Discourse Parsing

Discourse issues

Analysis techniques

DISCOURSE RELATIONS

A example

Jones has lots of experience.
He has been on the board for 10 years.
And he 's refused bribes.
So he's honest.
He would really make a good president.

[Cohen 1987]

Diagnosis

- **Relations between facts/assertions not explicitly expressed**
- **Cue phrases (here: and, so) only contribute to a limited extent – ambiguous!**

Challenges

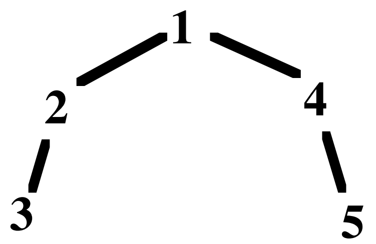
- **Reconstructing the intended argumentative structure (in analysis)**
- **Presenting arguments in a natural and understandable form (in generation)**

GENERATION – PRESENTING DISCOURSE RELATIONS

Some possible variations

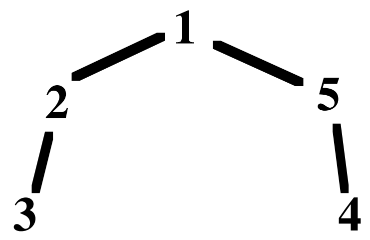
PRE-ORDER

1. Jones would make a good president.
2. He has lots of experience.
3. He has been on the board for 10 years..
4. And he's honest.
5. He's refused bribes.



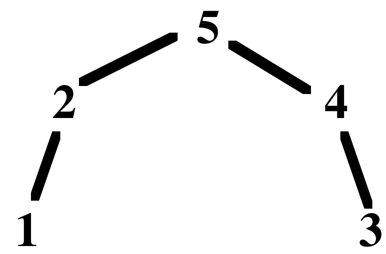
HYBRID

1. Jones would make a good president.
2. He has lots of experience.
3. He has been on the board for 10 years.
4. *And* he 's refused bribes.
5. *So* he's honest.



POST-ORDER

1. Jones has been on the board for 10 years.
2. He has lots of experience.
3. *And* he 's refused bribes.
4. *So* he's honest.
5. He would *really* make a good president.



Methods

- **Ordering and cue-phrase selection, embedded in sentence planning**
  (e.g., [Grote, Stede 1998])
- **Decisions guided by heuristics expressing aspects of linguistic/rhetorical adequacy**
  (e.g., [Scott, de Souza 1992])

INFERRING DISCOURSE RELATIONS FROM TEXT

Seminal method by Marcu [2000]

Shallow processing of unrestricted text

Based on empirical results obtained by a large number of researchers

Principled Procedure

1. Hypothesizing elementary units of text and rhetorical relations between them

The problem of rhetorical grounding

2. Propagating results by a well-constrained mathematical model

The problem of rhetorical structure derivation

RESOURCES FOR THE METHOD

Information exploited – observables in the text

Linguistics of *punctuation* – by itself 80% correctness

***Connectives* – approximately 1 marker for every 2 clauses sufficiently large**

Problems

Ambiguities between sentential and discourse function (e.g., *and*)

Connectives can signal more than one relation (e.g., *but:* CONTRAST, ANTITHESIS)

Connectives do not explicitly signal the size of the textual spans they relate

Evidence about the data (through corpus analyses) includes

***Marker* – the orthographic environment**

***Position* (in the textual unit) and *where to link it* (the textual unit related by it)**

***Rhetorical relation* (it expresses) and *status* (*Nucleaus* or *Satellite*)**

AN EXAMPLE

The underlying text

[John likes sweets.[1]][Most of all, John likes ice cream and chocolate.[2]]

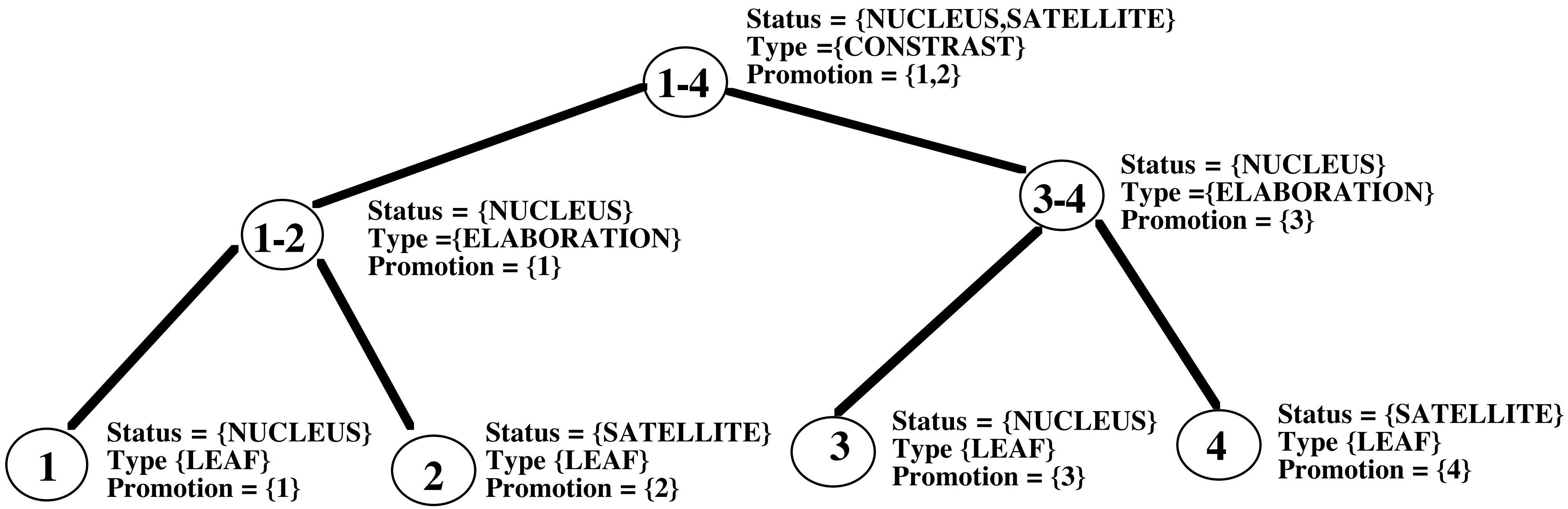
[*In contrast*, Mary likes fruit.[3]][*Especially* bananas and strawberries.[4]]

Relations hypothesized

1) **rhet_rel(CONSTRAST,1,3) ⊕ rhet_rel(CONSTRAST,1,4) ⊕ rhet_rel(CONSTRAST,2,3) ⊕ rhet_rel(CONSTRAST,2,4)**
2) **rhet_rel(ELABORATION,1,2)**
3) **rhet_rel(ELABORATION,4,1) ⊕ rhet_rel(ELABORATION,4,2) ⊕ rhet_rel(ELABORATION,4,3)**

1) **A CONSTRAST between some part preceding and some part following *in contrast***
2) **The second text span is about the same item (John) as the first one**
3) **The last text span is an ELABORATION (*especially*) of some part preceding it**

THE EXAMPLE YIELDS A SINGLE SOLUTION



Restrictions propagated

- **A CONTRAST must hold for text span 1, due to the promotion state**
- **The second ELABORATION must link 4 to 3, to avoid crossing CONTRAST**
- **Then the CONTRAST can only hold between 1 and 3, due to the promotion state**

ANALYSIS TECHNIQUES

Empirical investigations

2,100 text fragments manually annotated

(1,197 out of 2,100 cue phrase have a discourse function)

54 rhetorical relations annotated

(*Rhetorical Structure Theory* [Mann, Thompson 1987a] defines only 24)

Method

A proof-theoretic account of deriving rhetorical structures

12 Axioms (rewrite rules) describe coherent tree formation

Trees are assembled into larger trees in a bottom-up fashion

Preference metric used to disambiguate between multiple solutions

Best discourse trees are usually those that are skewed to the right

Motivated by results from psycholinguistics and text writing

RESULTS

Performance of the rhetorical parser

	Analysts		Program	
	Recall	**Precision**	**Recall**	**Precision**
Elementary spans	**87.9**	**87.9**	**51.2**	**95.9**
Spans	**89.6**	**89.6**	**63.5**	**87.7**
Nuclearity	**79.4**	**88.2**	**50.6**	**85.1**
Relations	**83.4**	**83.4**	**47.0**	**78.4**

Qualitative evaluations

- **Good discourse structures**

  at the paragraph level, for unambiguous discourse markers (especially not *and*)

- **Bad discourse structures**

  for incorrectly labeled intentional relations, for very large texts

SYNTAX-BASED TECHNIQUES (LeThanh et al. 2004)

Segmentation

Discourse segmentation rules according to phrasal categories

Rules selected which are in accordance with the syntactic structure

NP also treated as textual units when accompanied by a cue phrase

Discourse pasing

Syntactic information used to determine discourse relations and nuclearity roles

Example: reporting clause in nucleus, reported clause satellite of an elaboration

Sources of knowledge for the interpretation:

- **Syntactic information, NP-cues, VP-cues;**
- **cohesive devices (synonyms and hyponyms derived from WordNet)**

TEXT-LEVEL DISCOURSE ANALYSIS

Search space

Reduction through constraints about textual organization and adjacency

Marcu: recursively at each level of granularity

Composition driven by scores

- **Block-level score to connect text spans in the same textual unit**
- **Textual adjacency constraint**

Algorithm – beam search

Heuristic scores include cue scores, in dependency of the degree of certainty

Block level scores heavily penalize connections across block boundaries

Hypotheses stored, with block level scores dominating cue scores

Combination of best-first and shallow depth-first searching

EVALUATION

Corpus

10 short and 10 long documents (between 30 and 1284 words)

Texts and parses from Penn tree bank, 22 discourse relations (variant: 14 relations)

Composition driven by scores

Output accuracy	*System*	*Human*	*Difference*
1. Discourse segment	**86.9**	**98.7**	**11.8**
2. Combinations at sentence level	**66.3**	**88.3**	**22.0**
3. Nuclearity role at sentence level	**60.0**	**82.4**	**22.4**
4. Discourse relations (2 variants)	**52.2/53.0**	**69.0/74.5**	**16.8/21.5**
5. Text span combinations	**53.7**	**72.7**	**19.0**
6. Nuclearity on text level	**47.1**	**65.6**	**18.5**
7. Discourse relations on text level	**39.1/39.9**	**52.7/56.9**	**13.7/17.0**

RECENT TEXT-LEVEL DISCOURSE ANALYSIS

Data material

Rich linguistic features (contextual, constituent parse, dependency parse, lexical)

18 rhetorical relation classes, 78 finer-grained relations (RST Discourse Treebank)

4 classes, 16 types, 23 subtypes (Penn Discourse Treebank) - local context only

Techniques used

2 classifiers in cascade (1. relation y/n, and 2. if yes, which relation)

Examining the effectiveness of features (constituent parse features work best)

Recognition of implicit relations (no cue phrase)

Discourse production rules, semantic similarity

Major approaches

Lin et al. 2009, Hernault et al. 2010, Feng and Hirst 2012

GREEDY DISCOURSE ANALYSIS (Feng, Hirst 2014)

Motivation

Best discourse parsers (as to 2012) highly inefficient

Parsing of a longer paragraph may take several hours

Techniques used - 2 step procedure

Greedy bottom-up parsing (almost linear time complexity)

Post-editing phase to encounter for context information (e.g., depth of structures)

Use of intuitive contextual features

Development of context according to sequential flow of text captured better

Brief characterization

Better performance, post-editing doubles search time, but improves quality

PERFORMANCE COMPARISON

Processing times for paragraphs in the corpus

1. implementation of HILDA parser (a previous model, for comparison) (with new features)

2./3 new model without/with posteditig (PE)

Model	**Parsing Time (seconds)**		
	Avg	Min	Max
gSVM^{FH}	11.19	0.42	124.86
gCRF	5.52	0.05	40.57
gCRF^{PE}	10.71	0.12	84.72

EVALUATION COMPARISON

1./2. Best model so far (1.), another (reimplemented greedy model (2.)

3./4. The newmodel without (3.), with (4.) post-editing phase

Model	Span	Nuc	Relation	
			Acc	MAFS
*j*CRF	82.5	68.4	55.7	N/A
$g\mathrm{SVM}^{FH}$	82.8	67.1	52.0	27.4/23.3
*g*CRF	84.9*	69.9*	57.2*	35.3/31.3
$g\mathrm{CRF}^{PE}$	**85.7***†	**71.0***†	**58.2***†	**36.2/32.3**
Human	88.7	77.7	65.8	N/A

$*$: significantly better than $g\mathrm{SVM}^{FH}$ ($p < .01$)
†: significantly better than *g*CRF ($p < .01$)

TEXT-LEVEL DISCOURSE DEPENDENCY PARSING (Li, Wang, Chao, Li 2014)

Motivation

Design of production rules difficult (unless with syntactic parsing)

Different levels of discourse units require different features (no uniform approach)

Reduction of complexity through functionality rather than constituency

Techniques used

Prerequisite – corpus with annotations of relations (converted into dependencies)

Parsing means finding the best-scoring dependency tree

(maximum spanning tree - MST)

Based on Eisner's dependency parsing algorithm, complexity O(n^3)

(parses left and right dependents of discourse units independently)

REPRESENTATION OF LINGUISTIC KNOWLEDGE

Features in two elementary discourse units connected by a relation (same as most others)

1. **WORD: first and last word, first and last bigram**
2. **POS: first one and two POS tags**
3. **Position: whether both of units are in same sentence, position in embedding nodes**
4. **Length: of the units**
5. **Syntactic: POS tags of the dominating nodes**
6. **Semantic similarity: between the units, according to Wordnet**

Categories of discourse relations

19 course-grained relations

111 fine-grained relations

PERFORMANCE USING COARSE-GRAINED RELATIONS

Method	Features	Unlabeled Acc.	Labeled Acc.
Eisner	1+2	0.3602	0.2651
	1+2+3	0.7310	0.4855
	1+2+3+4	0.7370	0.4868
	1+2+3+4+5	0.7447	0.4957
	1+2+3+4+5+6	0.7455	0.4983
MST	1+2	0.1957	0.1479
	1+2+3	0.7246	0.4783
	1+2+3+4	0.7280	0.4795
	1+2+3+4+5	0.7340	0.4915
	1+2+3+4+5+6	0.7331	0.4851

PERFORMANCE USING FINE-GRAINED RELATIONS

Method	Feature types	Unlabeled Acc.	Labeled Acc.
Eisner	1+2	0.3743	0.2421
	1+2+3	0.7451	0.4079
	1+2+3+4	0.7472	0.4041
	1+2+3+4+5	0.7506	0.4254
	1+2+3+4+5+6	0.7485	0.4288
MST	1+2	0.2080	0.1300
	1+2+3	0.7366	0.4054
	1+2+3+4	0.7468	0.4071
	1+2+3+4+5	0.7494	0.4288
	1+2+3+4+5+6	0.7460	0.4309

EVALUATION

S **blank tree structure**

N **nuclearity indication**

R **tree structure with relation indication (no nuclearity)**

	S	N	R
Our-coarse	82.9	73.0	60.6
Our-fine	83.4	73.8	57.8
Percep-coarse	82.3	72.6	59.4
HILDA-manual	83.0	68.4	55.3
HILDA-seg	72.3	59.1	47.8
LeThanh	53.7	47.1	39.9
Marcu	44.8	30.9	18.8
Human	88.1	77.5	66.0

A RECENT COMPARISON (9 systems)

Micro-averaged F1 scores on labelled attachment decisions (original Parseval)

parser	S	N	R	F
HHN16 HILDA	**65.1**	**54.6**	**44.7**	**44.1**
SHV15 D*	**65.3**	**54.2**	**45.1**	**44.2**
JCN15 1S-1S	**65.1**	**55.5**	**45.1**	**44.3**
FH14 gCRF*	***68.6***	***55.9***	**45.8**	**44.6**
BPS16	**59.5**	**47.2**	**34.7**	**34.3**
LLC16	**64.5**	**54.0**	**38.1**	**36.6**
BCS17 mono	**61.9**	**53.4**	**44.5**	**44.0**
BCS17 cross+dev	**62.7**	**54.5**	**45.5**	**45.1**
JE14 DPLP**	**64.1**	**54.2**	***46.8***	***46.3***
human	**78.7**	**66.8**	**57.1**	**55.0**