Elhadad:2003)

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:

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

- Estimation for the "insertion" state:

$$p_{s_m}(w'|w) \stackrel{def}{=} \frac{1 - \max_{i<m} p_{s_i}(w'|w)}{\sum_{u \in V}(1 - \max_{i<m} p_{s_i}(u|w))}.$$

Model Construction

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

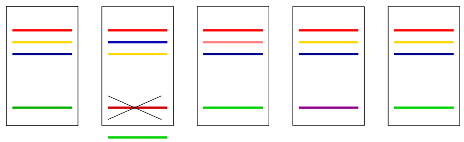
From Clusters to States

- Each large cluster constitutes a state
- Agglomerate small clusters into an *"insert"* state

Viterbi re-estimation

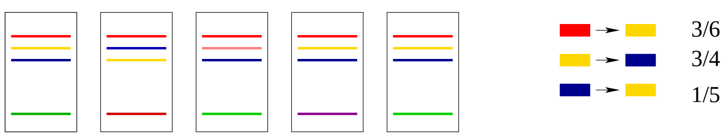
Goal: incorporate ordering information

- Decode the training data with Viterbi decoding



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

Estimating Transition Probabilities



$$p(s_j|s_i) = \frac{g(c_i, c_j) + \delta_2}{g(c_i) + \delta_2 m}$$

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

Information Ordering: Algorithm

Input: set of sentences

- Produce all permutations of the set
- Rank them based on the content model

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)

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:

$$p(s \in summary | s \in source) = \frac{summary\_count(s)}{source\_count(s)},$$

- Given a new text, decode it and extract sentences corresponding to "summary" states

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

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

Evaluation: Data

Domain	Average Length	Vocabulary Size	Token/ type
Earthquake	10.4	1182	13.158
Clashes	14	1302	4.464
Drugs	10.3	1566	4.098
Finance	13.7	1378	12.821
Accidents	11.5	2003	5.556

Baselines for Summarization

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

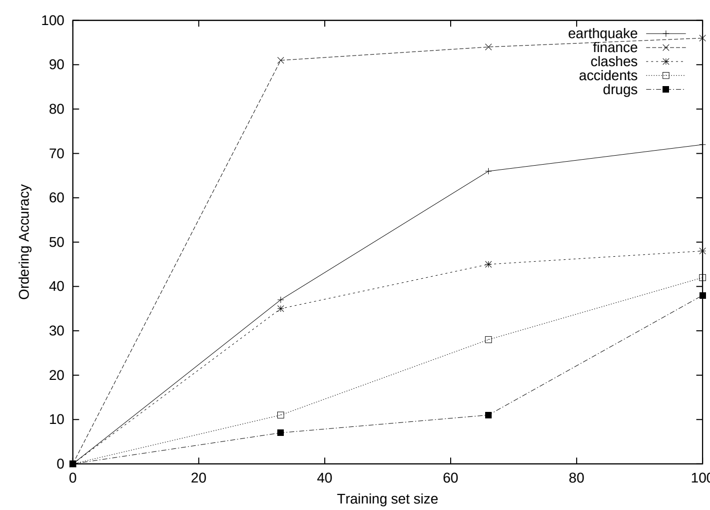
Results: Summarization

Summarizer	Extraction accuracy
Content-based	**88%**
Sentence classifier (words + location)	76%
Leading n sentences	69%

Results: Ordering

Domain	Algorithm	Prediction Accuracy	Rank	τ
Earthquake	Content	**72%**	**2.67**	**0.81**
	Lapata '03	24%	(N/A)	0.48
	Bigram	4%	485.16	0.27
Clashes	Content	**48%**	**3.05**	**0.64**
	Lapata '03	27%	(N/A)	0.41
	Bigram	12%	635.15	0.25
Drugs	Content	**38%**	**15.38**	0.45
	Lapata '03	27%	(N/A)	**0.49**
	Bigram	11%	712.03	0.24
Finance	Content	**96%**	**0.05**	**0.98**
	Lapata '03	17%	(N/A)	0.44
	Bigram	66%	7.44	0.74
Accidents	Content	**41%**	**10.96**	**0.44**
	Lapata '03	10%	(N/A)	0.07
	Bigram	2%	973.75	0.19

Ordering: Learning Curve



Summarization: Learning Curve