Discourse Segmentation

Ling575
Discourse and Dialogue
April 20, 2011

Roadmap

- Project updates and milestones
- Automatic Discourse Segmentation
  - Linear Segmentation
    - Unsupervised techniques
    - Supervised techniques
    - Segmentation evaluation
  - Discourse Parsing
  - Discourse Relation Extraction
    - D-LTAG

Project Presentations

- Spread over April 27, May 4
- Literature review
  - At least 3 papers
  - Identify 1 as primary
    - Everyone should read
- Relation to project, project plan

Recognizing Discourse Structure

- Decompose text into subunits

Recognizing Discourse Structure

- Decompose text into subunits
- Questions:
  - What type of structure is derived?

Recognizing Discourse Structure

- Decompose text into subunits
- Questions:
  - What type of structure is derived?
    - Sequential spans, hierarchical trees, arbitrary graphs
  - What is the granularity of the subunits?

Recognizing Discourse Structure

- Decompose text into subunits
- Questions:
  - What type of structure is derived?
    - Sequential spans, hierarchical trees, arbitrary graphs
  - What is the granularity of the subunits?
    - Clauses? Sentences? Paragraphs?
  - What information guides segmentation?

Recognizing Discourse Structure

- Decompose text into subunits
- Questions:
  - What type of structure is derived?
    - Sequential spans, hierarchical trees, arbitrary graphs
  - What is the granularity of the subunits?
    - Clauses? Sentences? Paragraphs?
  - What information guides segmentation?
    - Local cue phrases?

Recognizing Discourse Structure

- Decompose text into subunits
- Questions:
  - What type of structure is derived?
    - Sequential spans, hierarchical trees, arbitrary graphs
  - What is the granularity of the subunits?
    - Clauses? Sentences? Paragraphs?
  - What information guides segmentation?
    - Local cue phrases? Lexical cohesion?

Recognizing Discourse Structure

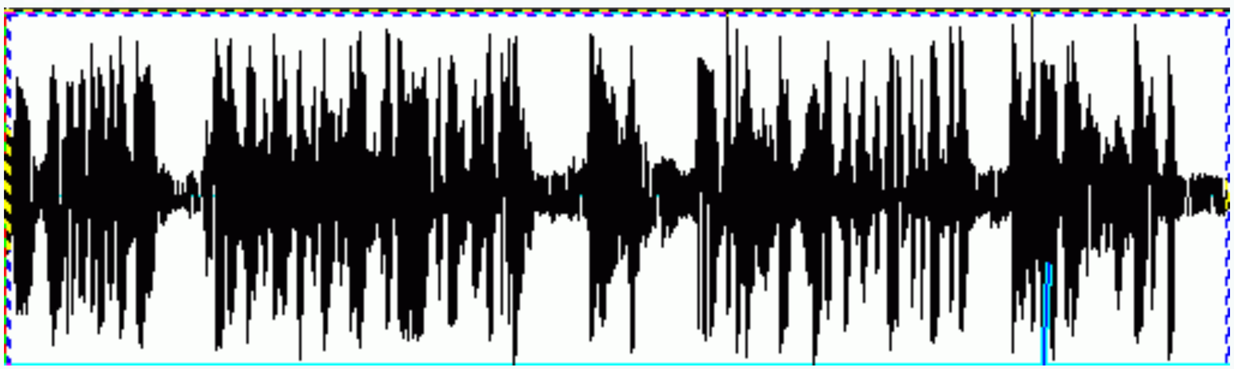
- Decompose text into subunits
- Questions:
  - What type of structure is derived?
    - Sequential spans, hierarchical trees, arbitrary graphs
  - What is the granularity of the subunits?
    - Clauses? Sentences? Paragraphs?
  - What information guides segmentation?
    - Local cue phrases? Lexical cohesion?
      - How is the information modeled? Learned?

Recognizing Discourse Structure

- Decompose text into subunits
- Questions:
  - What type of structure is derived?
    - Sequential spans, hierarchical trees, arbitrary graphs
  - What is the granularity of the subunits?
    - Clauses? Sentences? Paragraphs?
  - What information guides segmentation?
    - Local cue phrases? Lexical cohesion?
      - How is the information modeled? Learned?
  - How do we evaluate the results?

Discourse Topic Segmentation

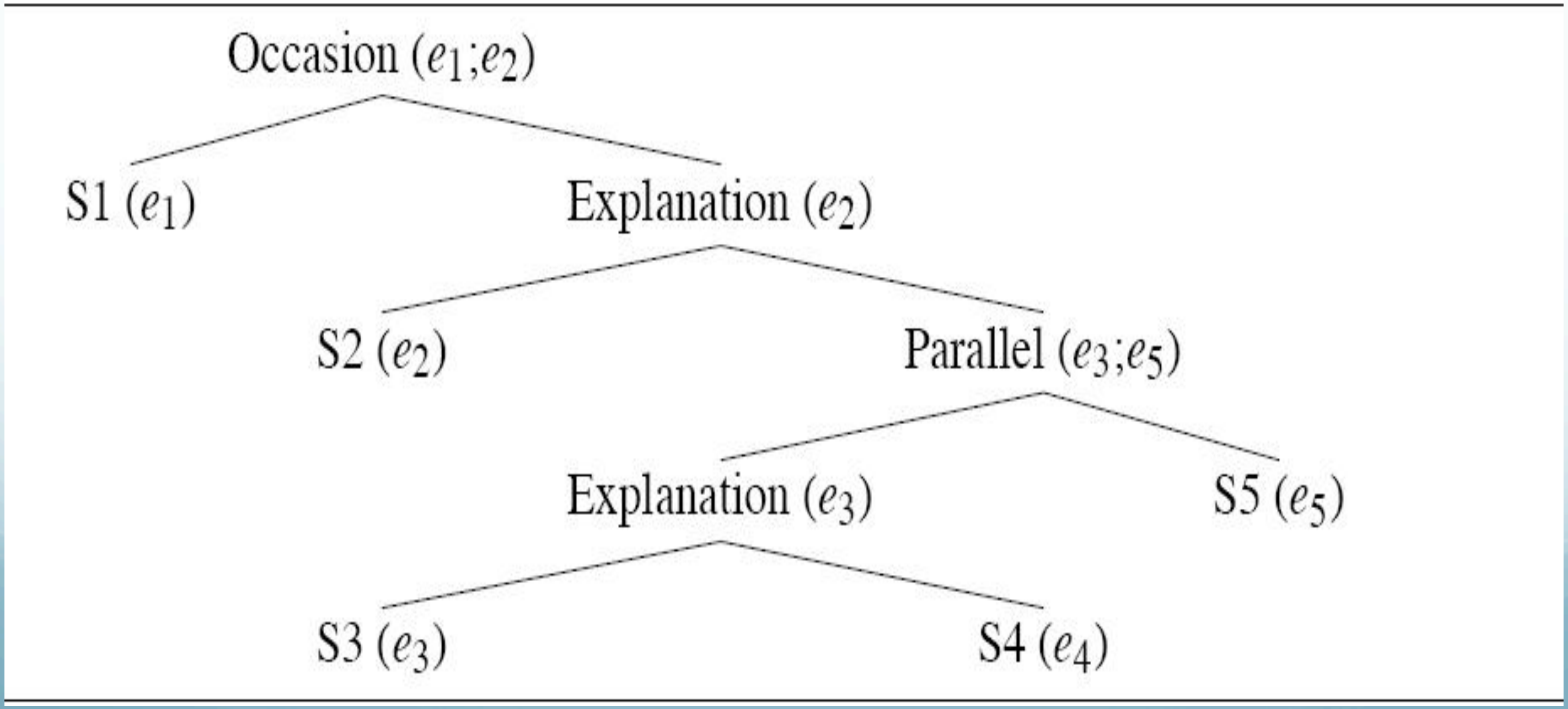
- Separate news broadcast into component stories

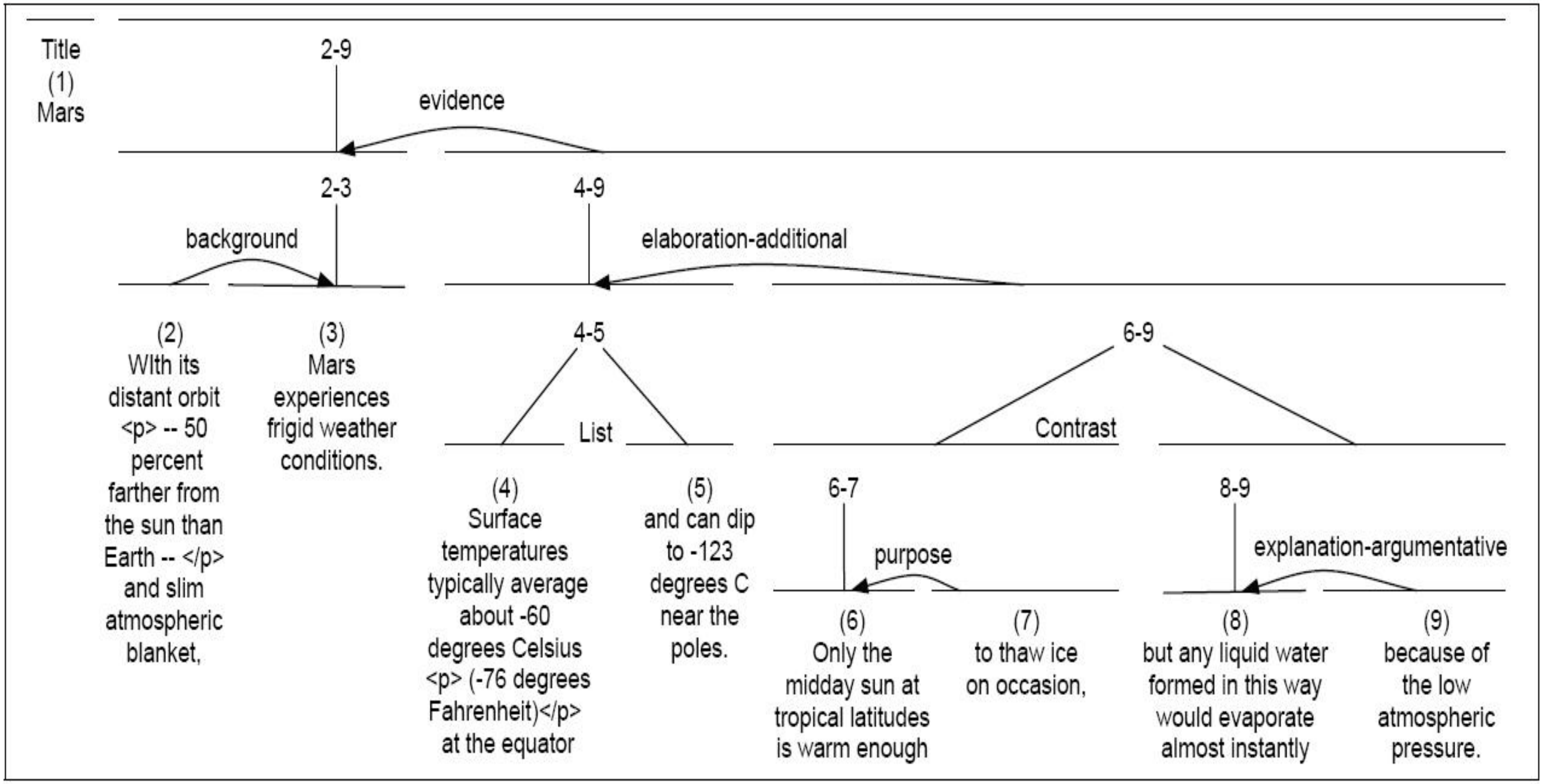


On "World News Tonight" this Thursday, another bad day on stock markets, all over the world global economic anxiety. ||
Another massacre in Kosovo, the U.S. and its allies prepare to do something about it. Very slowly. ||
And the millennium bug, Lubbock Texas prepares for catastrophe, Bangalore in India sees only profit.||

Coherence Analysis

S1: John went to the bank to deposit his paycheck.
S2: He then took a train to Bill's car dealership.
S3: He needed to buy a car.
S4: The company he works now isn't near any public transportation.
S5: He also wanted to talk to Bill about their softball league.




Title
(1)
Mars
2-9
evidence
2-3
4-9
background
elaboration-additional
(2)
WIth its distant orbit <p> -- 50 percent farther from the sun than Earth -- </p> and slim atmospheric blanket,
(3)
Mars experiences frigid weather conditions.
4-5
List
6-9
Contrast
(4)
Surface temperatures typically average about -60 degrees Celsius <p> (-76 degrees Fahrenheit)</p> at the equator
(5)
and can dip to -123 degrees C near the poles.
6-7
purpose
8-9
explanation-argumentative
(6)
Only the midday sun at tropical latitudes is warm enough
(7)
to thaw ice on occasion,
(8)
but any liquid water formed in this way would evaporate almost instantly
(9)
because of the low atmospheric pressure.

TextTiling

- Structure:
  - Linear segmentation
- Units:
  - ‘Sections’ from sentences, paragraphs
- Information:
  - Lexical cohesions, word-level
- Evaluation:
  - Accuracy, WindowDiff

TextTiling (Hearst '97)

- Lexical cohesion-based segmentation
  - Boundaries at dips in cohesion score
  - Tokenization, Lexical cohesion score, Boundary ID

TextTiling (Hearst '97)

- Lexical cohesion-based segmentation
  - Boundaries at dips in cohesion score
  - Tokenization, Lexical cohesion score, Boundary ID
- Tokenization
  - Units?

TextTiling (Hearst '97)

- Lexical cohesion-based segmentation
  - Boundaries at dips in cohesion score
  - Tokenization, Lexical cohesion score, Boundary ID
- Tokenization
  - Units?
    - White-space delimited words
    - Stopped
    - Stemmed
    - 20 words = 1 pseudo sentence

Lexical Cohesion Score

- Similarity between spans of text
  - b = ‘Block’ of 10 pseudo-sentences before gap
  - a = ‘Block’ of 10 pseudo-sentences after gap
  - How do we compute similarity?

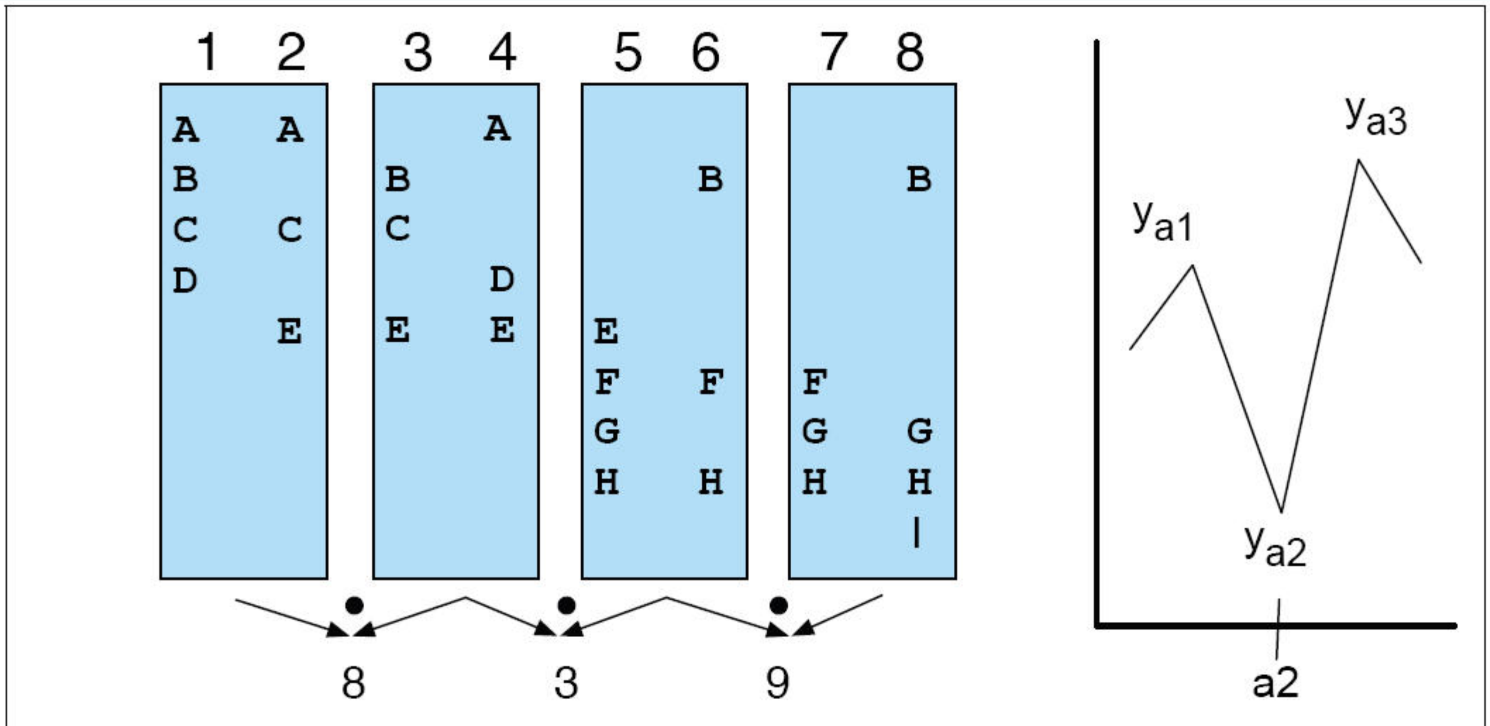
Lexical Cohesion Score

- Similarity between spans of text
  - b = ‘Block’ of 10 pseudo-sentences before gap
  - a = ‘Block’ of 10 pseudo-sentences after gap
  - How do we compute similarity?
    - Vectors and cosine similarity (again!)

$$sim_{cosine}(\vec{b},\vec{a}) = \frac{\vec{b} \bullet \vec{a}}{|\vec{b}||\vec{a}|} = \frac{\sum_{i=1}^{N} b_i \times a_i}{\sqrt{\sum_{i=1}^{N} b_i^2} \sqrt{\sum_{i=1}^{N} a_i^2}}$$

Segmentation

- Depth score:
  - Difference between position and adjacent peaks
  - E.g., $(y_{a1}-y_{a2})+(y_{a3}-y_{a2})$



(a) (b)

Evaluation

Evaluation

- Contrast with reader judgments
  - Alternatively with author or task-based
  - 7 readers, 13 articles: “Mark topic change”
    - If 3 agree, considered a boundary

Evaluation

- Contrast with reader judgments
  - Alternatively with author or task-based
  - 7 readers, 13 articles: "Mark topic change"
    - If 3 agree, considered a boundary
- Run algorithm – align with nearest paragraph
  - Contrast with random assignment at frequency

Evaluation

- Contrast with reader judgments
  - Alternatively with author or task-based
  - 7 readers, 13 articles: “Mark topic change”
    - If 3 agree, considered a boundary
- Run algorithm – align with nearest paragraph
  - Contrast with random assignment at frequency
- Auto: 0.66, 0.61; Human:0.81, 0.71
  - Random: 0.44, 0.42

Discussion

- Overall: Auto much better than random
  - Often “near miss” – within one paragraph
    - 0.83,0.78
- Issues

Discussion

- Overall: Auto much better than random
  - Often "near miss" – within one paragraph
    - 0.83,0.78
- Issues: Summary material
  - Often not similar to adjacent paras
- Similarity measures

Discussion

- Overall: Auto much better than random
  - Often "near miss" – within one paragraph
    - 0.83,0.78
- Issues: Summary material
  - Often not similar to adjacent paras
- Similarity measures
  - Is raw tf the best we can do?
  - Other cues??
- Other experiments with TextTiling perform less well – Why?

Improving TextTiling

- Basic technique: Weighted word overlap

Improving TextTiling

- Basic technique: Weighted word overlap
- Issue?
  - Restricted to exact word match
    - Synonyms? Hyper/hyponyms?
    - Related words?

Improving TextTiling

- Basic technique: Weighted word overlap
- Issue?
  - Restricted to exact word match
    - Synonyms? Hyper/hyponyms?
    - Related words?
- How can we generalize?
  - Automatically create pseudo-words that capture
  - Latent Semantic Analysis (and related techniques)

Latent Semantic Analysis

- Represents words by their document occurrence patt.
- Create term x document matrix

Latent Semantic Analysis

- Represents words by their document occurrence patt.
- Create term x document matrix
- Perform dimensionality reduction
  - E.g. Singular Value Decomposition on matrix
  - Create best rank *k* approximation of original matrix

Latent Semantic Analysis

- Represents words by their document occurrence patt.
- Create term x document matrix
- Perform dimensionality reduction
  - E.g. Singular Value Decomposition on matrix
  - Create best rank *k* approximation of original matrix
- Yields representation of terms/documents in new space of latent semantic concepts
  - Can improve retrieval, semantic distance calculation

Latent Semantic Analysis

- Represents words by their document occurrence patt.
- Create term x document matrix
- Perform dimensionality reduction
  - E.g. Singular Value Decomposition on matrix
  - Create best rank *k* approximation of original matrix
- Yields representation of terms/documents in new space of latent semantic concepts
  - Can improve retrieval, semantic distance calculation
- Many dimensionality reduction variants (e.g. GLSA)

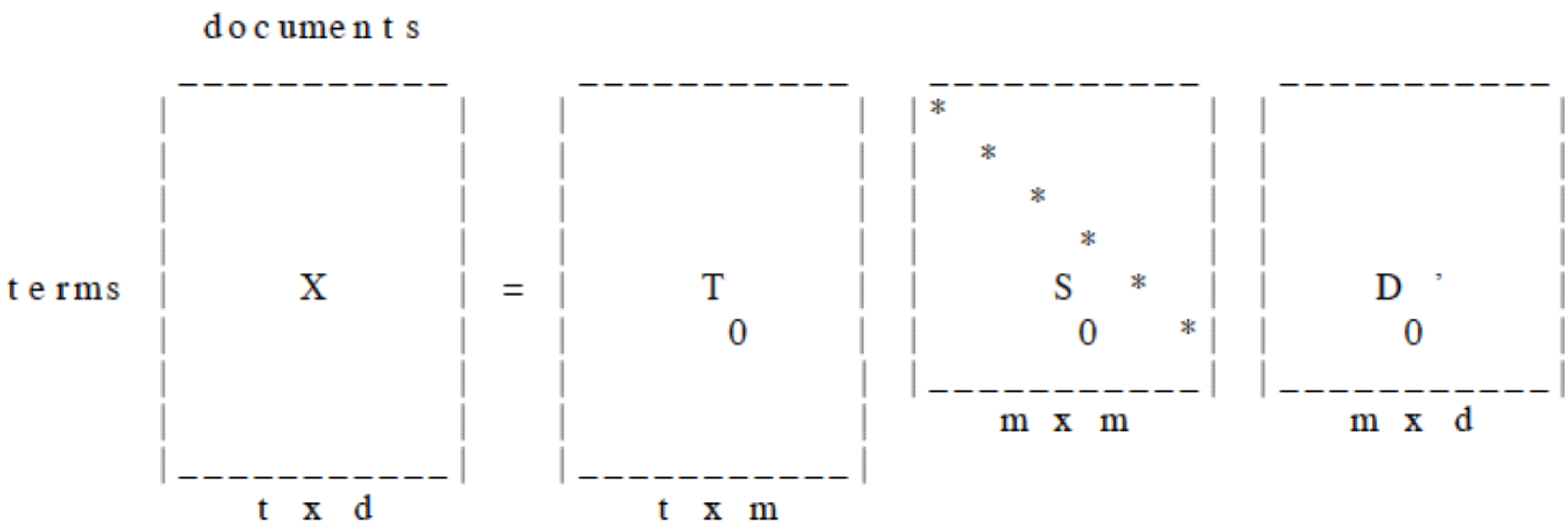
Latent Semantic Analysis

Sample Term by Document matrix

	access	*document*	*retrieval*	*information*	*theory*	*database*	*indexing*
Doc 1	x	x	x			x	x
Doc 2				x*	x		
Doc 3			x	x*			

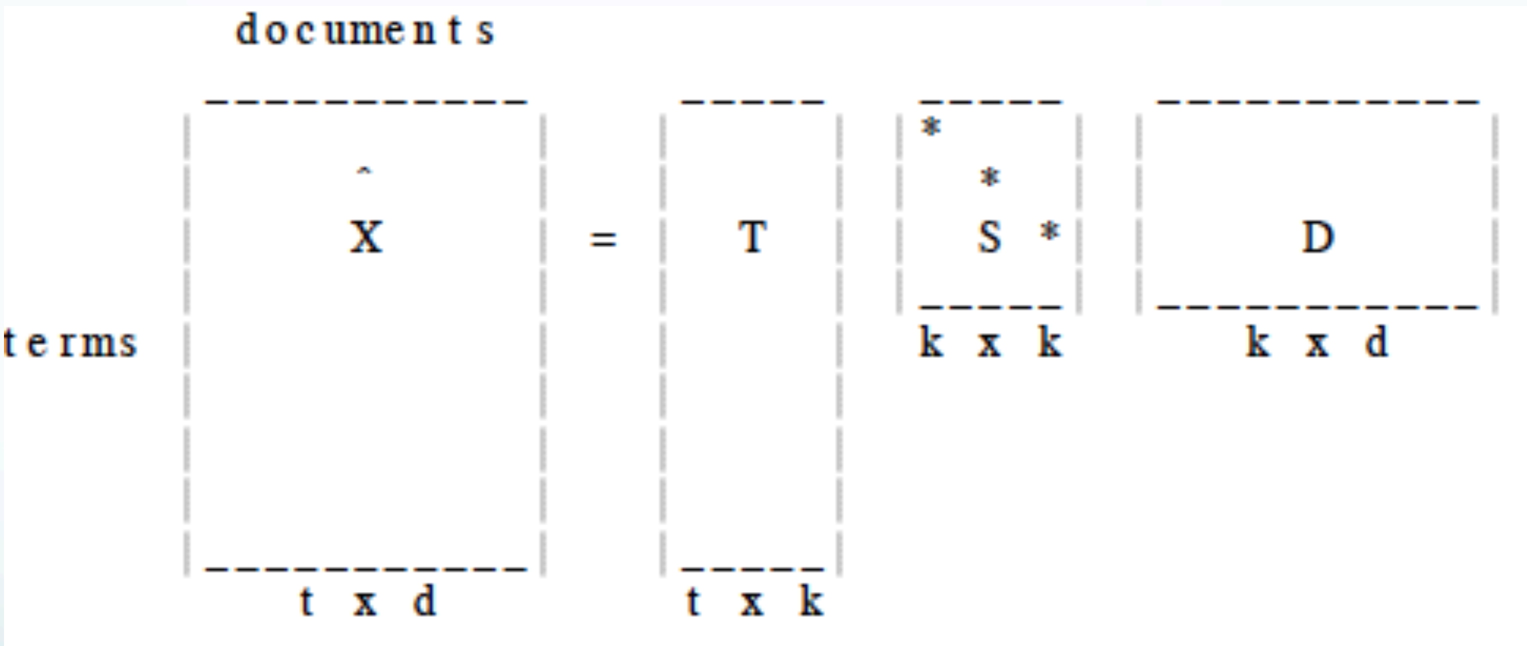
LSA

- SVD



LSA: Rank k Approximation

- Reduced rank:



LSA Text Segmentation

- Replace basic word based cosine similarity
  - Instead use similarity based on new representation
    - Reduced effect of lexical variation

LSA Text Segmentation

- Replace basic word based cosine similarity
  - Instead use similarity based on new representation
    - Reduced effect of lexical variation
- Compute average block similarity
- Perform peak-picking, assign boundaries at dips
- Improves significantly over raw TextTiling

LSA Text Segmentation

- Replace basic word based cosine similarity
  - Instead use similarity based on new representation
    - Reduced effect of lexical variation
- Compute average block similarity
- Perform peak-picking, assign boundaries at dips
- Improves significantly over raw TextTiling
- Choi2001, Bergsten 2006, Malioutov et al, 2007; Matveeva & Levow 2007; Eisenstein 2008

Supervised Segmentation

- Structure:
  - Linear segmentation
- Units:
  - Stories from sentences
- Information:
  - Lexical cohesion, word-level
  - Cues
- Evaluation:
  - WindowDiff

Supervised Segmentation

- Unsupervised techniques rely on block similarity
  - Wide windows of topical similarity

Supervised Segmentation

- Unsupervised techniques rely on block similarity
  - Wide windows of topical similarity
- Other sources of information?
  - Local indicators
    - Cue phrases:

Supervised Segmentation

- Unsupervised techniques rely on block similarity
  - Wide windows of topical similarity
- Other sources of information?
  - Local indicators
    - Cue phrases: Often domain specific
      - News: ‘Christiane Amanpour, reporting’
      - ‘This is CNN’, ‘Incorporated’
    - Trigram models
  - (Beeferman et al, 1999; Galley et al, 2003)

Supervised Segmentation

- Combines
  - Local context tri-gram language models

Supervised Segmentation

- Combines
  - Local context tri-gram language models
  - Trigger pairs:
    - Pairs of words where occurrence of 1st boosts that of 2nd

Supervised Segmentation

- Combines
  - Local context tri-gram language models
  - Trigger pairs:
    - Pairs of words where occurrence of 1st boosts that of 2nd
      - Appearance w/in 500 words greater than prior
      - E.g. Picket, Scab; Pulitzer, Prizewinning
    - Implemented in a log-linear model
  - Integrated as product of log-linear and trigram models

Segmentation Features

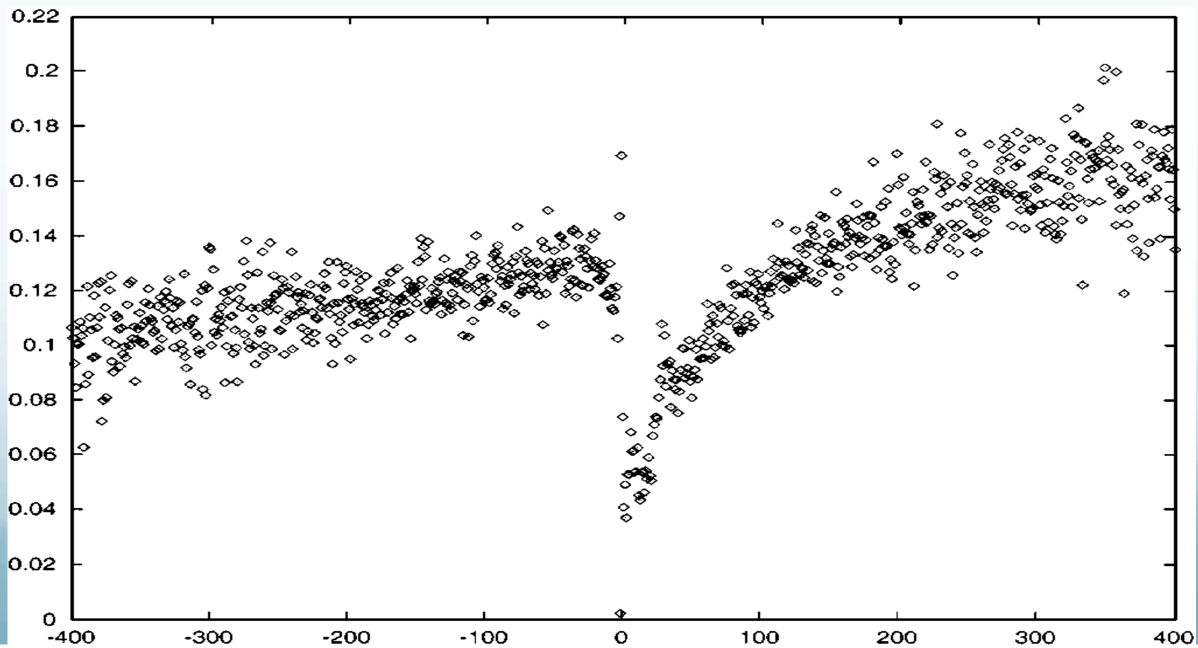
- Topicality features:
  - Idea:
    - Local tri-gram model more predictive than long-range

Segmentation Features

- Topicality features:
  - Idea:
    - Local tri-gram model more predictive than long-range
  - Implementation:
    - Compute log of ratio of trigger model to tri-gram model

Segmentation Features

- Topicality features:
  - Idea:
    - Local tri-gram model more predictive than long-range
  - Implementation:
    - Compute log of ratio of trigger model to tri-gram model



Segmentation Features II

- Word in some relative position:
  - I.e. Word X appears in next/previous Y words/sentences

Segmentation Features II

- Word in some relative position:
  - I.e. Word X appears in next/previous Y words/sentences
- Generates enormous numbers of possible features
  - 300K-500K
  - Filtered by feature selection process

Segmentation Features II

- Word in some relative position:
  - I.e. Word X appears in next/previous Y words/sentences
- Generates enormous numbers of possible features
  - 300K-500K
  - Filtered by feature selection process
- Why is selection/filtering necessary?

Segmentation Features II

- Word in some relative position:
  - I.e. Word X appears in next/previous Y words/sentences
- Generates enormous numbers of possible features
  - 300K-500K
  - Filtered by feature selection process
- Why is selection/filtering necessary?
  - Domain-specificity:
    - Broadcast news: ‘C.’ or ‘N.’ within some number of words
    - WSJ: ‘Incorporated’ or ‘Corporation’ within some window

Segmentation Evaluation

- Standard NLP evaluation measures:
  - Accuracy, F-measures
- Problems?

Segmentation Evaluation

- Standard NLP evaluation measures:
  - Accuracy, F-measures
- Problems?
  - Near-misses
    - Intuitively better if close to ‘true’ boundary
  - Fuzziness of boundaries

Segmentation Evaluation

- Standard NLP evaluation measures:
  - Accuracy, F-measures
- Problems?
  - Near-misses
    - Intuitively better if close to ‘true’ boundary
  - Fuzziness of boundaries
- Alternatives:
  - Give credit for near-miss

WindowDiff

- Pevzner & Hearst 2002
- Compares automatic to reference segmentation

WindowDiff

- Pevzner & Hearst 2002
- Compares automatic to reference segmentation
- Slides window of length k across segmentation
  - Compares # hypothesis boundaries to # of reference
    - Penalize window where h_i <> r_i; k= ½ average seg length

WindowDiff

- Pevzner & Hearst 2002
- Compares automatic to reference segmentation
- Slides window of length *k* across segmentation
  - Compares # hypothesis boundaries to # of reference
    - Penalize window where h_i <> r_i; *k*= ½ average seg length

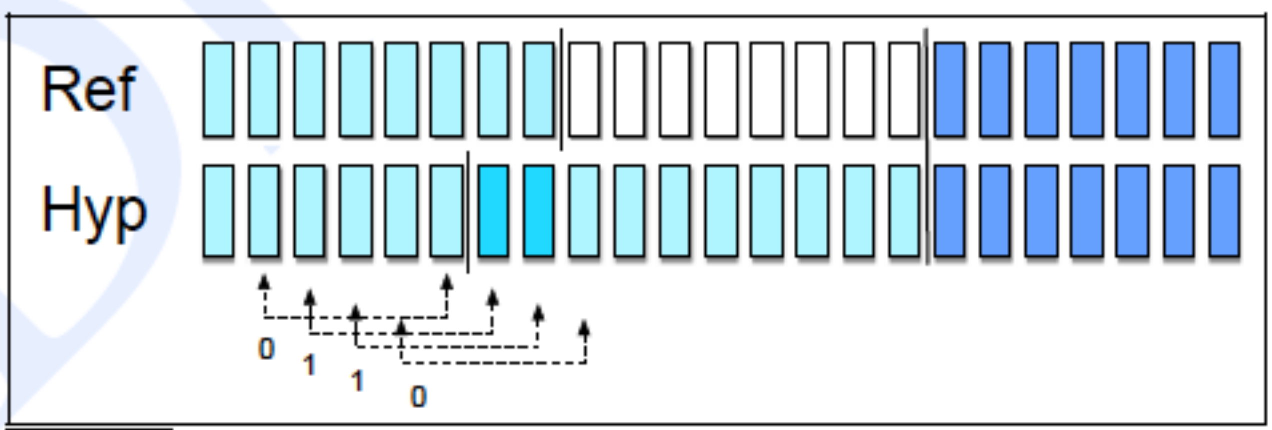


Figure 21.2 The WindowDiff algorithm, showing the moving window sliding over the hypothesis string, and the computation of $|r_i - h_i|$ at four positions. After Pevzner and Hearst (2002).

Other Systems

- Shriberg et al.
  - HMM's over topic models, language models, and prosodic cues
    - Contrasts in pitch, loudness; silence
- Galley et al, 2003:
  - LCSeg:
    - Lexical chains, cues from words, silence, overlap
    - Multi-party dialogue
- Multi-lingual:
  - English, Chinese, Dutch, Arabic,…

RST Parsing

- Structure:
  - Discourse structure (RST) tree, Relations
- Units:
  - Spans over clauses
- Information:
  - Lexical cohesion, word-level
- Evaluation:
  - Accuracy

RST Parsing (Marcu 1999)

- Learn and apply classifiers for
  - Segmentation and parsing of discourse

RST Parsing (Marcu 1999)

- Learn and apply classifiers for
  - Segmentation and parsing of discourse
- Assign coherence relations between spans

RST Parsing (Marcu 1999)

- Learn and apply classifiers for
  - Segmentation and parsing of discourse
- Assign coherence relations between spans
- Create a representation over whole text => parse
- Discourse structure
  - RST trees
    - Fine-grained, hierarchical structure
      - Clause-based units

Corpus-based Approach

- Training & testing on 90 RST trees
  - Texts from MUC, Brown (science), WSJ (news)
- Annotations:

Corpus-based Approach

- Training & testing on 90 RST trees
  - Texts from MUC, Brown (science), WSJ (news)
- Annotations:
  - Identify "edu"s – elementary discourse units
    - Clause-like units – key relation
    - Parentheticals – could delete with no effect

Corpus-based Approach

- Training & testing on 90 RST trees
  - Texts from MUC, Brown (science), WSJ (news)
- Annotations:
  - Identify "edu"s – elementary discourse units
    - Clause-like units – key relation
    - Parentheticals – could delete with no effect
  - Identify nucleus-satellite status

Corpus-based Approach

- Training & testing on 90 RST trees
  - Texts from MUC, Brown (science), WSJ (news)
- Annotations:
  - Identify “edu”s – elementary discourse units
    - Clause-like units – key relation
    - Parentheticals – could delete with no effect
  - Identify nucleus-satellite status
  - Identify relation that holds – I.e. elab, contrast…

Identifying Segments & Relations

- Key source of information:
  - Cue phrases
    - Aka discourse markers, cue words, clue words
  - Typically connectives
    - E.g. conjunctions, adverbs
  - Clue to relations, boundaries
  - Although, but, for example, however, yet, with, and....
    - John hid Bill's keys **because** he was drunk.

Cue Phrases

- Issues:
  - Ambiguous:
  - Insufficient:
    - Not all relations marked by cue phrases
    - Only 15-25% of relations marked by cues

Learning Discourse Parsing

- Train classifiers for:
  - EDU segmentation
  - Coherence relation assignment
  - Discourse structure assignment
    - Shift-reduce parser transitions

- Use range of features:
  - Cue phrases
  - Lexical/punctuation in context
  - Syntactic parses

RST Parsing Model

- Shift-reduce parser
  - Assumes segmented into edus
    - Edu -> Edt – minimal discourse tree unit, undefined

RST Parsing Model

- Shift-reduce parser
  - Assumes segmented into edus
    - Edu -> Edt – minimal discourse tree unit, undefined
  - Shift: Add next input edt to stack

RST Parsing Model

- Shift-reduce parser
  - Assumes segmented into edus
    - Edu -> Edt – minimal discourse tree unit, undefined
  - Shift: Add next input edt to stack
  - Reduce: pop top 2 edts, combine in new tree
    - Update: status, relation
    - Push new tree on stack

RST Parsing Model

- Shift-reduce parser
  - Assumes segmented into edus
    - Edu -> Edt – minimal discourse tree unit, undefined
  - Shift: Add next input edt to stack
  - Reduce: pop top 2 edts, combine in new tree
    - Update: status, relation
    - Push new tree on stack
- Requires 1 shift op and 6 reduce
  - Reduce ops: ns, sn, nn -> binary; below ops: non-binary

RST Parsing Model

- Shift-reduce parser
  - Assumes segmented into edus
    - Edu -> Edt – minimal discourse tree unit, undefined
  - Shift: Add next input edt to stack
  - Reduce: pop top 2 edts, combine in new tree
    - Update: status, relation
    - Push new tree on stack
- Requires 1 shift op and 6 reduce
  - Reduce ops: ns, sn, nn -> binary; below ops: non-binary
- Learn: relation+ops: 17*6 + 1

Learning Segmentation

- Per-word: Sentence, edu, or parenthetical bnd
- Features:

Learning Segmentation

- Per-word: Sentence, edu, or parenthetical bnd
- Features:
  - Context: 5 word window, POS: 2 before, after
    - Discourse marker present?
    - Abbreviations?

Learning Segmentation

- Per-word: Sentence, edu, or parenthetical bnd
- Features:
  - Context: 5 word window, POS: 2 before, after
    - Discourse marker present?
    - Abbreviations?
  - Global context:
    - Discourse markers, commas & dashes, verbs present
  - 2417 binary features/example

Segmentation Results

- Good news: Overall:~97% accuracy
  - Contrast: Majority (none): ~92%
  - Contrast: “DOT”= bnd: 93%
  - Comparable to alternative approaches
- Bad news:
  - Problem cases: Start of parenthetical, edu

Learning Shift-Reduce

- Construct sequence of actions from tree
- For a configuration:
  - 3 top trees on stack, next edt in input
  - Features: # of trees on stack, in input
    - Tree characteristics: size, branching, relations
    - Words and POS of 1st and last 2 lexemes in spans
    - Presence and position of any discourse markers
    - Previous 5 parser actions
    - Hearst-style semantic similarity across trees
    - Similarity of WordNet measures

Classifying Shift-Reduce

- C4.5 classifier
  - Overall: 61%
    - Majority: 50%, Random: 27%

Classifying Shift-Reduce

- C4.5 classifier
  - Overall: 61%
    - Majority: 50%, Random: 27%
- End-to-end evaluation:
  - Good on spans and N/S status
    - Poor on relations

Discussion

- Noise in segmentation degrades parsing
  - Poor segmentation -> poor parsing

Discussion

- Noise in segmentation degrades parsing
  - Poor segmentation -> poor parsing
- Need sufficient training data
  - Subset (27) texts insufficient
    - More variable data better than less but similar data

Discussion

- Noise in segmentation degrades parsing
  - Poor segmentation -> poor parsing
- Need sufficient training data
  - Subset (27) texts insufficient
    - More variable data better than less but similar data
- Constituency and N/S status good
  - Relation far below human

Evaluation

- Segmentation:
  - Good: 96%
    - Better than frequency or punctuation baseline

Evaluation

- Segmentation:
  - Good: 96%
    - Better than frequency or punctuation baseline
- Discourse structure:
  - Okay: 61% span, relation structure

Evaluation

- Segmentation:
  - Good: 96%
    - Better than frequency or punctuation baseline
- Discourse structure:
  - Okay: 61% span, relation structure
- Relation identification: poor

Issues

- Goal: Single tree-shaped analysis of all text
  - Difficult to achieve

Issues

- Goal: Single tree-shaped analysis of all text
  - Difficult to achieve
    - Significant ambiguity
    - Significant disagreement among labelers

Issues

- Goal: Single tree-shaped analysis of all text
  - Difficult to achieve
    - Significant ambiguity
    - Significant disagreement among labelers
    - Relation recognition is difficult
      - Some clear "signals", I.e. although
      - Not mandatory, only 25%

D-LTAG

Webber, Joshi, et al

D-LTAG

- Structure:
  - Tree structure, relations
- Units:
  - Clauses in local coherence relations
- Information:
  - Word pairs, word n-grams, polarity
- Evaluation:
  - F-measure on relations

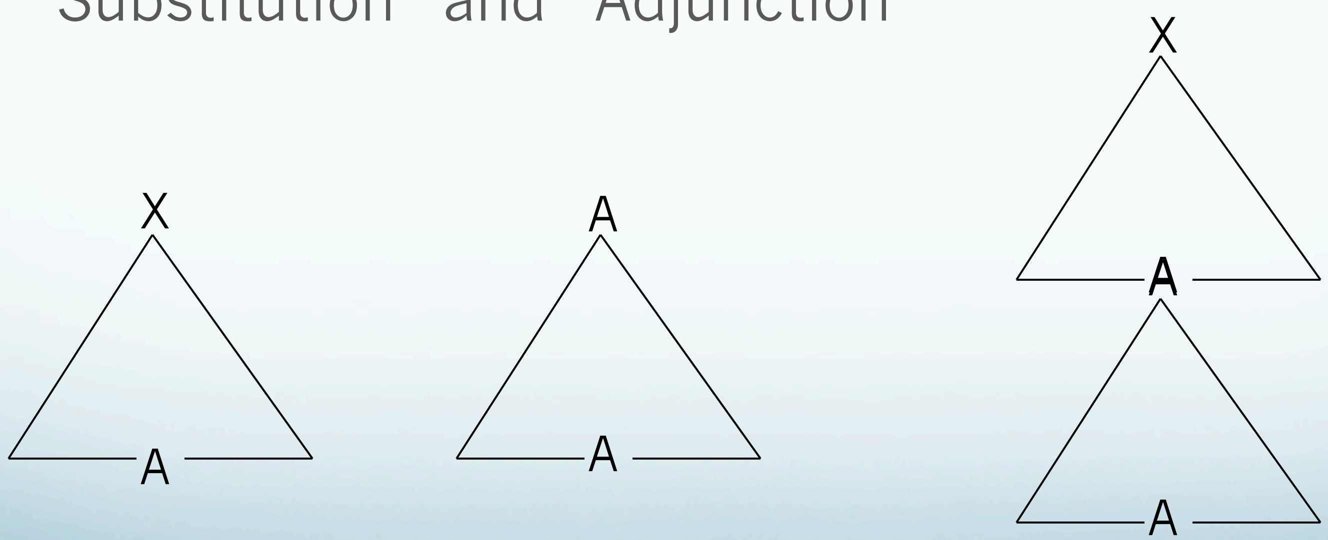
D-LTAG

- Discourse – handles discourse relations
- Lexicalized – builds on rules associated with words
- Tree Adjoining Grammar
  - Grammar model – mildly context-sensitive
    - Basic units are trees
    - Trees are combined by
      - Substitution
      - Adjunction

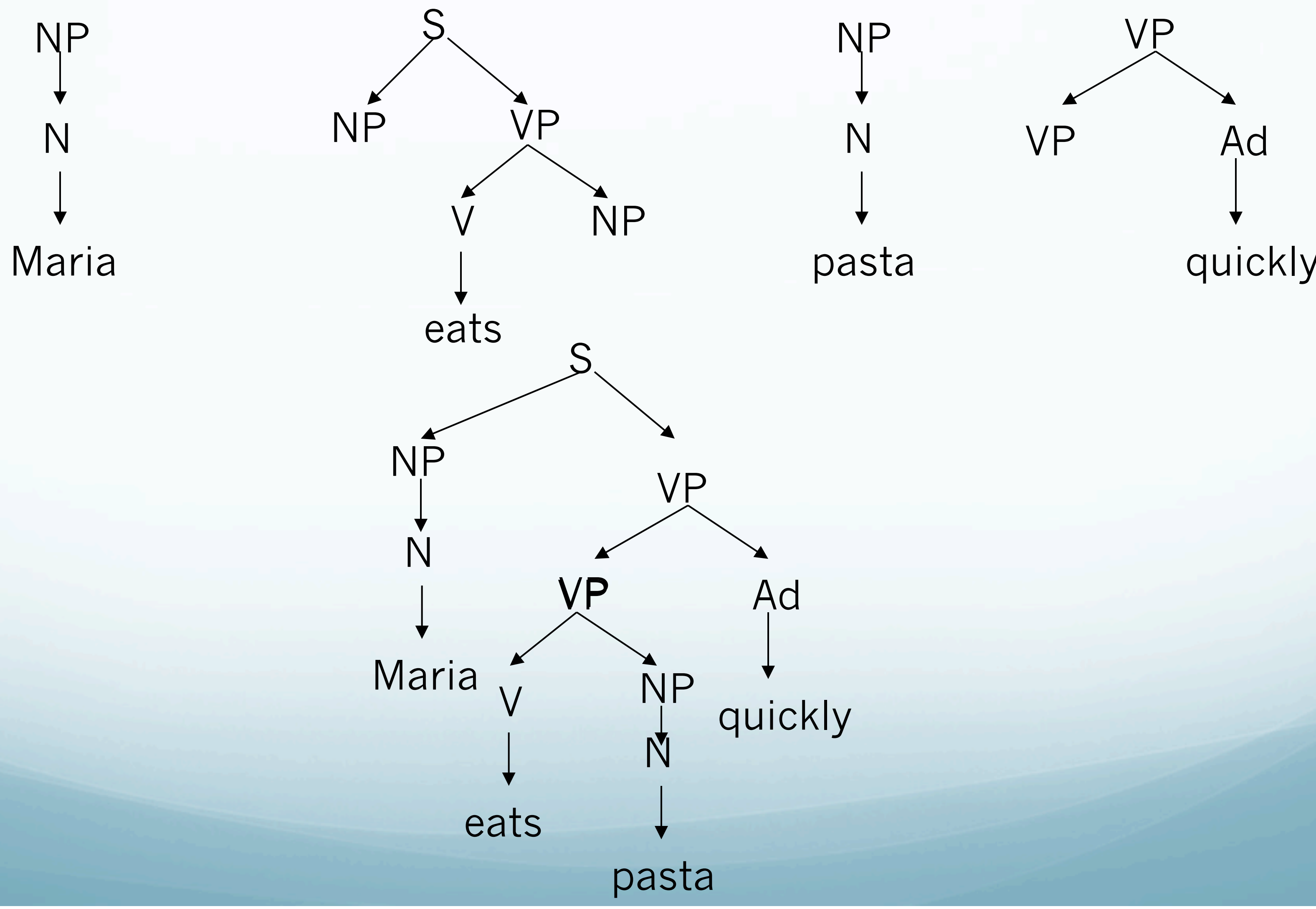
Tree Adjoining Grammars

- Mildly context-sensitive (Joshi, 1979)
  - Motivation:
    - Enables representation of crossing dependencies

- Operations for rewriting
  - "Substitution" and "Adjunction"



L-TAG Example



Dimensions of D-LTAG

- Discourse relations:
  - ‘Semantic classifications’ of lexical connectives (or implicit)
- Discourse structures:
  - Trees: predominantly binary (for discourse part)
- Discourse units:
  - Usually clauses, sequences
    - Exceptionally, VP coord, nominalization of discourse, anaphor, answer
- Discourse Segments:
  - Non-overlapping
- Discourse Relation Triggers:
  - Lexical elements and Structure

Lexicalization of Discourse

- D-LTAG associates discourse tree elements with words
- What words? Cue words (often)

Lexicalization of Discourse

- D-LTAG associates discourse tree elements with words
- What words? Cue words (often)
  - Specifically, discourse connectives:
    - Coordinating conjunctions
      - And, or, but

Lexicalization of Discourse

- D-LTAG associates discourse tree elements with words
- What words? Cue words (often)
  - Specifically, discourse connectives:
    - Coordinating conjunctions
      - And, or, but
    - Subordinating conjunctions
      - Although, because, however, etc, or null (implicit)

Lexicalization of Discourse

- D-LTAG associates discourse tree elements with words
- What words? Cue words (often)
  - Specifically, discourse connectives:
    - Coordinating conjunctions
      - And, or, but
    - Subordinating conjunctions
      - Although, because, however, etc, or null (implicit)
    - Parallel constructions
      - not only, but also

Lexicalization of Discourse

- D-LTAG associates discourse tree elements with words
- What words? Cue words (often)
  - Specifically, discourse connectives:
    - Coordinating conjunctions
      - And, or, but
    - Subordinating conjunctions
      - Although, because, however, etc, or null (implicit)
    - Parallel constructions
      - not only, but also
    - Discourse adverbials
      - Then, instead, ...

Connectives & Arguments

- Connectives viewed as predicate of 2 arguments
  - (from Webber 2006)
    - John loves Barolo.
    - So he ordered three cases of the ‘97.
    - But he had to cancel the order
    - Because he then discovered he was broke.

Connectives & Arguments

- Connectives viewed as predicate of 2 arguments
  - (from Webber 2006)
    - John loves Barolo.
    - So he ordered three cases of the ‘97.
    - But he had to cancel the order
    - Because he then discovered he was broke.
  - Conjunctions (So, But, Because)
    - Arg2 – current clause; arg1 – previous clause

Connectives & Arguments

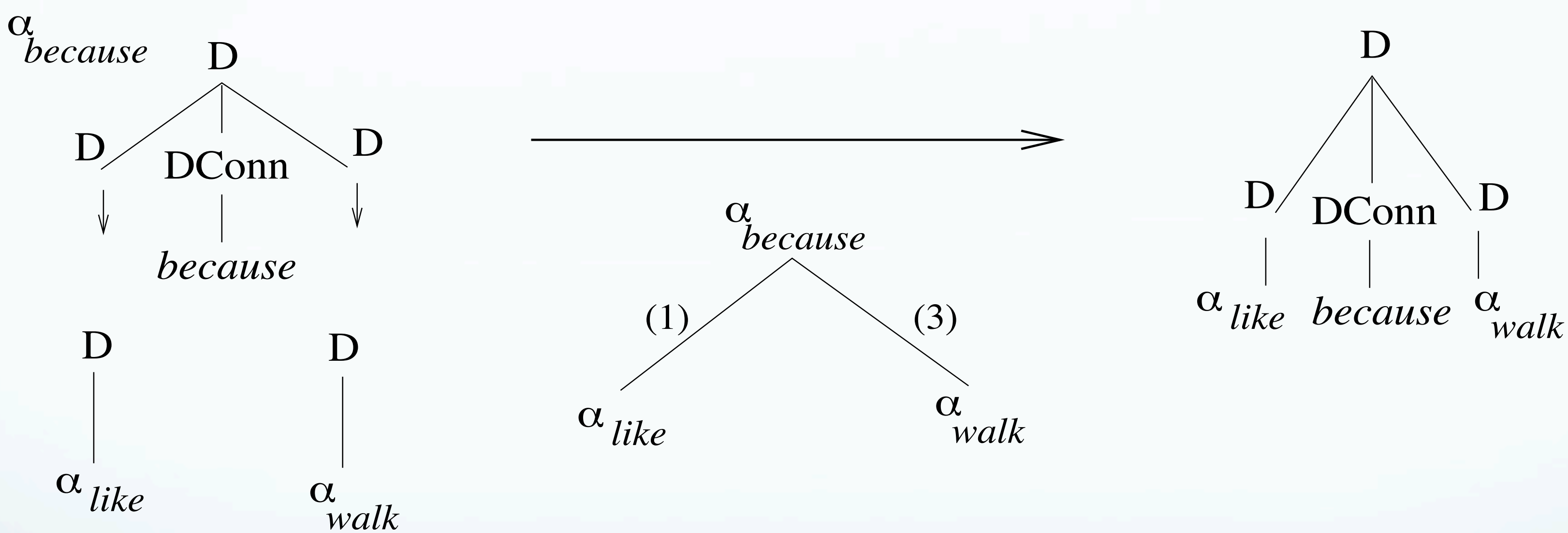
- Connectives viewed as predicate of 2 arguments
  - (from Webber 2006)
    - John loves Barolo.
    - So he ordered three cases of the ‘97.
    - But he had to cancel the order
    - Because he then discovered he was broke.
  - Conjunctions (So, But, Because)
    - Arg2 – current clause; arg1 – previous clause
  - Discourse adverbial ‘then’
    - Arg2 – current clause; arg1 -????

Connectives & Arguments

- Connectives viewed as predicate of 2 arguments
  - (from Webber 2006)
    - John loves Barolo.
    - So he ordered three cases of the '97.
    - But he had to cancel the order
    - Because he then discovered he was broke.
  - Conjunctions (So, But, Because)
    - Arg2 – current clause; arg1 – previous clause
  - Discourse adverbial 'then'
    - Arg2 – current clause; arg1 -????
      - Implicit anaphor – some prior clause or discourse element

Example: Structural Arguments to Conjunctions

- *John likes Mary* **because** **she walks Fido.**



Penn Discourse Treebank

- Explicit connectives:
  - In Washington, House aides said Mr. Phelan told congressmen that the collar, *which banned program trades through the Big Board's computer* **when** **the Dow Jones Industrial Average moved 50 points**, didn't work well.
  - A Chemical spokeswoman said *the second-quarter charge was "not material"* **and that no personnel changes were made** **as a result**.

Penn Discourse Treebank

- Explicit connectives:
  - In Washington, House aides said Mr. Phelan told congressmen that the collar, *which banned program trades through the Big Board's computer* **when** **the Dow Jones Industrial Average moved 50 points**, didn't work well.
  - A Chemical spokeswoman said *the second-quarter charge was "not material"* **and that no personnel changes were made** **as a result**.
- Implicit connectives:
  - *Some have raised their cash positions to record levels*. **Implicit=because (causal)** **High cash positions help buffer a fund when the market falls**.
  - *The projects already under construction will increase Las Vegas's supply of hotel rooms by 11,795, or nearly 20%, to 75,500*. **Implicit=so (consequence)** **By a rule of thumb of 1.5 new jobs for each new hotel room, Clark County will have nearly 18,000 new jobs**.

Annotation

- Applied as extension to Penn Treebank
  - Wall Street Journal
- Have trained D-LTAG parsers
- Available as Penn Discourse Treebank from LDC
  - PDTB 2.0

Genre	Total Inter-Sentential Discourse Rels	Total Explicit Inter-Sentential Connectives	Implicit Connectives
ESSAYS	3302	691 (20.9%)	2112 (64.0%)
SUMMARIES	916	95 (10.4%)	363 (39.6%)
LETTERS	433	85 (19.6%)	267 (61.7%)
NEWS	23017	4709 (20.5%)	13287 (57.7%)

Recognizing Implicit Relations

- Discourse parsing requires relation identification
  - Cue words/phrases very helpful
    - Relations disambiguatable at 93% by connective

Recognizing Implicit Relations

- Discourse parsing requires relation identification
  - Cue words/phrases very helpful
    - Relations disambiguatable at 93% by connective
  - But..
    - Only account for 25-30% of cases

Recognizing Implicit Relations

- Discourse parsing requires relation identification
  - Cue words/phrases very helpful
    - Relations disambiguatable at 93% by connective
  - But..
    - Only account for 25-30% of cases

- Relations are overwhelming ‘implicit’
  - However, identifiable by people
    - Annotated in PDTB

Implicit Relations

- The 101-year-old magazine has never had to woo advertisers with quite so much fervor before.
- It largely rested on its hard-to-fault demographics.

Implicit Relations

- The 101-year-old magazine has never had to woo advertisers with quite so much fervor before.
- **[because]** It largely rested on its hard-to-fault demographics.

Implicit Relations

- The 101-year-old magazine has never had to woo advertisers with quite so much fervor before.
- **[because]** It largely rested on its hard-to-fault demographics.

- Previous results had used synthetic implicits:
  - Delete existing connectives and classify
  - Accuracy not bad, but overestimates true implicits

PDTB Implicits

- Relations annotated between all adjacent sentences
- Hierarchy of relations:
  - Top-level: Comparison, Contigency, Expansion, Temporal
- Relation holds between 2 spans (args)
  - 1st sentence: Arg1; 2nd sentence: Arg2

Relation Features

- Narrowly focused funds grew wildly popular. They faded info oblivion after the crash.
- What relation?

Relation Features

  - Narrowly focused funds grew wildly popular. They faded info oblivion after the crash.
- What relation? **But/contrast**
- What words?

Relation Features

- Narrowly focused funds grew wildly popular. They faded info oblivion after the crash.

- What relation? But/contrast
- What words? **Popular/Oblivion**
  - Antonyms -> contrast

Relation Features

  - Narrowly focused funds grew wildly popular. They faded info oblivion after the crash.
- What relation? But/contrast
- What words? Popular/Oblivion
  - Antonyms -> contrast
- Word-pair features frequently used
- Problem:

Relation Features

    - Narrowly focused funds grew wildly popular. They faded info oblivion after the crash.
  - What relation? But/contrast
  - What words? Popular/Oblivion
    - Antonyms -> contrast
- Word-pair features frequently used
- Problem: too many pairs – lots of possible features
- Approach: Filtering
  - Stem; Use only most frequent stems – largely fn wds

Word-pair Analysis

- Sentence pairs from English Gigaword Corpus
  - Explicit relations with connectives removed

Word-pair Analysis

- Sentence pairs from English Gigaword Corpus
  - Explicit relations with connectives removed
- Contrast vs Other:
  - 5000 contrast, 2500 Causal, 2500 No-rel

Word-pair Analysis

- Sentence pairs from English Gigaword Corpus
  - Explicit relations with connectives removed
- Contrast vs Other:
  - 5000 contrast, 2500 Causal, 2500 No-rel
    - No-rel: sentences at least 3 sentences aparts
  - Extract all word-pairs
    - Remove those with < 5 occurrences
  - Rank by information gain in MALLET

Word-Pairs

- Highest information gain: (Pitler et al, 2009)
  - What do we see?

Comparison vs. Other			Contingency vs. Other		
the-but	s-but	the-in	the-and	in-the	the-of
of-but	for-but	but-but	said-said	to-of	the-a
in-but	was-but	it-but	a-and	a-the	of-the
to-but	that-but	the-it*	to-and	to-to	the-in
and-but	but-the	to-it*	and-and	the-the	in-in
a-but	he-but	said-in	to-the	of-and	a-of
said-but	they-but	of-in	in-and	in-of	s-and

Table 1: Word pairs with highest information gain.

Word-Pairs

- Highest information gain: (Pitler et al, 2009)
  - What do we see?
    - Lots of cue words!! Didn't they get deleted??
      - Only inter-sentence ones, not within sentence

Comparison vs. Other			Contingency vs. Other		
the-but	s-but	the-in	the-and	in-the	the-of
of-but	for-but	but-but	said-said	to-of	the-a
in-but	was-but	it-but	a-and	a-the	of-the
to-but	that-but	the-it*	to-and	to-to	the-in
and-but	but-the	to-it*	and-and	the-the	in-in
a-but	he-but	said-in	to-the	of-and	a-of
said-but	they-but	of-in	in-and	in-of	s-and

Table 1: Word pairs with highest information gain.

Word-Pairs

- Highest information gain: (Pitler et al, 2009)
  - What do we see?
    - Lots of cue words!! Didn't they get deleted??
      - Only inter-sentence ones, not within sentence
    - Function words: 'the-it'

Comparison vs. Other			Contingency vs. Other		
the-but	s-but	the-in	the-and	in-the	the-of
of-but	for-but	but-but	said-said	to-of	the-a
in-but	was-but	it-but	a-and	a-the	of-the
to-but	that-but	the-it*	to-and	to-to	the-in
and-but	but-the	to-it*	and-and	the-the	in-in
a-but	he-but	said-in	to-the	of-and	a-of
said-but	they-but	of-in	in-and	in-of	s-and

Table 1: Word pairs with highest information gain.

Word-Pairs

- Highest information gain: (Pitler et al, 2009)
  - What do we see?
    - Lots of cue words!! Didn't they get deleted??
      - Only inter-sentence ones, not within sentence
    - Function words: 'the-it'

Comparison vs. Other			Contingency vs. Other		
the-but	s-but	the-in	the-and	in-the	the-of
of-but	for-but	but-but	said-said	to-of	the-a
in-but	was-but	it-but	a-and	a-the	of-the
to-but	that-but	the-it*	to-and	to-to	the-in
and-but	but-the	to-it*	and-and	the-the	in-in
a-but	he-but	said-in	to-the	of-and	a-of
said-but	they-but	of-in	in-and	in-of	s-and

Table 1: Word pairs with highest information gain.

New Features

- Intuition: popular/oblivion
  - Contrast in lexical semantics – antonymy
  - Also contrast in polarity: positive vs negative
    - Might help Comparison

New Features

- Intuition: popular/oblivion
  - Contrast in lexical semantics – antonymy
  - Also contrast in polarity: positive vs negative
    - Might help Comparison
- Polarity tags:
  - Each sentiment word (and negation) gets MPQA tag

New Features

- Intuition: popular/oblivion
  - Contrast in lexical semantics – antonymy
  - Also contrast in polarity: positive vs negative
    - Might help Comparison
- Polarity tags:
  - Each sentiment word (and negation) gets MPQA tag
- Inquirer tags:
  - General Inquirer lexicon classes: (verbs only)
    - Polarity, Comparison, Rise/Fall, Pain/Pleasure
      - Categories have fewer sparseness problems than words

More Features

- Money/Percent/Num:
  - Counts of each; combination in each pair
    - Why?

More Features

- Money/Percent/Num:
  - Counts of each; combination in each pair
    - Why? Domain-dependent? WSJ!
- WSJ-LM: Rank of relation for uni/bigram in PDTB

More Features

- Money/Percent/Num:
  - Counts of each; combination in each pair
    - Why? Domain-dependent? WSJ!
- WSJ-LM: Rank of relation for uni/bigram in PDTB
- Expl-LM: Rank from EG corpus

More Features

- Money/Percent/Num:
  - Counts of each; combination in each pair
    - Why? Domain-dependent? WSJ!
- WSJ-LM: Rank of relation for uni/bigram in PDTB
- Expl-LM: Rank from EG corpus
- Verbs:
  - # of verbs in same Levin class
  - Average VP length
  - Main verb POS tag

And More Features

- First/Last Words: First/List 1-3 words in each Arg

And More Features

- First/Last Words: First/List 1-3 words in each Arg
- Modality: presence of modals; specific types

And More Features

- First/Last Words: First/List 1-3 words in each Arg
- Modality: presence of modals; specific types
- Context: previous/following explicit relation

And More Features

- First/Last Words: First/List 1-3 words in each Arg
- Modality: presence of modals; specific types
- Context: previous/following explicit relation
- Word-pairs:
  - Derived from EG corpus
  - Derived from Implicit spans in PDTB
  - Only PDTB implicit word pairs with information gain> 0
  - Derived from Explicit spans in PDTB

Experiments

- Classifiers:
  - MALLET: Naïve Bayes, MaxEnt, AdaBoost
  - Train balanced 1-vs-all classifiers for 4 classes
  - Test on natural distribution

Features	Comp. vs. Not	Cont. vs. Other	Exp. vs. Other	Temp. vs. Other	Four-way
Money/Percent/Num	19.04 (43.60)	18.78 (56.27)	22.01 (41.37)	10.40 (23.05)	(63.38)
Polarity Tags	16.63 (55.22)	19.82 (76.63)	71.29 (59.23)	11.12 (18.12)	(65.19)
WSJ-LM	18.04 (9.91)	0.00 (80.89)	0.00 (35.26)	10.22 (5.38)	(65.26)
Expl-LM	18.04 (9.91)	0.00 (80.89)	0.00 (35.26)	10.22 (5.38)	(65.26)
Verbs	18.55 (26.19)	36.59 (62.44)	59.36 (52.53)	12.61 (41.63)	(65.33)
First-Last, First3	21.01 (52.59)	36.75 (59.09)	63.22 (56.99)	15.93 (61.20)	(65.40)
Inquirer tags	17.37 (43.8)	15.76 (77.54)	70.21 (58.04)	11.56 (37.69)	(62.21)
Modality	17.70 (17.6)	21.83 (76.95)	15.38 (37.89)	11.17 (27.91)	(65.33)
Context	19.32 (56.66)	29.55 (67.42)	67.77 (57.85)	12.34 (55.22)	(64.01)
Random	9.91	19.11	64.74	5.38	

Discussion

- Overall:
  - First/Last words generally among best features
  - Context features also help

Discussion

- Overall:
  - First/Last words generally among best features
  - Context features also help
- Surprises:
  - Polarity features some of worst for contrast
    - Polarity pairs as common in non-contrast as contrast

Discussion

- Overall:
  - First/Last words generally among best features
  - Context features also help
- Surprises:
  - Polarity features some of worst for contrast
    - Polarity pairs as common in non-contrast as contrast
- Word-pairs:
  - Best features: From implicit pairs, w/info gain

Discussion

- Overall:
  - First/Last words generally among best features
  - Context features also help
- Surprises:
  - Polarity features some of worst for contrast
    - Polarity pairs as common in non-contrast as contrast
- Word-pairs:
  - Best features: From implicit pairs, w/info gain
- Combining features improves 6-16% absolute
- However, overall accuracy still not great