Intelligent Systems (AI-2)

Computer Science cpsc422, Lecture 29

Nov, 17, 2017

422 big picture: Where are we?

Hybrid: Det +Sto

Prob CFG
Prob Relational Models
Markov Logics

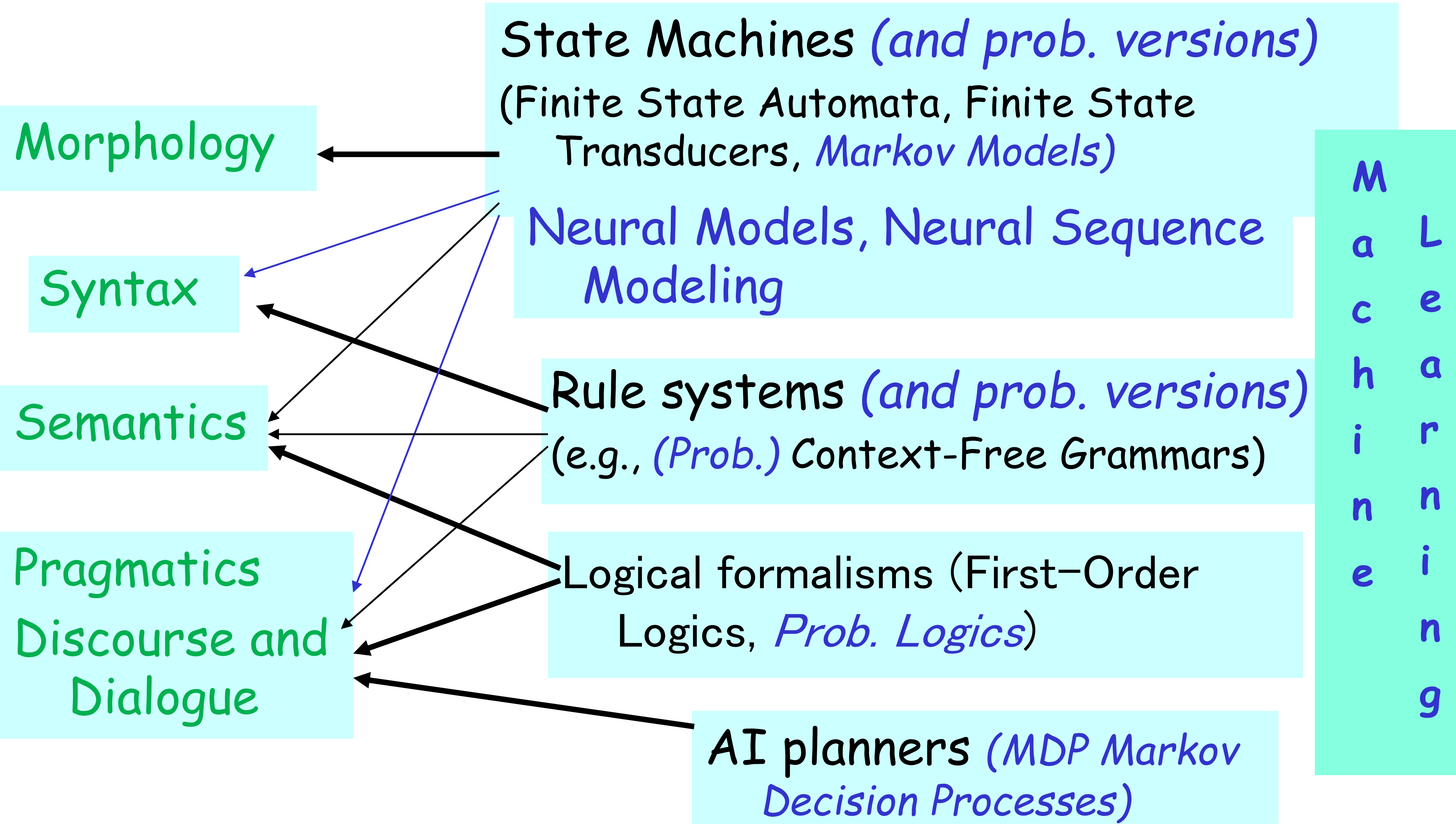
	Deterministic	Stochastic
Query	*Logics* *First Order Logics* *Ontologies* • Full Resolution • SAT	*Belief Nets* Approx. : Gibbs *Markov Chains and HMMs* Forward, Viterbi…. Approx. : Particle Filtering *Undirected Graphical Models* *Markov Networks* *Conditional Random Fields*
Planning		*Markov Decision Processes and Partially Observable MDP* • Value Iteration • Approx. Inference *Reinforcement Learning*

Applications of AI

Representation

Reasoning Technique

NLP: Knowledge–Formalisms Map
(including probabilistic formalisms)



Discovering Discourse Structure: Computational Tasks

The bank was hamstrung in its efforts to face the challenges of a changing market by its links to the government, analysts say.

Discourse Segmentation

The bank was hamstrung in its efforts

to face the challenges of a changing market by its links to the government,

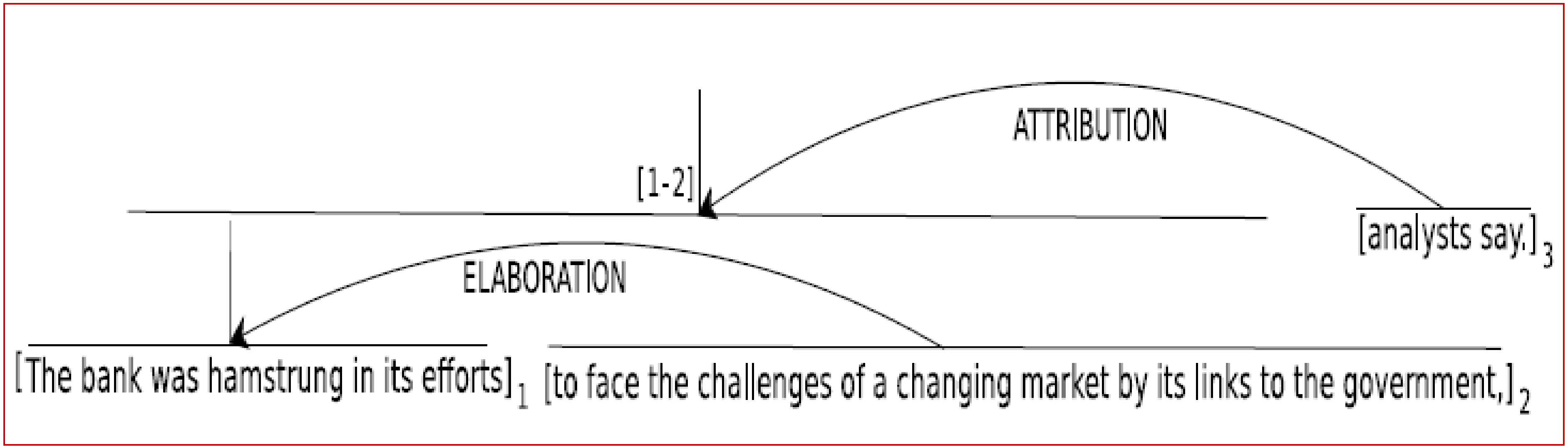
analysts say.

1

2

3

Discourse Parsing



Some general points

- Intelligent Systems are complex; often integrating many R&R systems + Machine Learning
- Conditional random fields, Syntactic parsing / Wordnet / Lexical Chains, Logistic regression….
- Discourse Parsing is a task we are very good at, here at UBC ;-)
  - 2 grad students working on/with it right now
  - (several in the past)
- Demo

Applications

- Detect Controversiality in online asynchronous conversations - 2014
- Summarize evaluative text (e.g., customer reviews) (journal paper 2016)
- "Using Discourse Structure Improves Machine Translation Evaluation“ ACL 2014
- Others ACL 2017 improvements in text categorization UofW

Some recent extensions

- **Coling 2016 – Semi-supervised data enrichment**
- **SigDial 2017 – joint neural model with Sentiment**

Current Work My group

- Improve on Coling paper – using a framework called data programming (Smart ensembling based on graphical models)
- Applied discourse features in detecting dementia from user generated text (did nto work ☹)

Parser	S	N	R	F
JE14 gCRF	84.3	69.4	56.9	56.2
FH14	82.0	68.2	57.8	57.6
JCN15 1S-1S	82.6	68.3	55.8	55.4
LLC16	82.2	66.5	51.4	50.6
BCS17 mono	81.0	67.7	55.7	55.3
BCS17 cross+dev	81.3	68.1	56.3	56.0
WLW17	86.0	72.4	59.7	58.9
Ours, before DP	85.9	72.0	56.7	56.0
Ours, after DP	85.9	72.0	57.7	56.8
Ours, before DP	85.9	72.0	58.1	57.2
Ours, after DP	85.9	72.0	58.4	57.6
Human	88.3	77.3	65.4	64.7

Table 1: Micro-Averaged F_1 score

et al., 2017) in four categories: Span (S) refer to discourse structures without label, Nuclearity (N) refer to discourse structures with nucleus/satellite label, Relation (R) refer to discourse structures with 18 coarse-grained relation label, and both (F) refer to discourse structures with both the relation label of nucleus label.

lowing their naming convention. ”**JCN15 1S-1S** is a two stage (sentence then intra-sentential level) CKY chart parser with Dynamic Conditional Random Field (DCRF) models (Joty et al., 2015). **FH14gCRF** is a two stage (sentence then intra-sentential level) bottom-up, greedy parser with linear-chain CRF models (Feng and Hirst, 2014). **LLC16** is a CKY chart parser with a hierarchical neural network model (attention-based hierarchical bi-LSTM) (Li et al., 2016). **BCS17 mono** is a transition-based parser that uses a feed-forward neural network model while **BCS17 cross+dev** is an variant of it with cross-lingual RST-DT data (Braud et al., 2017). **JE14 concat** is a shift-reduce parser that uses an SVM model (Ji and Eisenstein, 2014).” We also add a more recent discourse parser, **WLW17**, a four stage (structure and nucleus, then sentence, then intra-sentiental, then intra-paragraph) shift-reduce parser based on SVM (Wang et al., 2017).

State-of the art 2017

Graph Structure

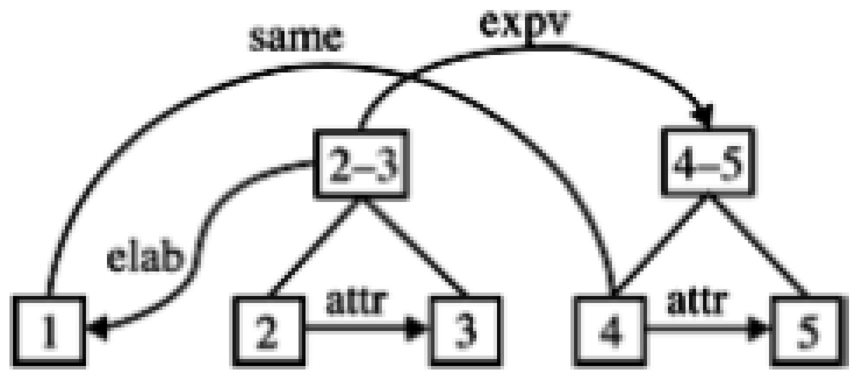


Figure 4.12: Graph representation of Example 4.33 (from Wolf and Gibson [2005, p. 267]).

- **Example 4.33** (1) Mr. Baker's assistant for inter-American affairs, Bernard Aronson, (2) while maintaining (3) that the Sandinistas had also broken the cease-fire, (4) acknowledged: (5) "It's never very clear who starts what."

Some Questions

Some Questions

- " Our discourse parser assumes that the input text has been already segmented into elementary discourse units. " Why do we need this kind of assumption?
- What is is the disadvantage when using DCRFs for sequence modelling compared to Hidden Markov Models and MRFs?
- method works for blogs or emails
- Could this be easily modified to detect the unnecessary words/sentences of a body of text?
- Discourse structure can also play important roles in sentiment analysis". Is there any work in progress in your lab that is related to this?

Some Questions

- Features, n-grams,
- **also seperately parse for DT of distinct paragraphs before building the final K probable discourse trees for the document?**
- Graph structure of discourse
- **How can the document parser account for different writing styles? ie**
- **aren't issues like "leaky boundaries" much more likely to arise in less formal writing, like a short story, than in a how-to-do manual?**
-

Our Discourse Parser

Discourse Parsing State of the art limitations:

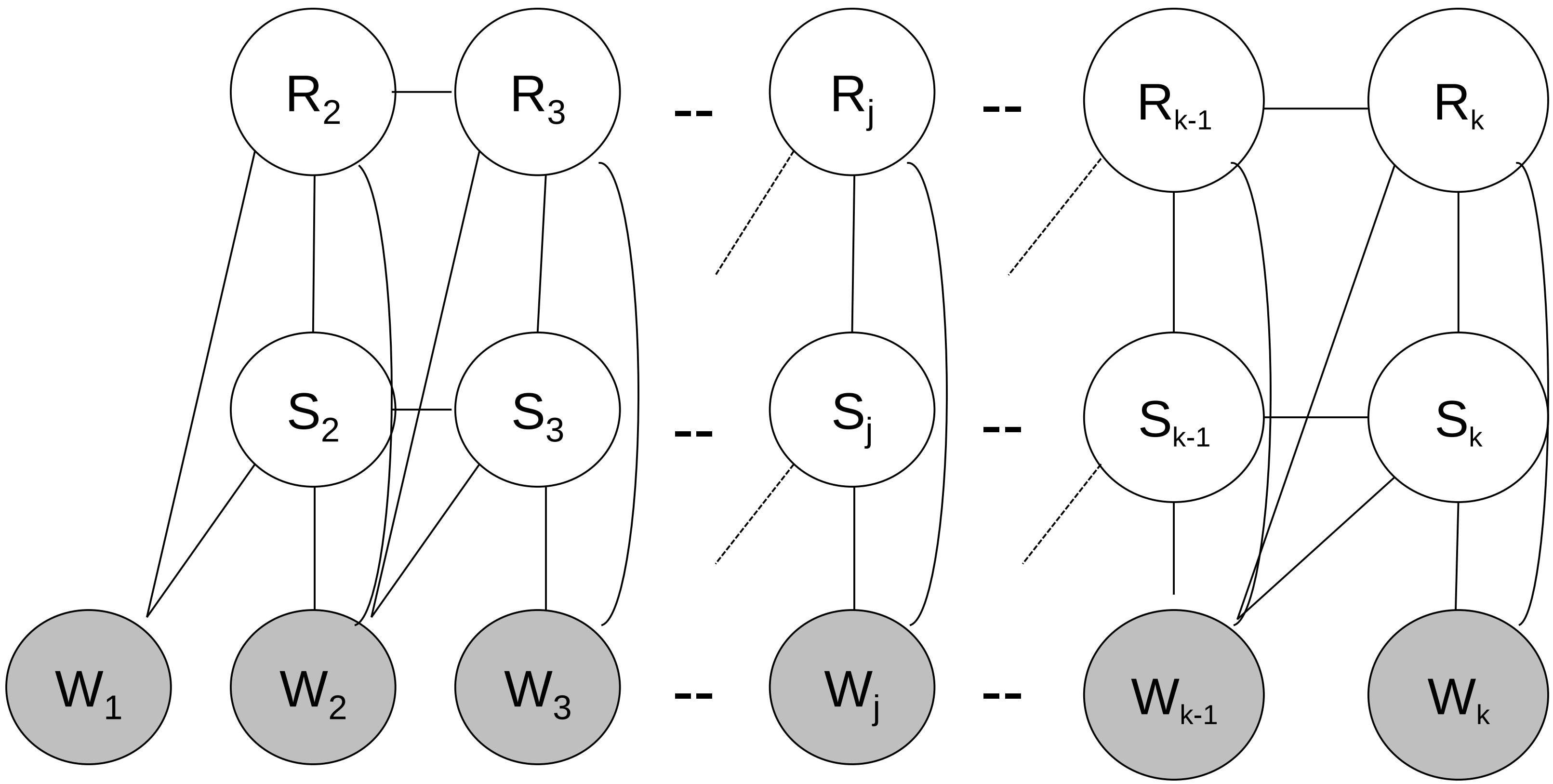
- Structure and labels determined separately
- Do not consider sequential dependency
- Suboptimal algorithm to build structure

Our Discourse Parser addresses these limitations

- Layered CRFs + CKY-like parsing

Our Parsing Model

Model **structure** and **label** jointly



Relation at level i
R ϵ {1 .. M}

Structure at level i
S ϵ {0, 1}

Spans at level i

Dynamic Conditional Random Field (DCRF) [Sutton et al, 2007]

Models sequential dependencies

Discourse Parsing: Evaluation

Corpora/Datasets

RST-DT corpus (Carlson & Marcu, 2001)	Instructional corpus (Subba & Di-Eugenio, 2009)
• 385 news articles -Train: 347 (7673 sentences) -Test: 38 (991 sentences)	•176 how-to-do manuals 3430 sentences

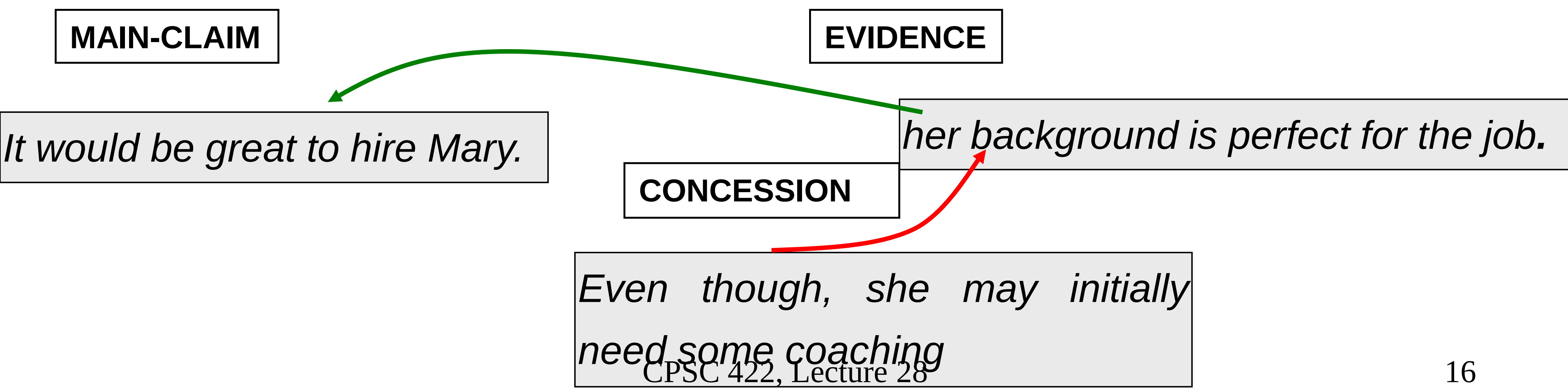
Excellent Results (beat state-of-the-art by a wide margin): [EMNLP-2012, ACL-2013**]**

Discourse Parsing: Example

NSERC Business Intelligence Network

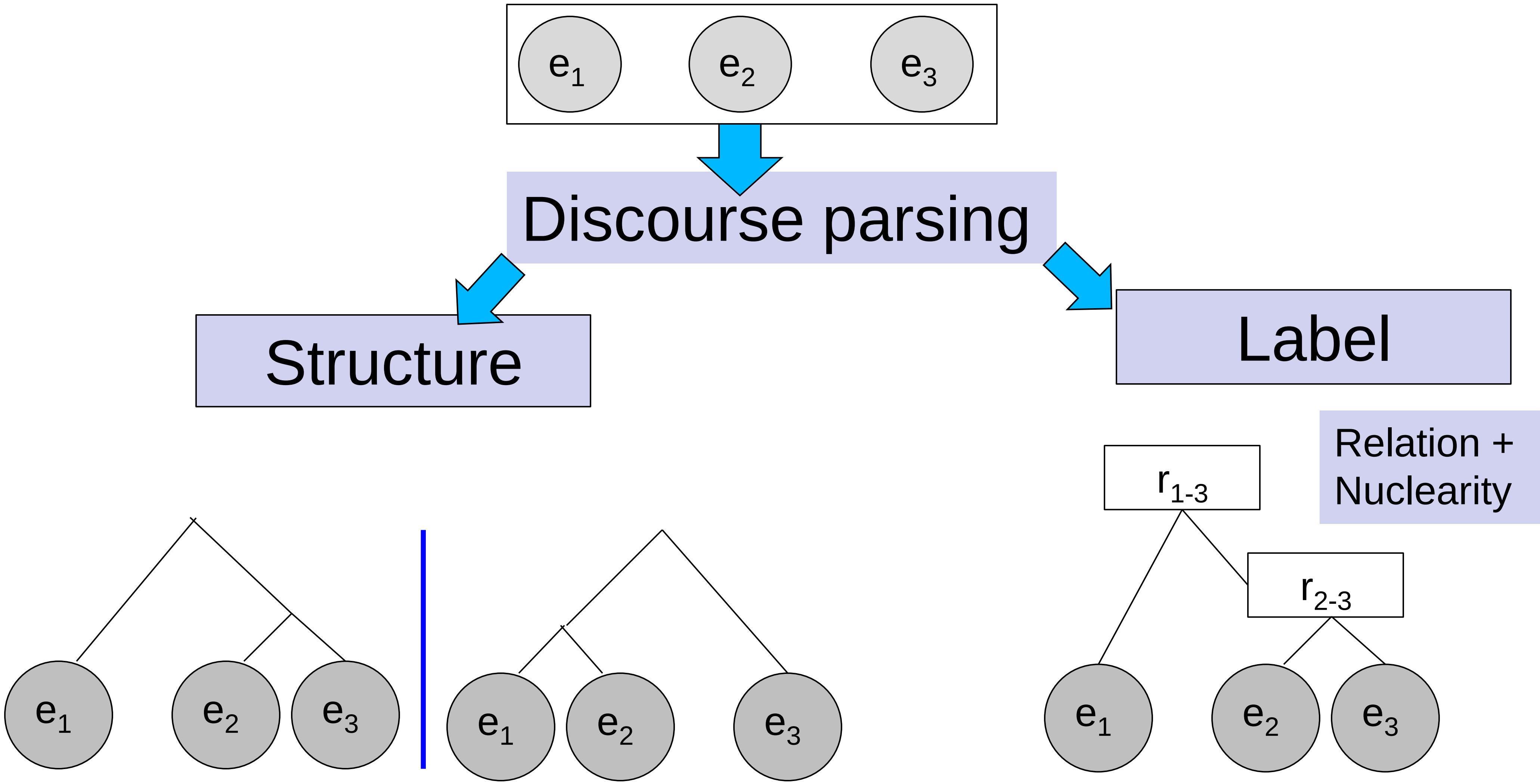
- What is the main claim in a message and how it is expanded/supported by the other claims

"It would be great to hire Mary. Even though she may initially need some coaching , her background is perfect for the job."



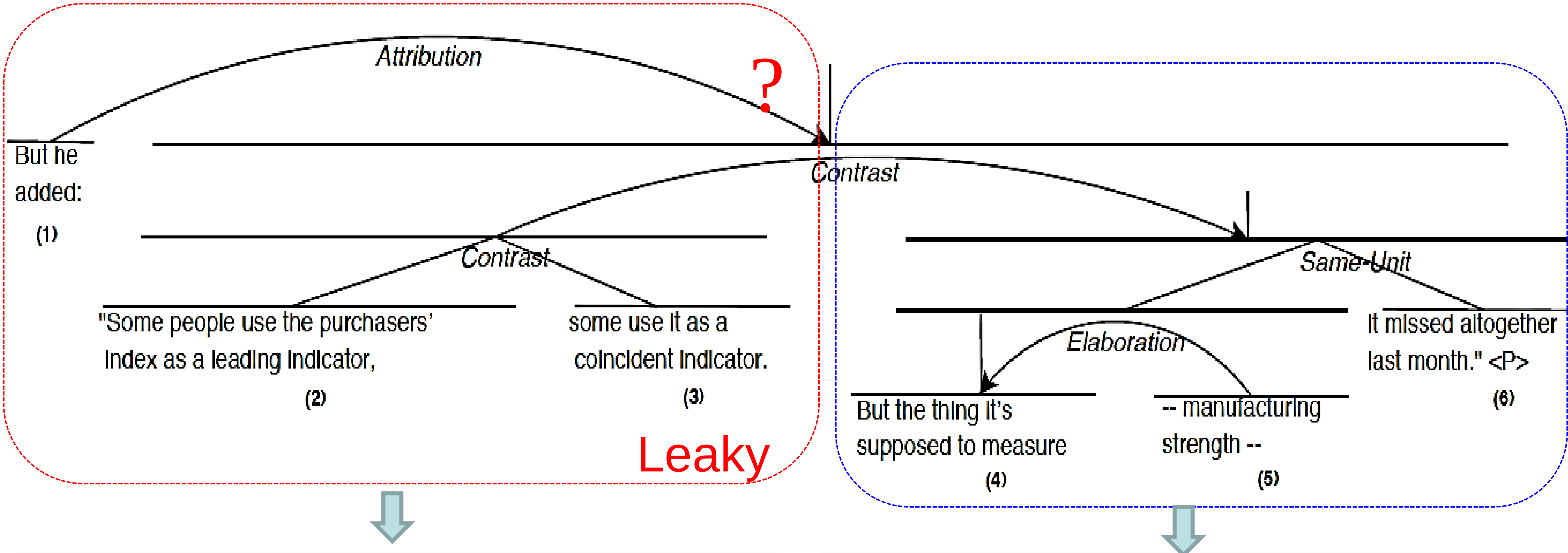
Discourse Parsing Task

Assume a text is already segmented into EDUs.



Observations (1)

- Nb of valid trees grows exponentially with the Nb of EDUs.

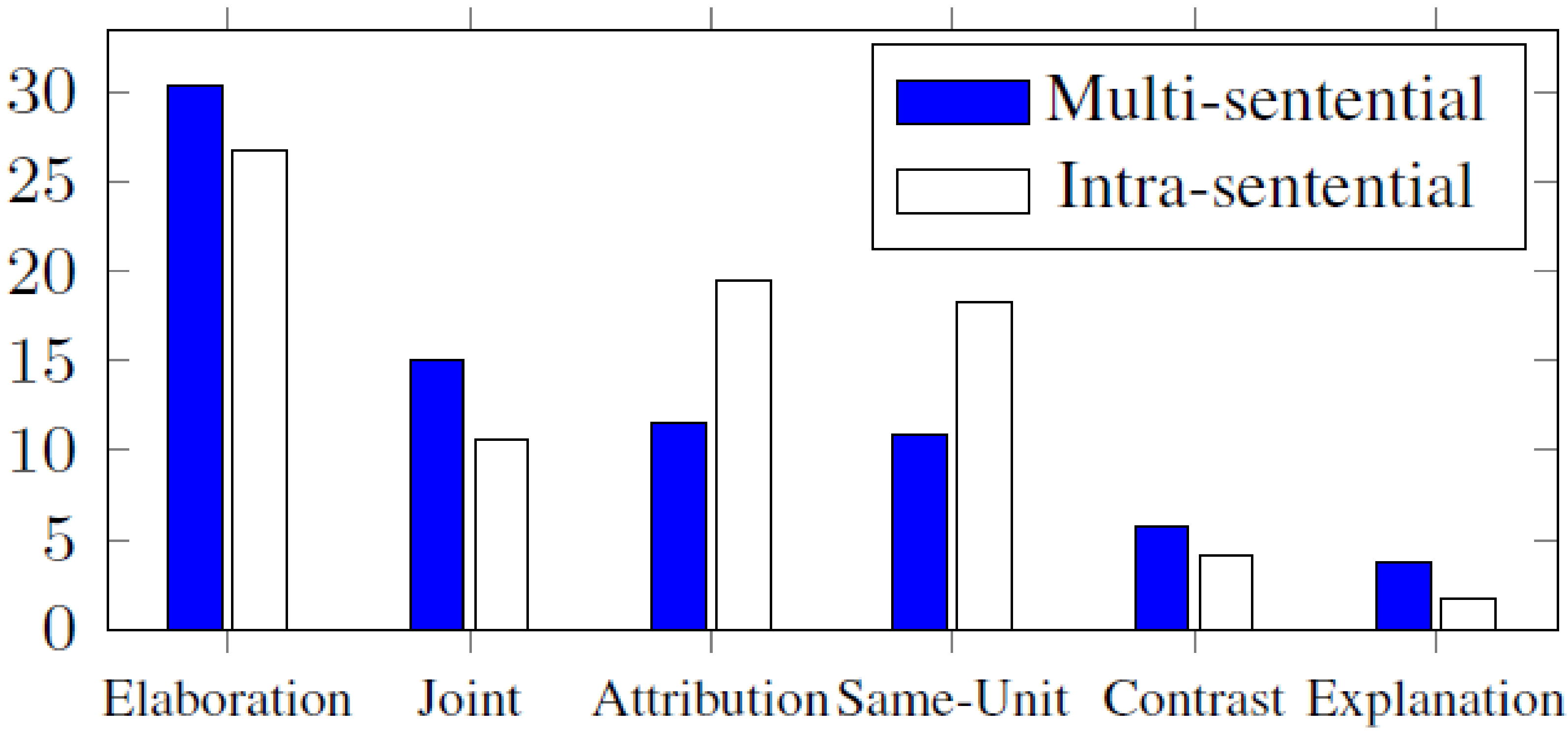


• 80% of the 5% merge with the adjacent sentences.
• Sliding window: build DTs for two adjacent sentences.

• More than 95% sentences have a well-defined DT.
• Build DTs for sentences first, then build on top of those.

Observations (2)

- Single model or two different models?
- Relations are distributed differently.



- Features don’t generalize.

Previous Work (1)

Soricut & Marcu, (2003)	
SPADE — Sentence level	Segmenter & Parser
Generative approach	√
Lexico-syntactic features	√
Structure & Label dependent	X
Sequential dependencies	X
Hierarchical dependencies	X

Hernault et al. (2010)	
HILDA — Document level	Segmenter & Parser
Discriminative approach	√
Structure & Label Jointly	X
Optimal	X
Sequential dependencies	X
Separate models	X

Newspaper (WSJ) articles

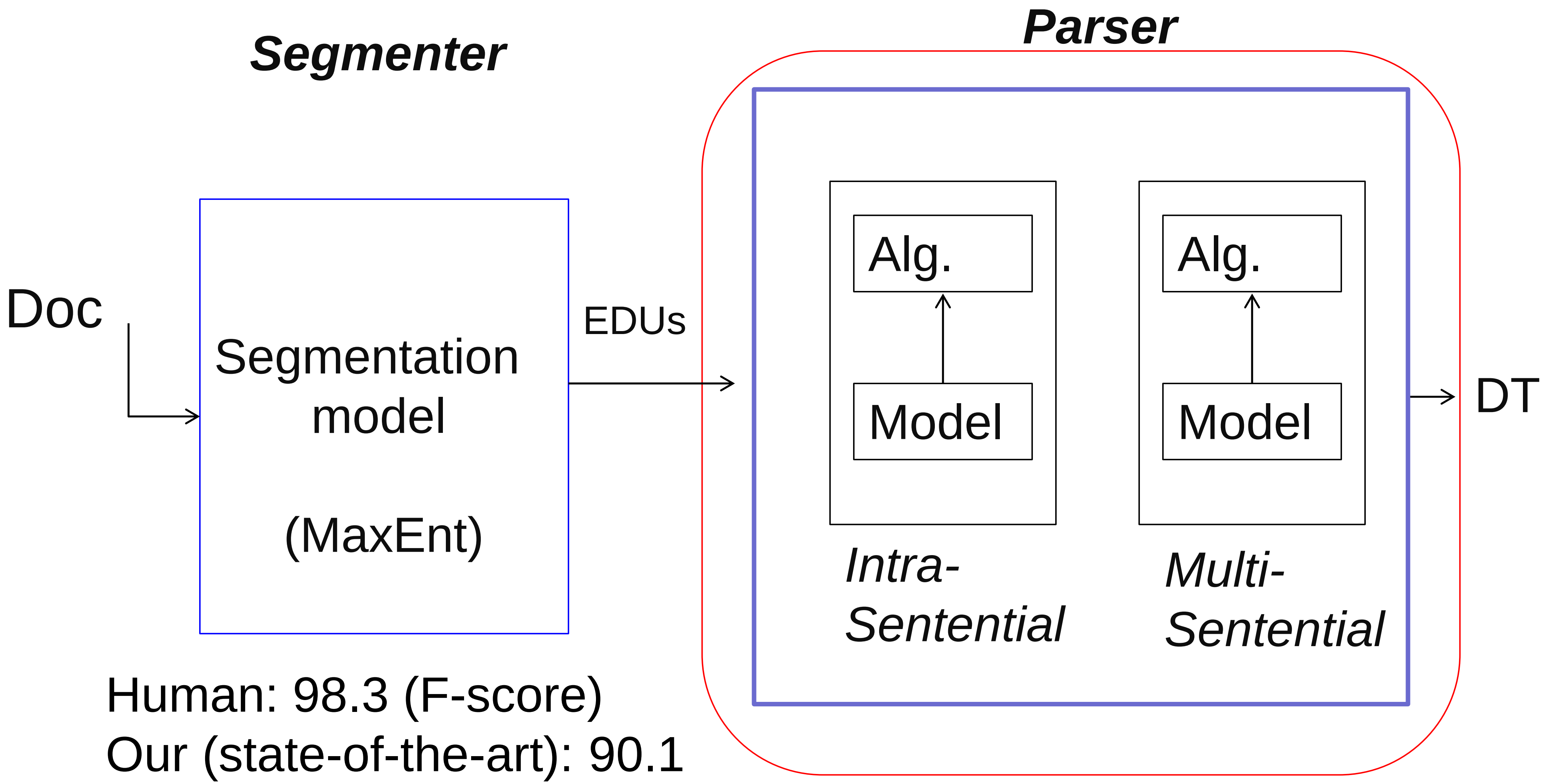
Previous Work (2)

Feng & Hirst, (2012)	Subba & Di-Eugenio, (2009)
HILDA	Shift-reduce — Only Parser
More linguistic features	Sentence + Document level
Dependency & constituency Contextual Discourse rules Lexical similarity Cue phrases	ILP-based classifier √ Compositional semantics √ Optimal X Sequential dependencies X Hierarchical dependencies X Separate models X
WSJ articles	Instructional manuals

Requirements from Our Parser

- Joint modeling of structure and relation
- Sequential and hierarchical dependencies
- Discriminate intra- vs. multi-sentential parsing
- Handle leaky sentence boundaries
- Optimal parsing algorithm
- Support k-best discourse parsing and reranking

Our Discourse Analysis Framework: CODRA

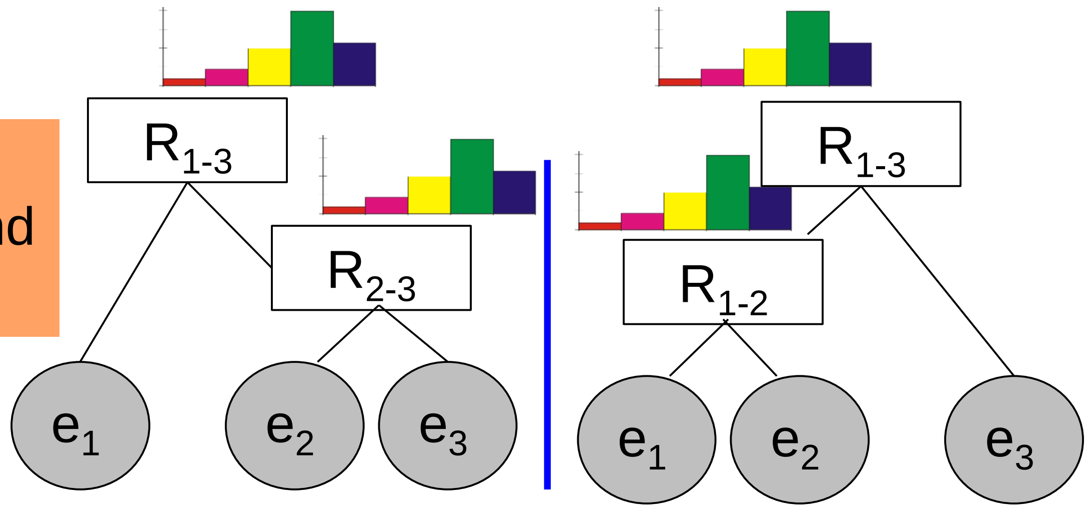


Human: 98.3 (F-score)
Our (state-of-the-art): 90.1

Our Discourse Parser

Parsing model (different for intra- and multi-sentential)

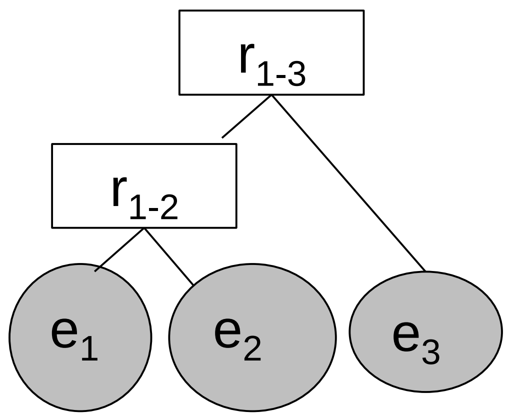
Assign probabilities to candidate DTs and their constituents.



R ranges over set of relations

Parsing algorithm (same for intra- and multi-sentential)

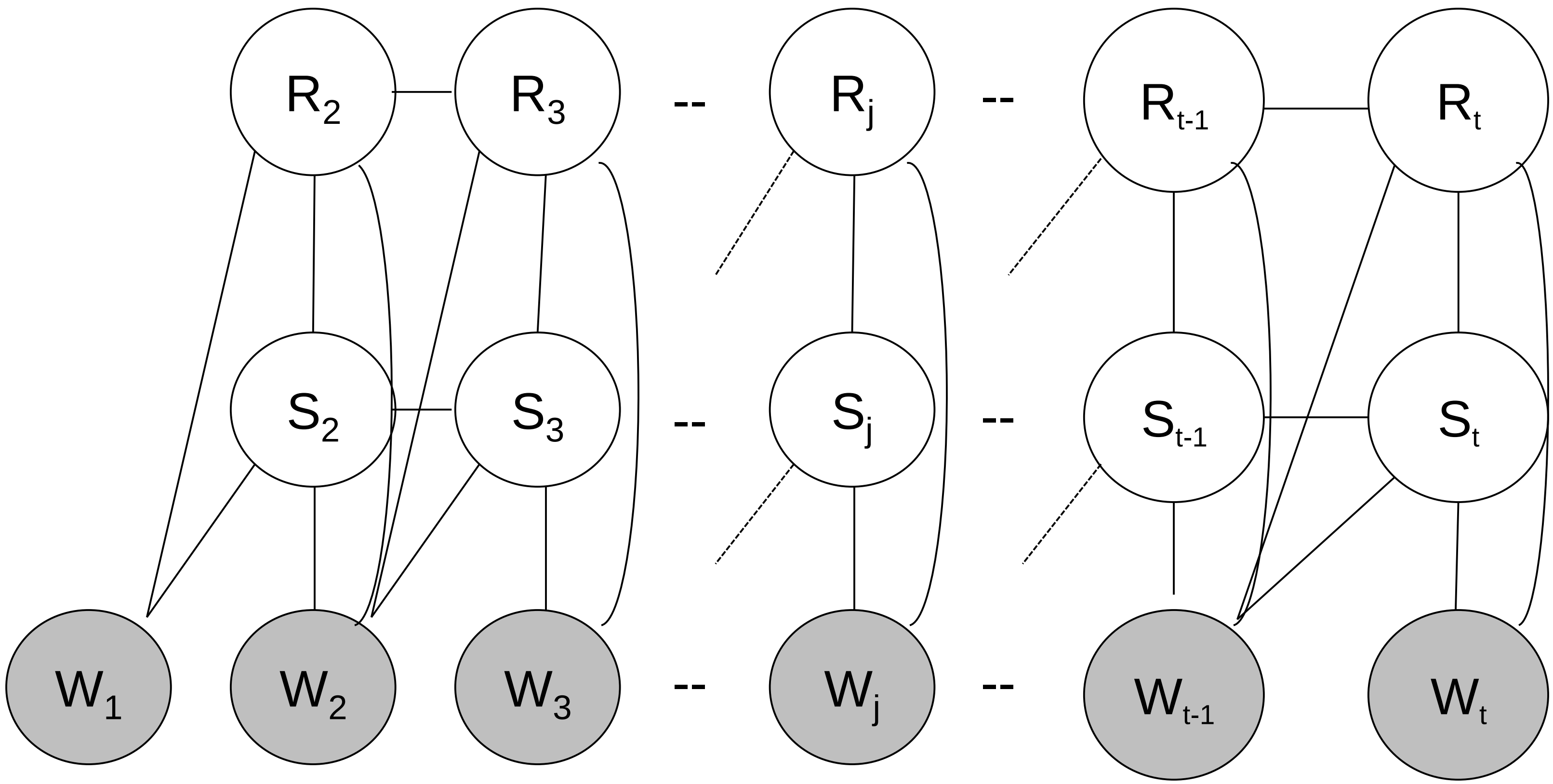
Find the (k-)most probable DT(s)



The Sentence-level Parsing Model

Structure and **label** jointly

Sequential dependencies



Relation at level i
R ϵ {1 .. M}

Structure at level i
S ϵ {0, 1}

Units at level i

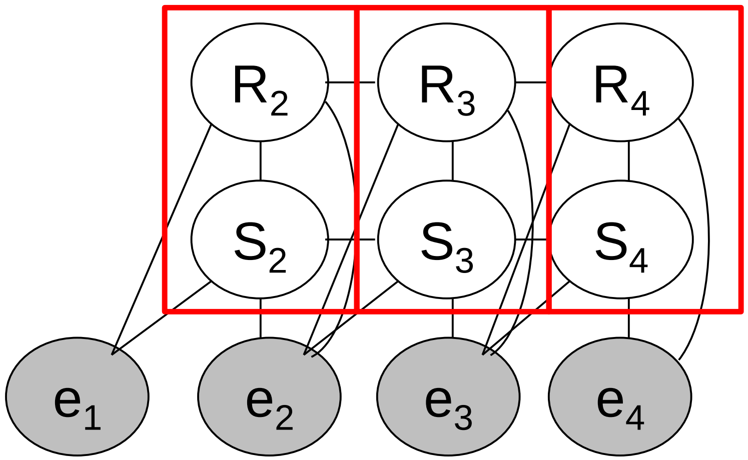
$$P(R_{2:t}, S_{2:t}|\mathbf{x}, \Theta_s) = \frac{1}{Z(\mathbf{x}, \Theta_s)} \prod_{i=2}^{t-1} \phi(R_i, R_{i+1}|\mathbf{x}, \Theta_{s,r}) \psi(S_i, S_{i+1}|\mathbf{x}, \Theta_{s,s}) \omega(R_i, S_i|\mathbf{x}, \Theta_{s,c})$$

Dynamic Conditional Random Field (DCRF) [Sutton et al, 2007]

Obtaining Probabilities (Sentence-level)

Apply DCRF recursively at different levels and compute posterior marginals of relation-structure pairs

Level 1

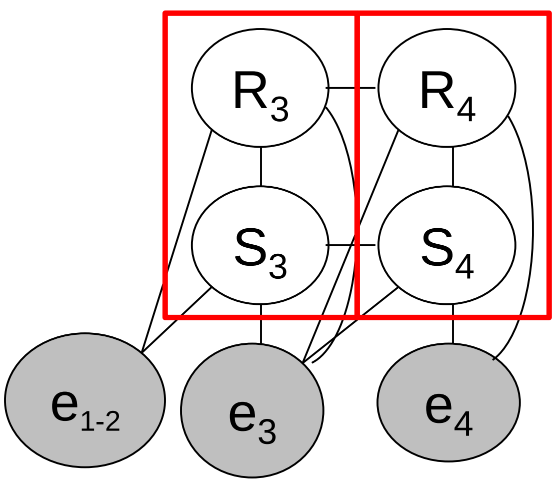


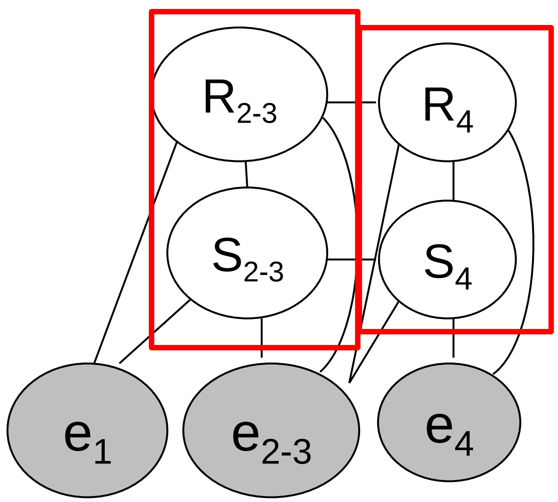
$P(R_2,S_2=1|e1,e2,e3,e4,\Theta)$

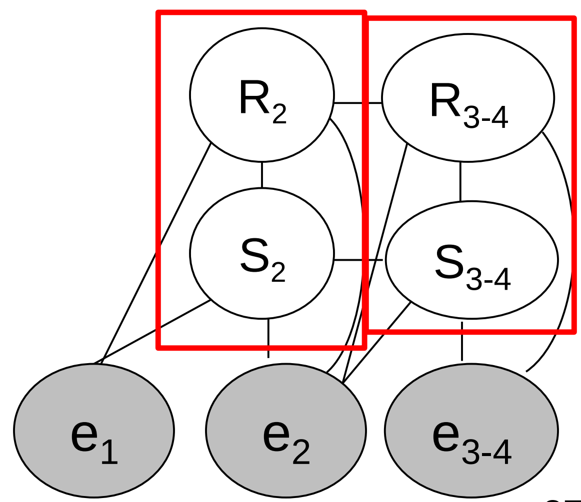
$P(R_3,S_3=1|e1,e2,e3,e4,\Theta)$

$P(R_4,S_4=1|e1,e2,e3,e4,\Theta)$

Level 2



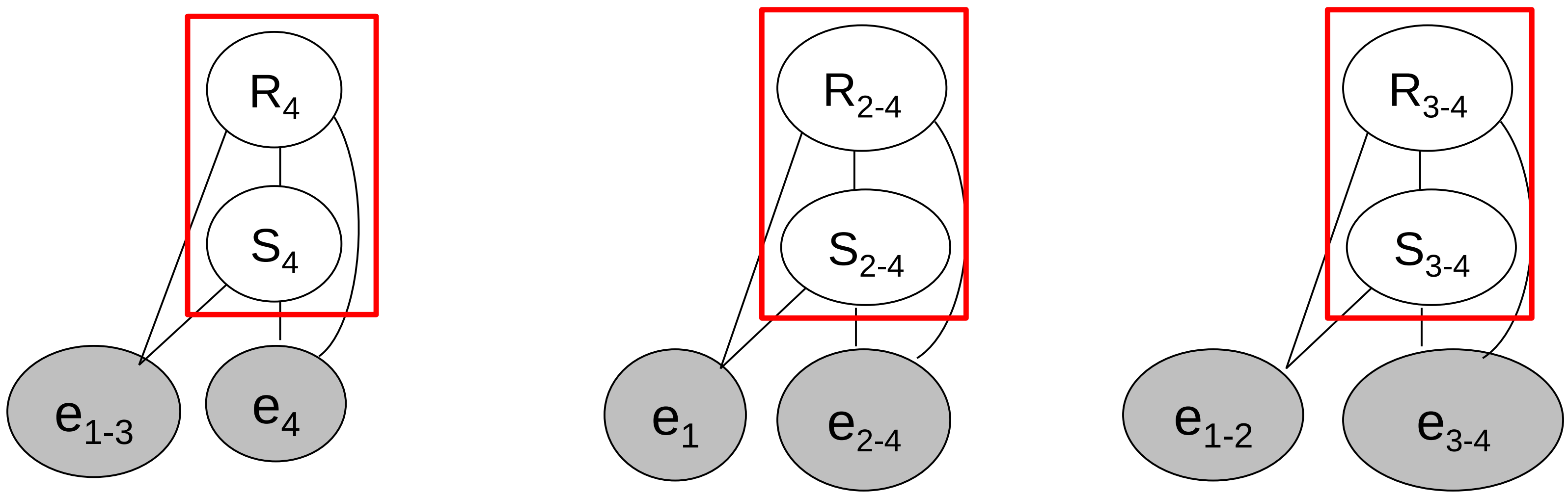




Obtaining Probabilities (Sentence-level)

Apply DCRF recursively at different levels and compute posterior marginals of relation-structure pairs

Level 3



Multi-sentential Parsing

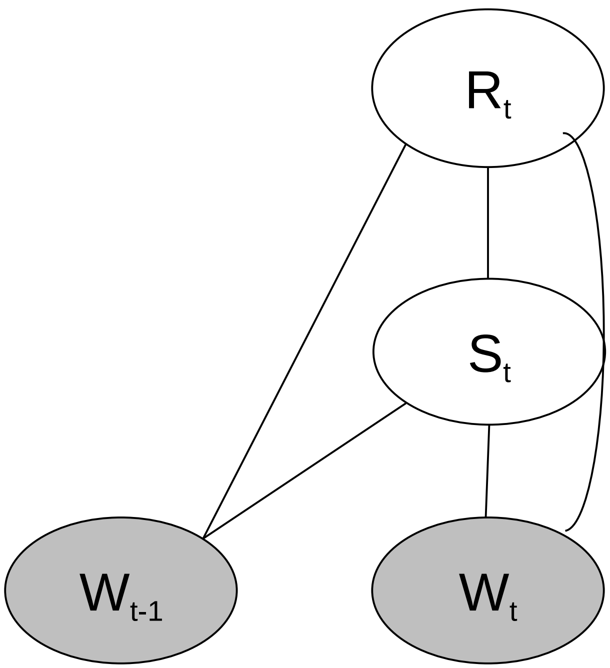
- Why not the same model used for intra-sentential?

  - "Fat" chain-structure=> exact inference=> Forwards-backwards
  - Forwards-backwards costs $O(TM^2)$ for a sequence.
  - # of possible sequences for a doc. with n sentences: $O(n^3)$

    - Total training cost: $O(D\ TM^2\ n^3)$

    Not scalable to document-level

Multi-sentential Parsing Model



Relation
R ϵ {1 .. M}

Structure
S ϵ {0, 1}

Units at level i

- Model **structure** and **label** jointly
- **Break the chain-structure**
- **CRF**
- Inference costs $O(M^2)$
- Allows balancing

$$P(R_t, S_t|\mathbf{x}, \Theta_d) = \frac{1}{Z(\mathbf{x}, \Theta_d)}\phi(R_t, S_t|\mathbf{x}, \Theta_d)$$

Dramatically reduces learning time

Features Used in Parsing Models

8 Organizational features	***Intra & Multi-Sentential***
Number of EDUs in *unit 1* (or *unit 2*).	
Number of tokens in *unit 1* (or *unit 2*).	
Distance of unit 1 in EDUs to the *beginning* (or to the *end*).	
Distance of unit 2 in EDUs to the *beginning* (or to the *end*).	
4 Text structural features	***Multi-Sentential***
Number of sentences in *unit 1* (or *unit 2*).	
Number of paragraphs in *unit 1* (or *unit 2*).	
8 N-gram features $N \in \{1, 2, 3\}$	***Intra & Multi-Sentential***
Beginning (or *end*) lexical N-grams in unit 1.	
Beginning (or *end*) lexical N-grams in unit 2.	
Beginning (or *end*) POS N-grams in unit 1.	
Beginning (or *end*) POS N-grams in unit 2.	
5 Dominance set features	***Intra-Sentential***
Syntactic labels of the *head* node and the *attachment* node.	
Lexical heads of the *head* node and the *attachment* node.	
Dominance relationship between the two units.	
9 Lexical chain features	***Multi-Sentential***
Number of chains spanning unit 1 and unit 2.	
Number of chains start in unit 1 and end in unit 2.	
Number of chains *start* (or *end*) in *unit 1* (or in *unit 2*).	
Number of chains skipping both unit 1 and unit 2.	
Number of chains skipping *unit 1* (or *unit 2*).	
2 Contextual features	***Intra & Multi-Sentential***
Previous and *next* feature vectors.	
2 Sub-structural features	***Intra & Multi-Sentential***
Root nodes of the *left* and *right* rhetorical sub-trees.	

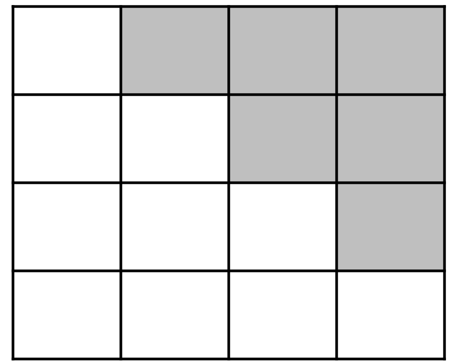
Hierarchical dependencies

Parsing Algorithm (1)

Probabilistic CKY-like bottom-up algorithm

4 EDUs

4X4



A

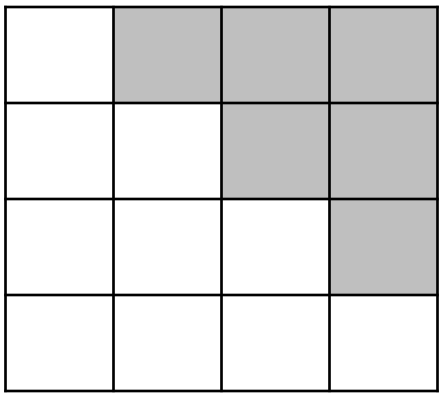
$$A[i,j] = P(r^*[U_i^b, U_{m^*}^e, U_j^e]), \text{ where}$$

$$(m^*, r^*) = \underset{i \le m < j\ ;\ R}{\operatorname{argmax}} P(R[U_i^b, U_m^e, U_j^e]) \times A[i,m] \times A[m+1,j]$$

R ranges over set of relations

Finds global optimal

Parsing Algorithm (2)



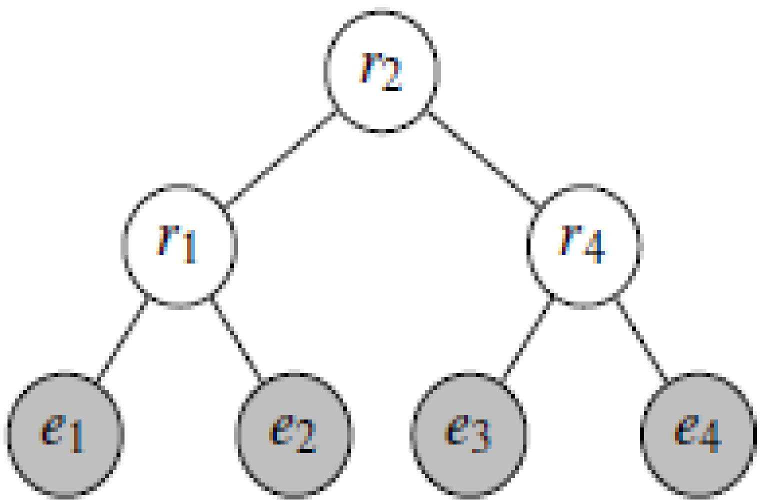
A

	1	1	2
		2	2
			3

B

	r_1	r_3	r_2
		r_2	r_3
			r_4

C



k-best Parsing

- Extension to k-best is straight-forward
- Store and keep track of k-best candidates at each step

- Complexity for n discourse units and M relations

1-best parsing

Time: $O(n^3M)$
Space: $O(n^2)$

k-best parsing

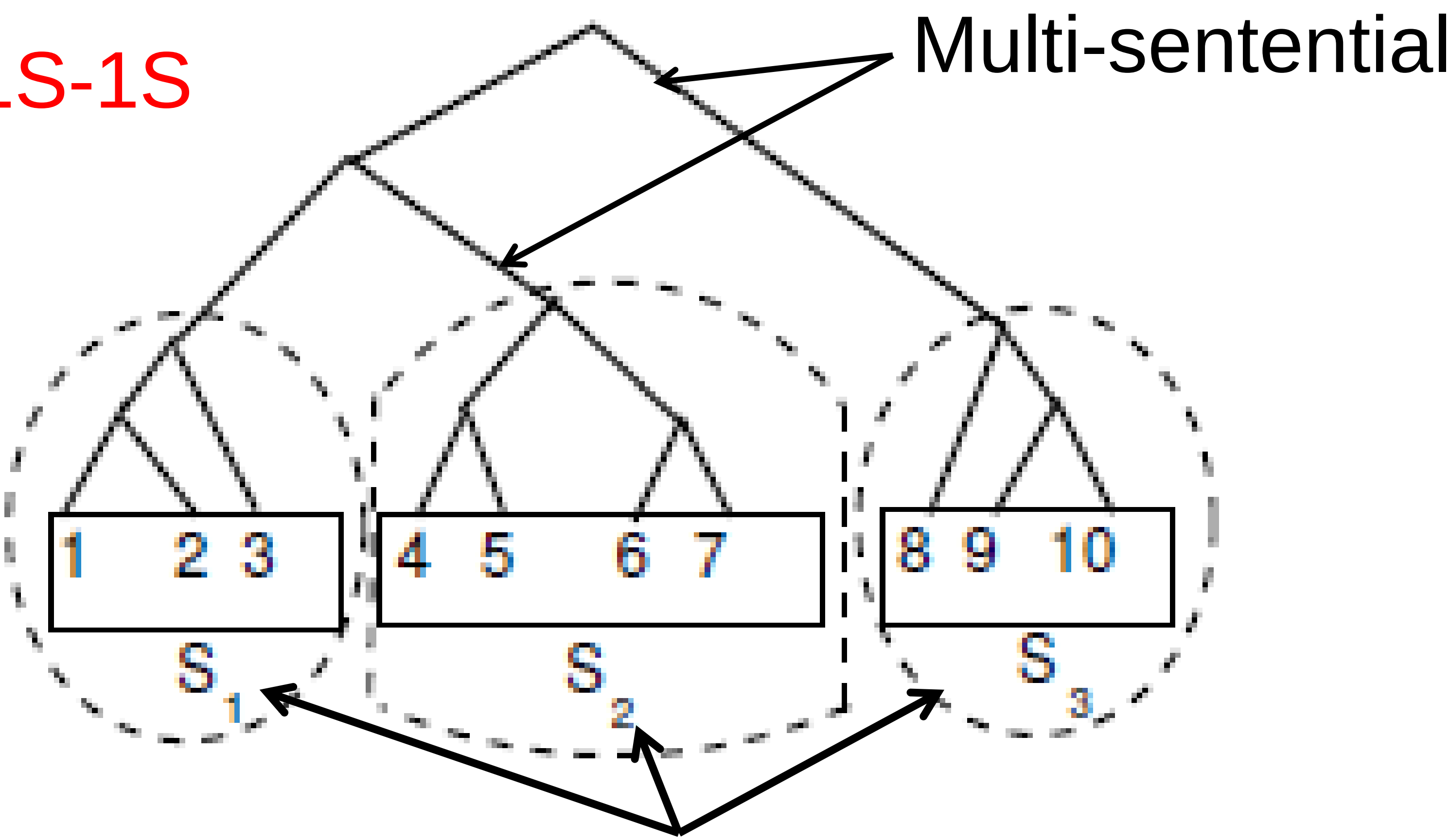
Time: $O(n^3Mk^2 \log k)$
Space: $O(n^2k)$

- See (Huang and Chiang, 2005) for cleverer ways to reduce the complexity.

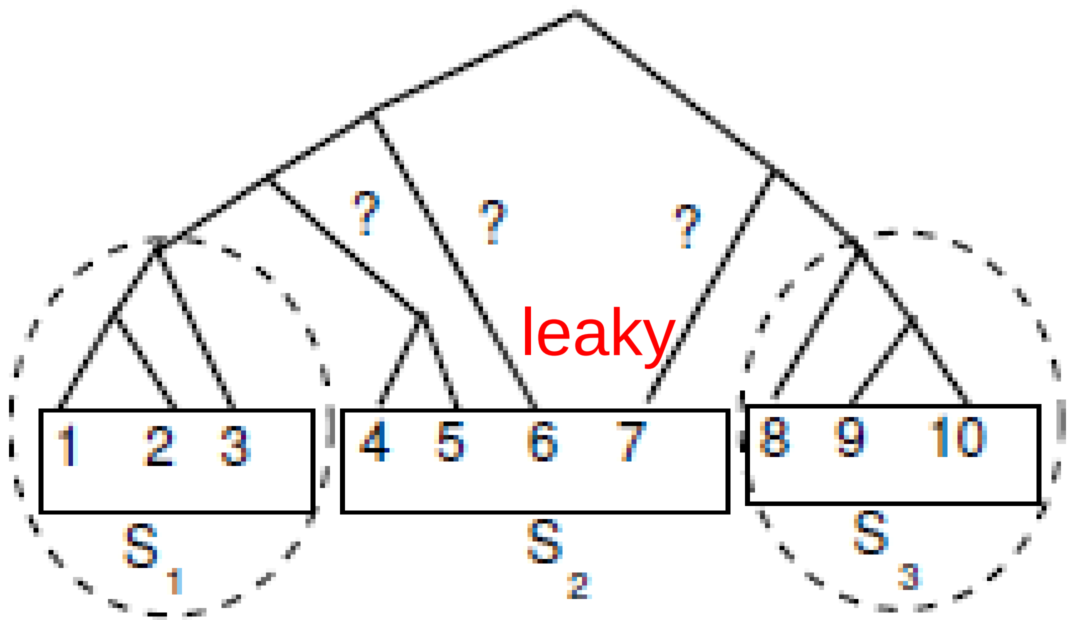
Combining Intra and Multi-sentential (1)

- Most sentences have a well-defined DT.
- Build DTs for sentences first, then build on top of those

1S-1S



Combining Intra- and Multi-sentential (2)



- 5% sentences have leaky boundaries in RST-DT.
- 12% sentences have leaky boundaries in Instructional domain.

- 80% of the 5% merge with the adjacent sentences in RST-DT.
- 75% of the 12% merge with the adjacent sentences in Inst. dom.

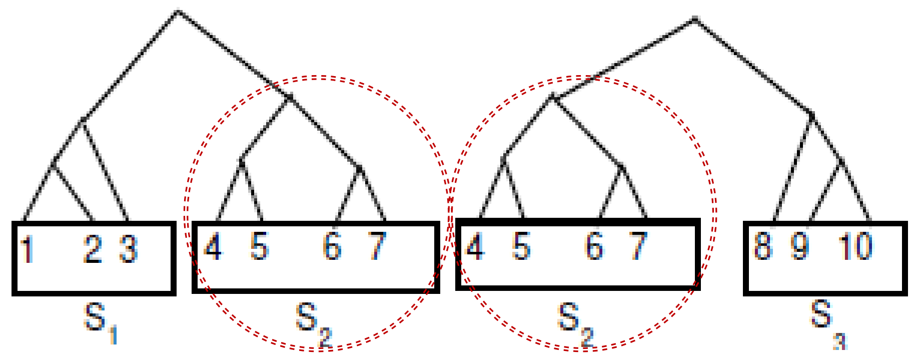
Combining Intra- and Multi-sentential (3)

Apply intra-sentential parser to each window of 2 sentences and consolidate the decisions to generate sentence-level sub-trees.

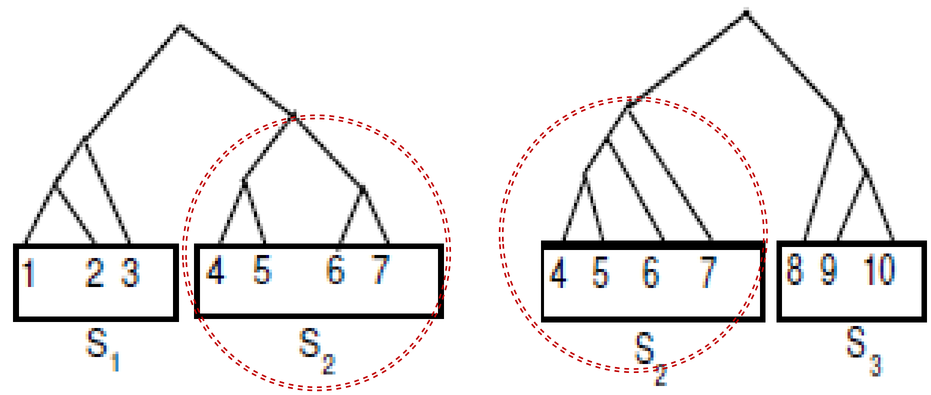
- Consolidation:

Sliding Window

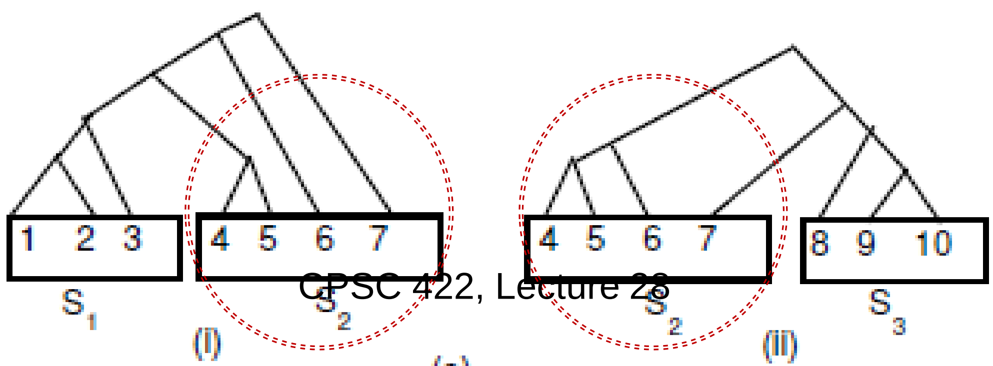
a) Same in both DTs.



b) Different but no cross.



c) Cross.



Experiments: Corpora/Datasets

RST-DT corpus

(Carlson & Marcu, 2001)

- 385 news articles

  -Train: 347 (7673 sent.)

  -Test: 38 (991 sent.)

Relations

- 18 relations
- 41 with Nucleus-Satellite

Instructional corpus

(Subba & Di-Eugenio, 2009)

- 176 how-to-do manuals

  3430 sentences

Relations

- 26 primary relations
- Reversal of non-commutative as separate relations
- 76 with Nucleus-Satellite

Experiments: Intra-sentential Parsing

Results based on manual segmentation

	RST-DT				Instructional	
	Test set		10-fold	Doubly	S&E	10-fold
Scores	**SPADE**	**OUR**	**OUR**	**Human**	**ILP**	**OUR**
Span	93.5	**96.5**	95.7	95.7	92.9	**98.3**
Nuclearity	85.8	**89.4**	88.6	90.4	71.8	**89.4**
Relation	67.6	**79.8**	78.9	83.0	63.0	**75.8**

Our model outperforms the state-of-the-art by a wide margin, especially on relation labeling

Experiments: Intra-sentential Parsing

Parsing based on automatic segmentation

	RST-DT			Instructional
	Test set		10-fold	10-fold
Scores	**SPADE**	**DCRF**	**DCRF**	**DCRF**
Span	76.7	**82.4**	80.1	**76.9**
Nuclearity	70.2	**76.6**	75.2	**67.6**
Relation	58.0	**67.5**	66.8	**57.5**

- Our model outperforms SPADE by a wide margin
- Inaccuracies in segmentation affects parsing on Instructional corpus

Experiments: Document-level Parsing

Results based on manual segmentation

	RST-DT				Instructional		
Scores	**HILDA**	**OUR (1-1)**	**OUR (SW)**	**Human**	**ILP**	**OUR (1-1)**	**OUR (SW)**
Span	74.7	**82.6**	**83.8**	88.7	70.4	**80.7**	**82.5**
Nuclearity	60.0	**68.3**	**68.9**	77.7	49.5	**63.0**	**64.8**
Relation	44.3	**55.8**	**55.9**	65.8	35.4	**43.5**	**44.3**

- Our model outperforms the state-of-the-art by a wide margin.
- Not significant difference between 1S-1S and SW in RST-DT.

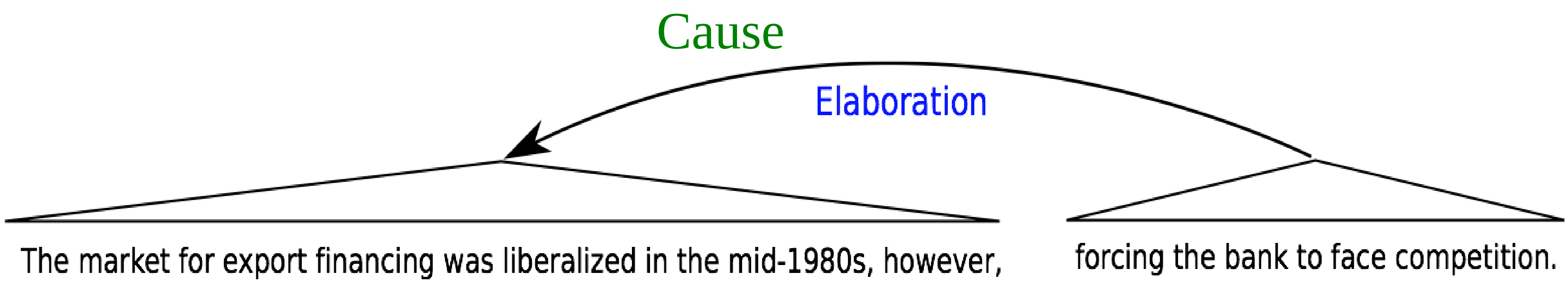
Error Analysis: Relation Labeling

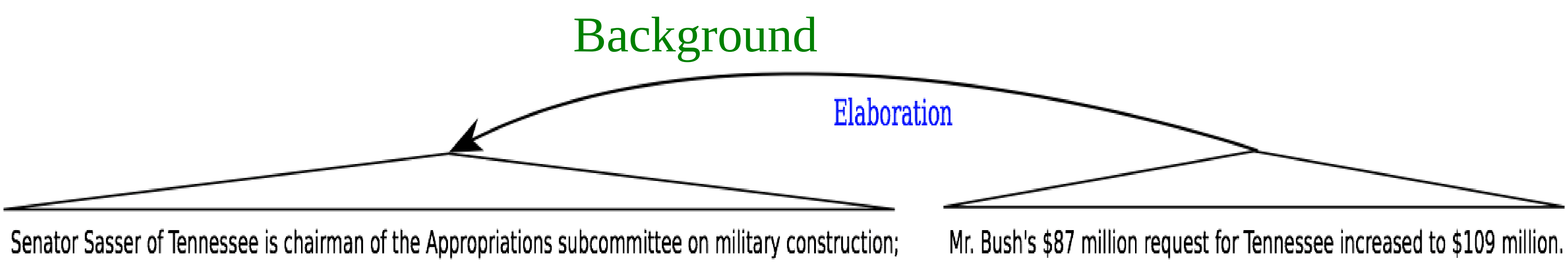
	T-C	T-O	T-CM	M-M	CMP	EV	SU	CND	EN	CA	TE	EX	BA	CO	JO	S-U	AT	EL
T-C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
T-O	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T-CM	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2	0	0	7
M-M	0	0	0	10	0	0	0	0	0	0	0	1	1	0	0	0	1	3
CMP	0	0	0	1	4	0	0	1	0	1	0	3	3	0	1	1	0	2
EV	0	0	0	0	0	0	0	0	0	0	0	2	0	0	2	0	2	11
SU	0	0	0	0	0	0	8	0	0	0	0	0	0	0	1	0	0	12
CND	0	0	0	0	0	0	0	22	0	0	0	0	1	3	0	0	3	2
EN	0	0	0	0	0	0	0	1	24	1	0	0	0	0	0	0	1	7
CA	0	0	0	0	0	0	0	0	2	3	0	4	2	2	7	0	3	11
TE	0	0	0	1	0	0	0	1	2	0	7	1	9	1	9	0	3	4
EX	0	0	0	1	0	0	0	0	1	5	0	12	0	1	3	0	3	12
BA	0	0	0	1	0	0	0	1	0	1	4	1	19	2	6	1	5	12
CO	0	0	0	1	2	0	0	2	0	1	3	2	2	33	7	0	0	9
JO	0	0	0	0	0	0	1	2	0	1	1	1	1	2	57	1	0	13
S-U	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	85	1	0
AT	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	3	272	9
EL	0	1	0	0	0	0	0	0	14	6	1	8	1	0	8	2	2	359

- Most frequent ones confuse less frequent ones
- Hard to distinguish semantically similar relations

Error Analysis: Examples

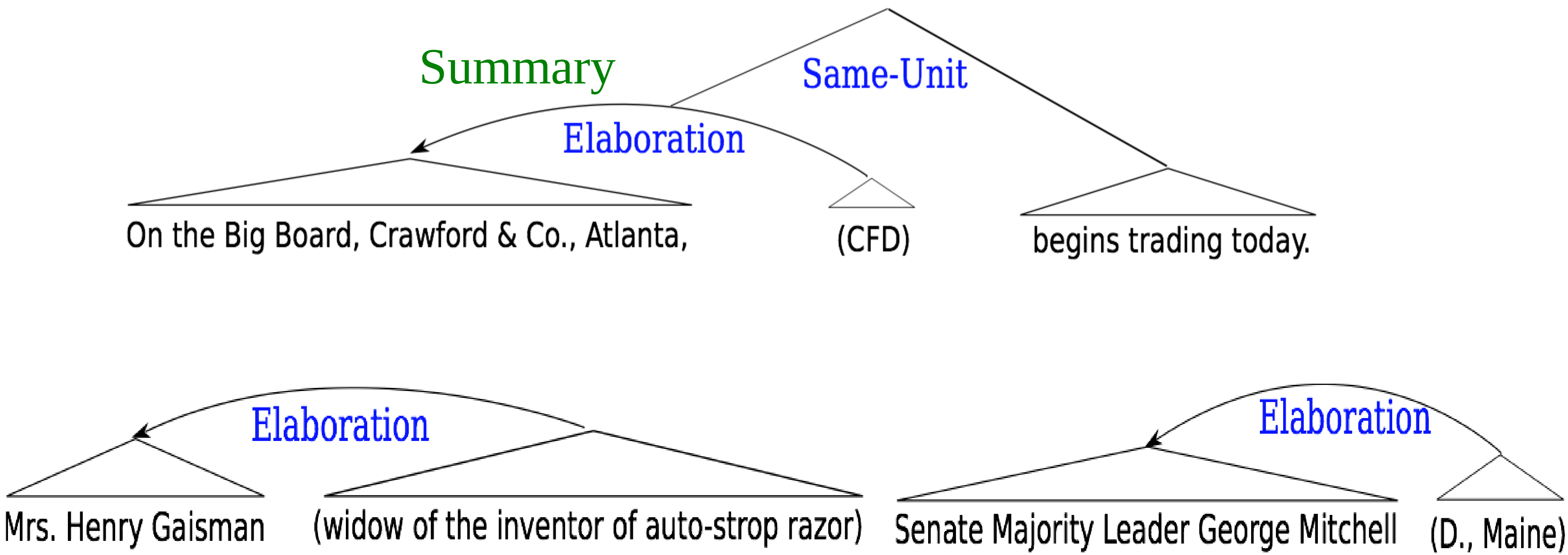
Confusion between Cause/Background and Elaboration





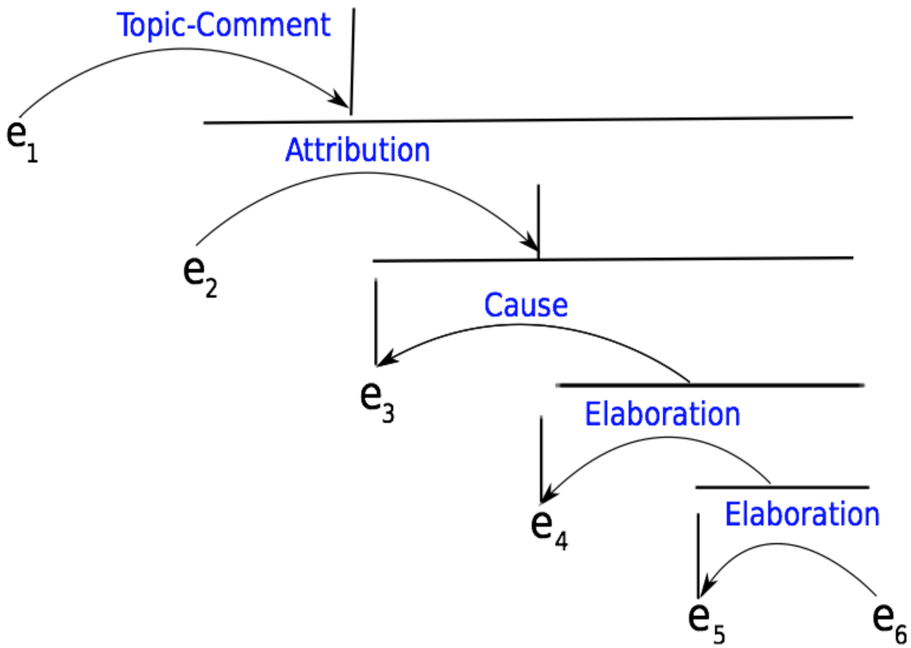
Error Analysis: Examples

Confusion between Summary and Elaboration

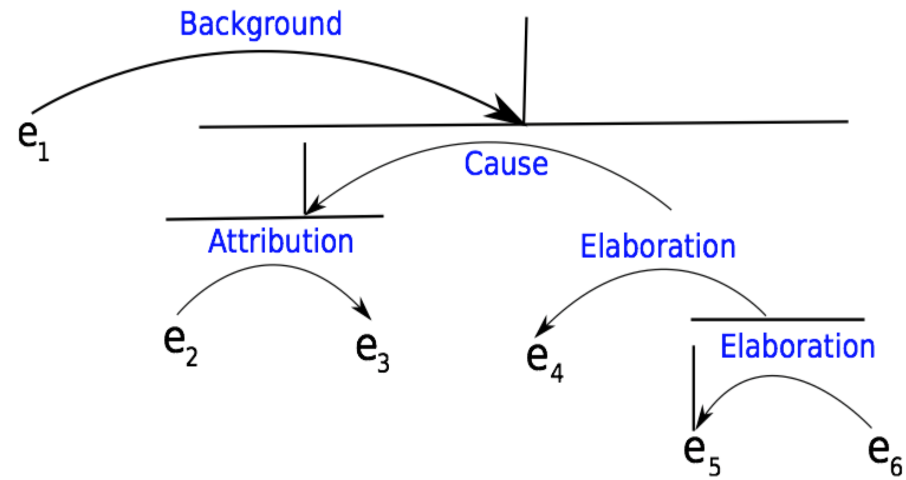


Error Analysis: Examples

Long range structural dependencies



(a) A human-annotated discourse tree.



(b) A system-generated discourse tree.

[what's more,]$_{e_1}$ [he believes]$_{e_2}$ [seasonal swings in the auto industry this year aren't occurring at the same time in the past,]$_{e_3}$ [because of production and pricing differences]$_{e_4}$ [that are curbing the accuracy of seasonal adjustments]$_{e_5}$] [built into the employment data.]$_{e_6}$

K-best reranking with Tree Kernels

A Novel Discriminative Framework for Sentence-Level Discourse Analysis

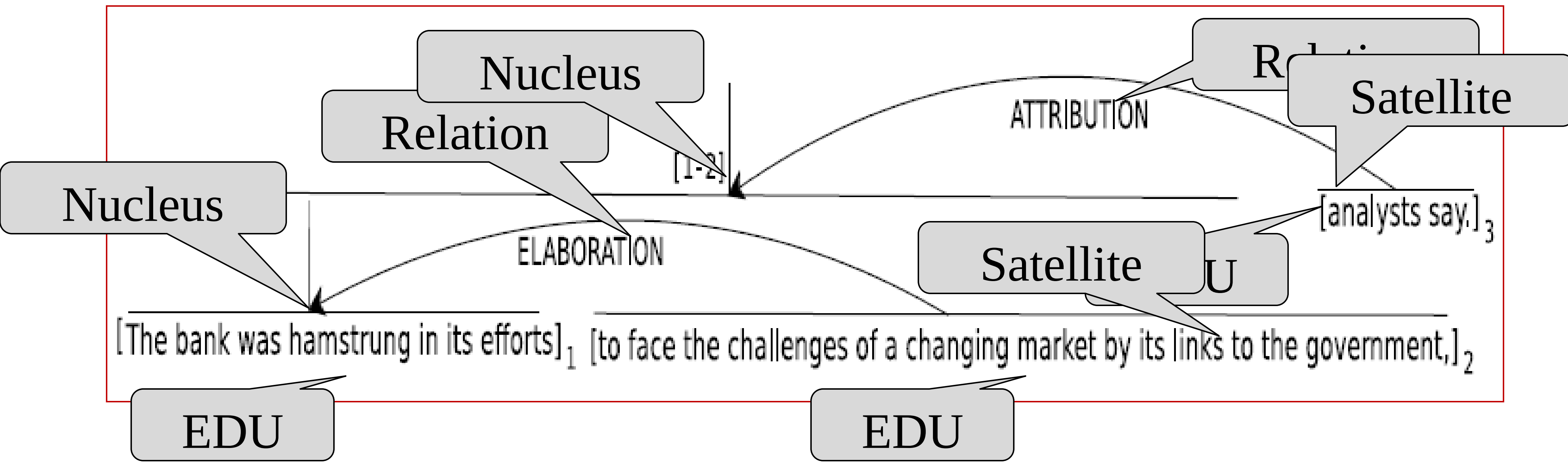
Shafiq Joty

In collaboration with

Giuseppe Carenini, Raymond T. Ng

Discourse Analysis in RST

The bank was hamstrung in its efforts to face the challenges of a changing market by its links to the government, analysts say.



Motivation

- ✓ Text summarization (Marcu, 2000)
- ✓ Text generation (Prasad et al., 2005)
- ✓ Sentence compression (Sporleder & Lapata, 2005)
- ✓ Question Answering (Verberne et al., 2007)

Outline

- Previous work
- Discourse parser
- Discourse segmenter
- Corpora/datasets
- Evaluation metrics
- Experiments
- Conclusion and future work

Previous Work (1)

Soricut & Marcu, (2003)

SPADE

Sentence level

Segmenter & Parser

- Generative approach √
- Lexico-syntactic features √
- Structure & Label dependent X
- Sequential dependencies X
- Hierarchical dependencies X

Hernault et al. (2010)

HILDA

Document level

Segmenter & Parser

- SVMs √
- Large feature set √
- Optimal
- Sequential dependencies X

Newspaper articles

Previous Work (2)

Subba & Di-Eugenio, (2009)

Shift-reduce $\left\{ \begin{array}{l} \text{Only} \\ \text{Parser} \end{array} \right\}$

Sentence + Document level

ILP-based classifier √
Compositional semantics √
Optimal X
Sequential dependencies X
Hierarchical dependencies X

Instructional manuals

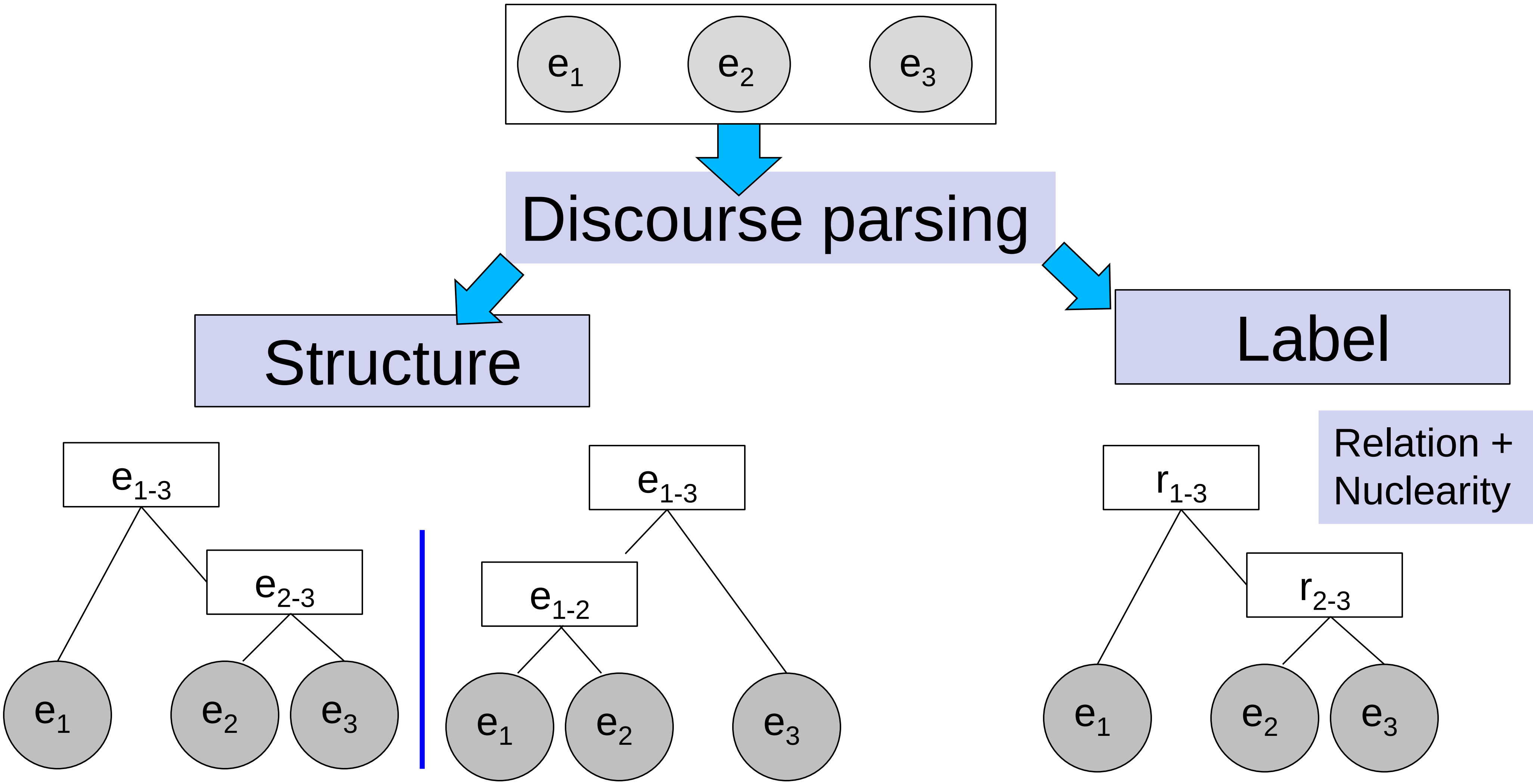
Fisher & Roark (2007)

Binary log-linear $\left\{ \begin{array}{l} \text{Only} \\ \text{Segmenter} \end{array} \right\}$

State-of-the-art performance
Parse-tree features are important

Discourse Parsing

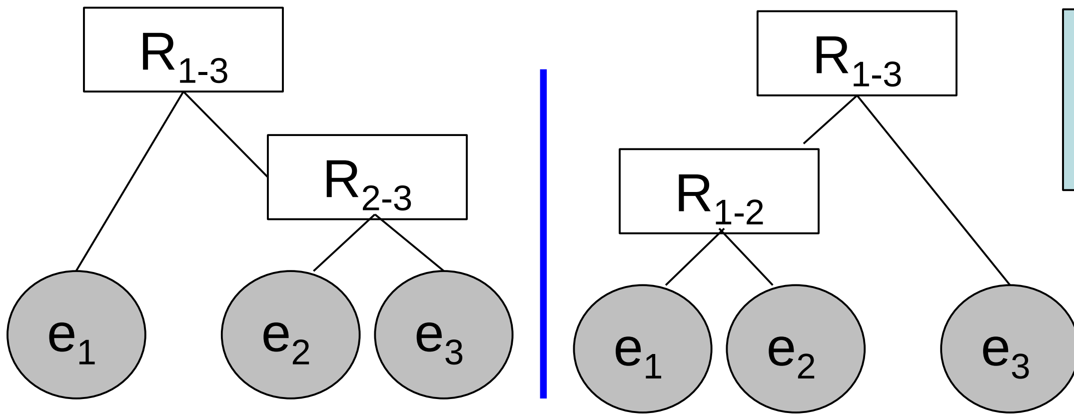
Assume a sentence is already segmented into EDUs.



Our Discourse Parser

Parsing model

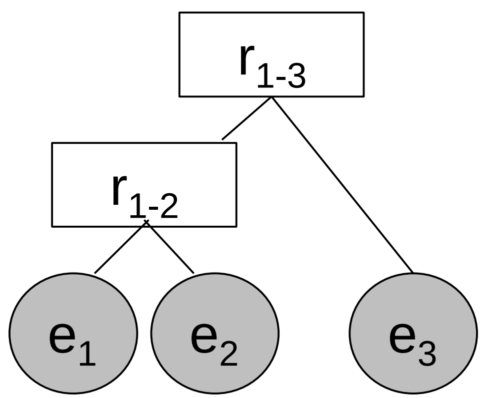
Assign probabilities to DTs.



R ranges over set of relations

Parsing algorithm

Find the most probable DT

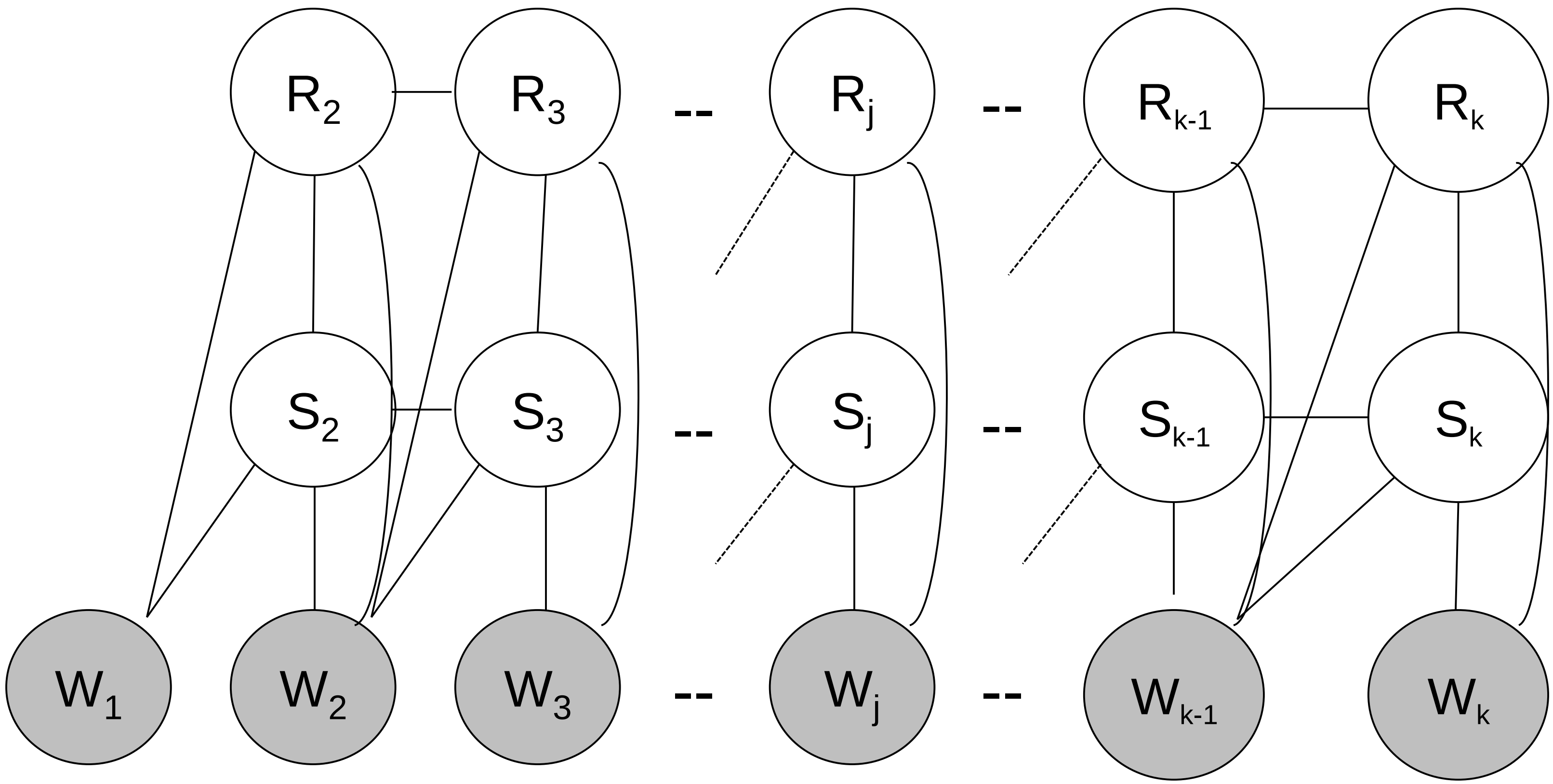


Requirements for Our Parsing Model

- ✓ Discriminative
- ✓ Joint model for Structure and Label
- ✓ Sequential dependencies
- ✓ Hierarchical dependencies
- ✓ Should support an optimal parsing algorithm

Our Parsing Model

Model **structure** and **label** jointly



Relation at level i
$R \in \{1 .. M\}$

Structure at level i
$S \in \{0, 1\}$

Spans at level i

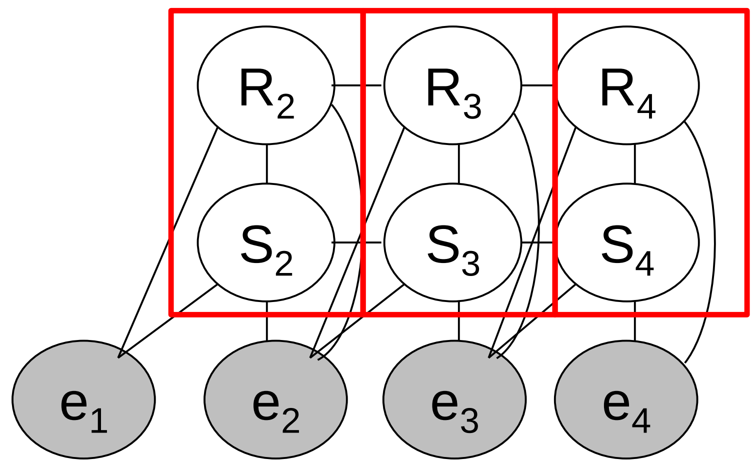
Dynamic Conditional Random Field (DCRF) [Sutton et al, 2007]

Models sequential dependencies

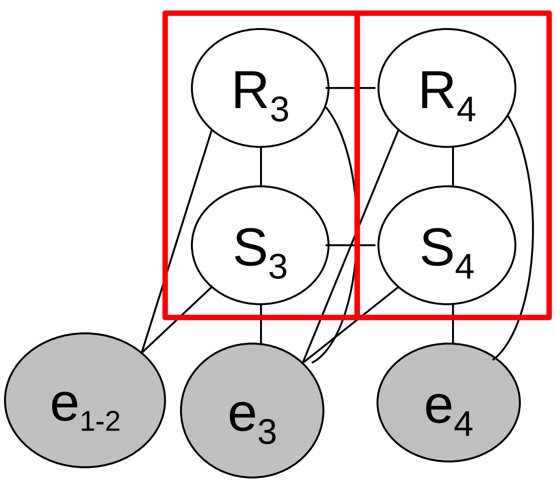
Obtaining probabilities

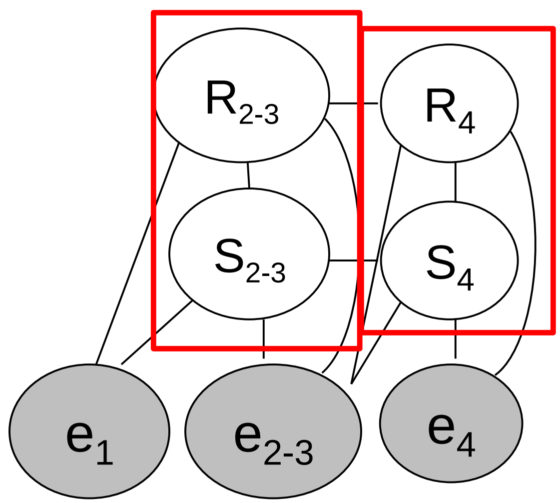
Apply DCRF recursively at different levels and compute posterior marginals of relation-structure pairs

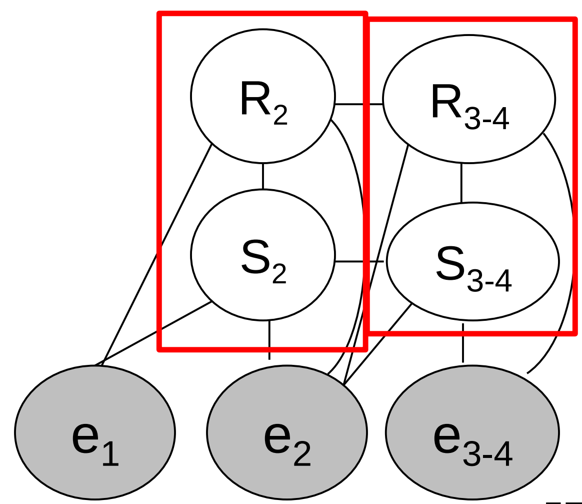
Level 1



Level 2



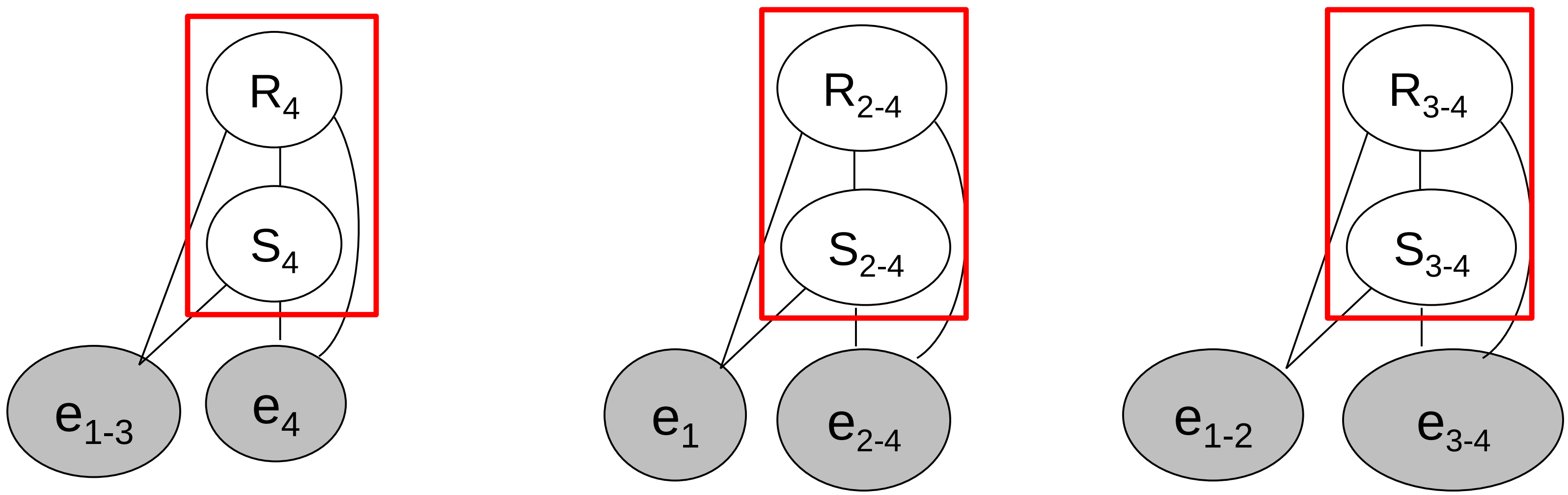




Obtaining probabilities

Apply DCRF recursively at different levels and compute posterior marginals of relation-structure pairs

Level 3



Features Used in Parsing Model

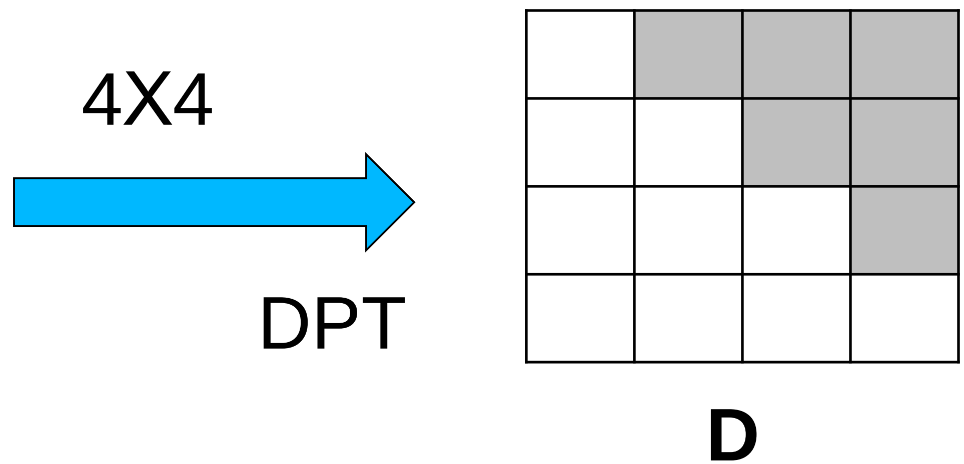
8 organizational features
Relative number of EDUs in *span 1* and *span 2*.
Relative number of tokens in *span 1* and *span 2*.
Distances of span 1 in EDUs to the *beginning* and to the *end*.
Distances of span 2 in EDUs to the *beginning* and to the *end*.
8 N-gram features
Beginning and *end* lexical N-grams in span 1.
Beginning and *end* lexical N-grams in span 2.
Beginning and *end* POS N-grams in span 1.
Beginning and *end* POS N-grams in span 2.
5 dominance set features (SPADE)
Syntactic labels of the *head* node and the *attachment* node.
Lexical heads of the *head* node and the *attachment* node.
Dominance relationship between the two text spans.
2 contextual features
Previous and *next* feature vectors.
2 substructure features
Root nodes of the *left* and *right* rhetorical subtrees.

Hierarchical dependencies

Parsing Algorithm

Probabilistic CKY-like bottom-up algorithm

4 EDUs



$$(k^*, r^*) = \underset{i \le k \le j \; ; \; R \in \{1 \cdots M\}}{\operatorname{argmax}} P(R[U_i(0), U_k(1), U_j(1)]) \times D[i,k] \times D[k+1, j] \quad (3.9)$$

Finds global optimal

Outline

- Previous work
- Discourse parser
- Discourse segmenter
- Corpora/datasets
- Evaluation metrics
- Experiments
- Conclusion and future work

Discourse Segmentation

The bank was hamstrung in its efforts to face the challenges of a changing market by its links to the government, analysts say.

Discourse Segmentation

The bank was hamstrung in its efforts

EDU

to face the challenges of a changing market by its links to the government,

EDU

analysts say.

EDU

Segmentation is the primary source of inaccuracy (Soricut & Marcu, 2003)

Our Discourse Segmenter

- **Binary classification:** boundary or no-boundary
- **Logistic Regression** with L_2 regularization
- **Bagging** to deal with sparse boundary tags

Features used

SPADE features
Chunk and POS features
Positional features
Contextual features

Corpora/Datasets

RST-DT corpus (Carlson & Marcu, 2001)	Instructional corpus (Subba & Di-Eugenio, 2009)
• 385 news articles -Train: 347 (7673 sentences) -Test: 38 (991 sentences)	•176 how-to-do manuals 3430 sentences
Relations • 18 relations • 39 with Nucleus-Satellite	**Relations** • 26 primary relations • Reversal of non-commutative as separate relations • 70 with Nucleus-Satellite

Evaluation Metrics

Metrics for parsing accuracy
(Marcu, 2000)

- Unlabeled (Span)
- Labeled (Nuclearity, Relation)

Precision, Recall
F-measure

Metric for segmentation accuracy
(Soricut & Marcu, 2003; Fisher & Roark, 2007)

Intra-sentence EDU boundary

Precision, Recall
F-measure

Experiments (1)

Parsing based on manual segmentation

	RST-DT				Instructional	
	Test set		10-fold	Doubly	S&E	10-fold
Scores	**SPADE**	**DCRF**	**DCRF**	**Human**	**ILP**	**DCRF**
Span	93.5	**94.6**	93.7	95.7	92.9	**97.7**
Nuclearity	85.8	**86.9**	85.2	90.4	71.8	**87.2**
Relation	67.6	**77.1**	75.4	83.0	63.0	**73.6**

Our model outperforms the state-of-the-art by a wide margin, especially on relation labeling

Experiments (2)

Discourse segmentation

	RST-DT						Instructional	
	Test set				10-fold		10-fold	10-fold
Scores	**HILDA**	**SPADE**	**F&R**	**LR**	**SPADE**	**LR**	**SPADE**	**LR**
Precision	77.9	83.8	**91.3**	88.0	83.7	**87.5**	65.1	**73.9**
Recall	70.6	86.8	89.7	**92.3**	86.2	**89.9**	82.8	**89.7**
F-measure	74.1	85.2	**90.5**	90.1	84.9	**88.7**	72.8	**80.9**

Human agreement (F-measure): 98.3

- Our model outperforms SPADE and comparable to F&R
- We use fewer features than F&R

Experiments (3)

Parsing based on automatic segmentation

	RST-DT			Instructional
	Test set		10-fold	10-fold
Scores	**SPADE**	**DCRF**	**DCRF**	**DCRF**
Span	76.7	**80.3**	78.7	**71.9**
Nuclearity	70.2	**73.6**	72.2	**64.3**
Relation	58.0	**65.4**	64.2	**54.8**

- Our model outperforms SPADE by a wide margin
- Inaccuracies in segmentation affects parsing on Instructional corpus

Error analysis (Relation labeling)

	TO	EV	SU	MA	COMP	EX	COND	TE	CA	EN	BA	CONT	JO	SA	AT	EL
TO	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	2
EV	0	0	0	0	0	0	0	0	0	0	0	0	1	1	3	2
SU	0	0	6	0	0	0	0	0	0	0	0	1	2	0	0	10
MA	0	0	0	10	0	1	0	1	0	0	0	0	2	0	1	7
COMP	0	0	0	1	1	1	0	0	2	0	3	2	1	0	0	6
EX	0	0	0	0	0	9	0	0	4	1	2	0	0	1	4	1
COND	0	0	0	0	0	0	20	3	0	1	1	1	1	2	6	7
TE	0	0	0	0	0	0	0	11	1	0	5	0	9	4	2	9
CA	0	0	0	1	0	4	0	1	5	4	1	1	6	1	6	3
EN	0	0	0	1	0	0	0	1	0	24	2	0	1	1	1	9
BA	0	0	0	0	1	1	2	7	1	0	15	2	7	4	6	15
CONT	0	0	0	0	1	1	2	1	0	0	4	26	4	6	5	6
JO	0	0	0	0	0	2	0	3	1	0	3	1	43	7	4	13
SA	0	0	2	0	0	0	3	2	0	3	0	0	0	80	3	31
AT	0	1	0	0	0	3	3	2	2	0	2	2	1	15	276	20
EL	1	0	1	3	2	3	2	5	5	11	5	6	14	9	19	295

- Most frequent ones confuse less frequent ones
- Hard to distinguish semantically similar relations

Conclusion

- Discriminative framework for discourse analysis.
- Our parsing model:
  - ✓ Discriminative
  - ✓ Structure and label jointly
  - ✓ Sequential and hierarchical dependencies
  - ✓ Supports an optimal parsing algorithm

- Our approach outperforms the state-of-the-art by a wide margin.

Future Work

- Extend to multi-sentential text.
- Can segmentation and parsing be done jointly?