# A Novel Discriminative Framework for Sentence-Level Discourse Analysis

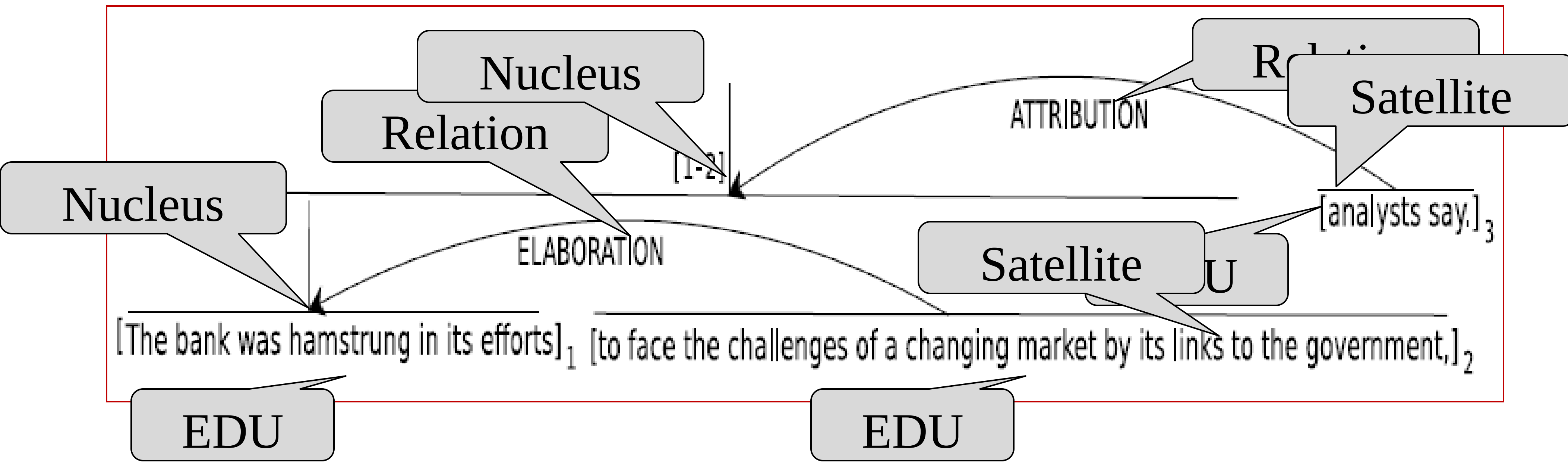
**Shafiq Joty**

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.*



# 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,*

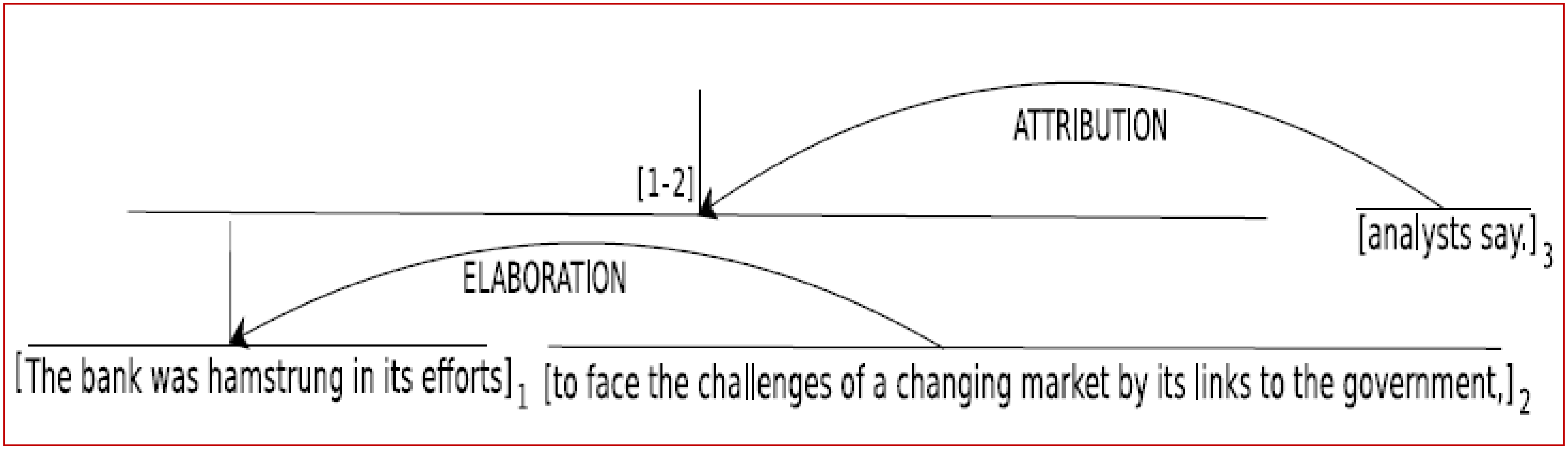
*analysts say.*

1

2

3

**Discourse Parsing**



# 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

Segmenter & Parser

Sentence level

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

## Hernault et al. (2010)

HILDA

Segmenter & Parser

Document level

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

Newspaper articles

# Previous Work (2)

**Subba & Di-Eugenio, (2009)**

Shift-reduce { Only Parser }

Sentence + Document level

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

Instructional manuals

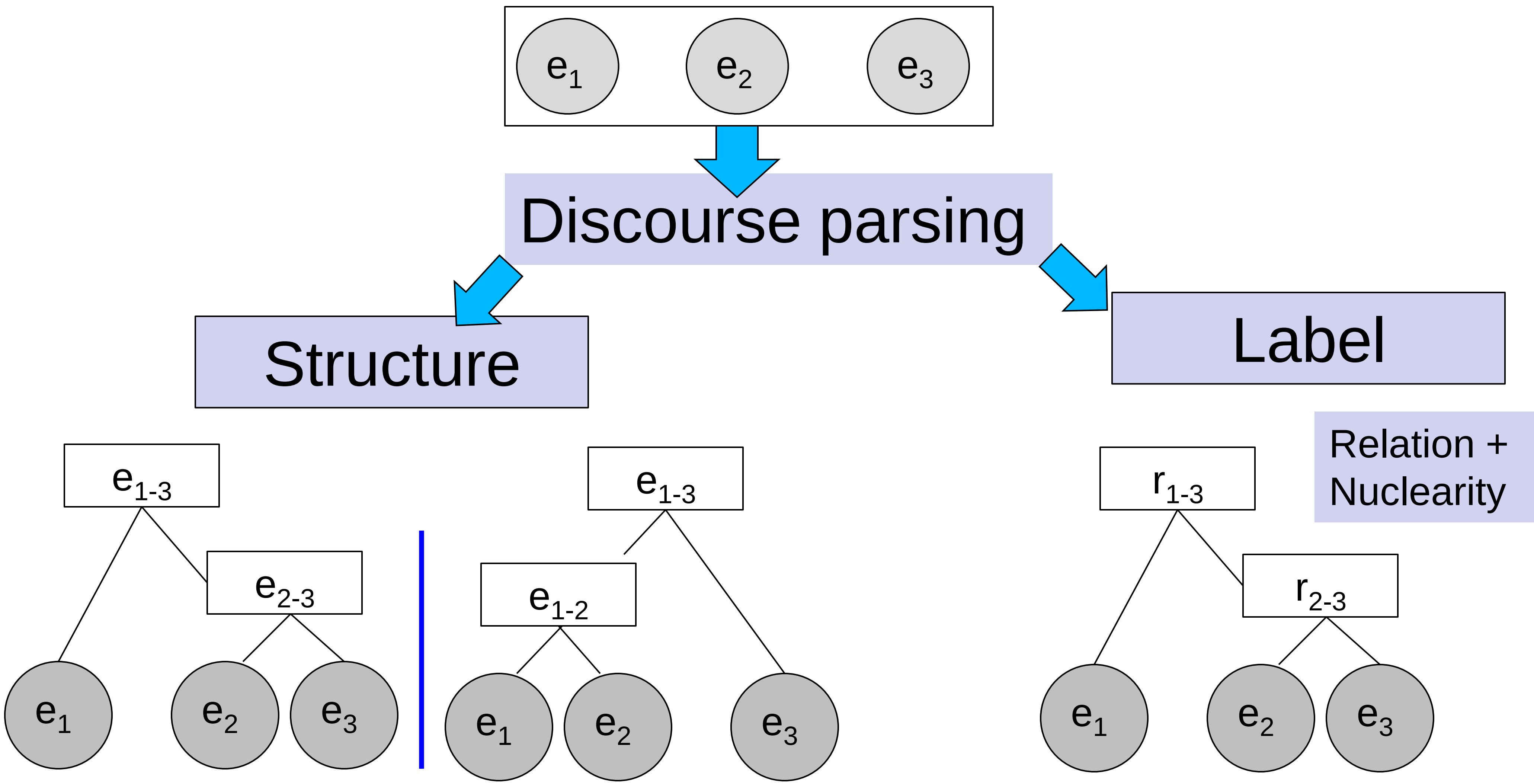
**Fisher & Roark (2007)**

Binary log-linear { Only Segmenter }

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

# Discourse Parsing

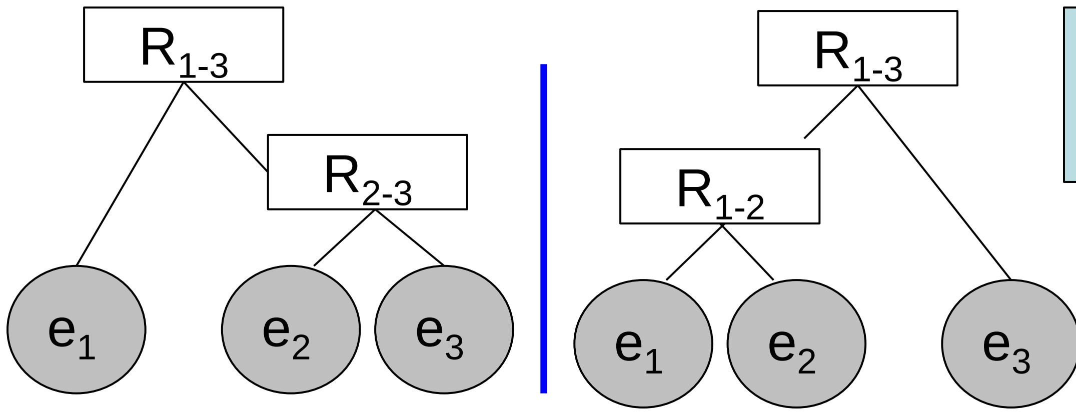
Assume a sentence is already segmented into EDUs.



# Our Discourse Parser

## Parsing model

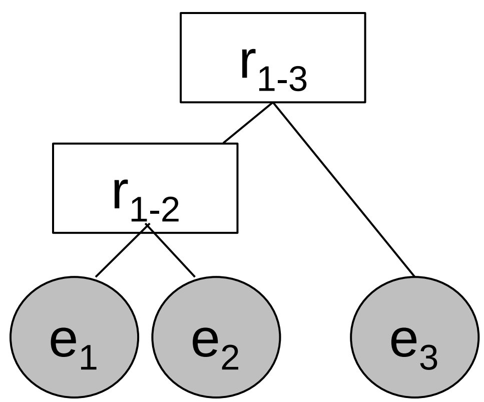
Assign probabilities to DTs.




R ranges over set of relations

## Parsing algorithm

Find the most probable DT

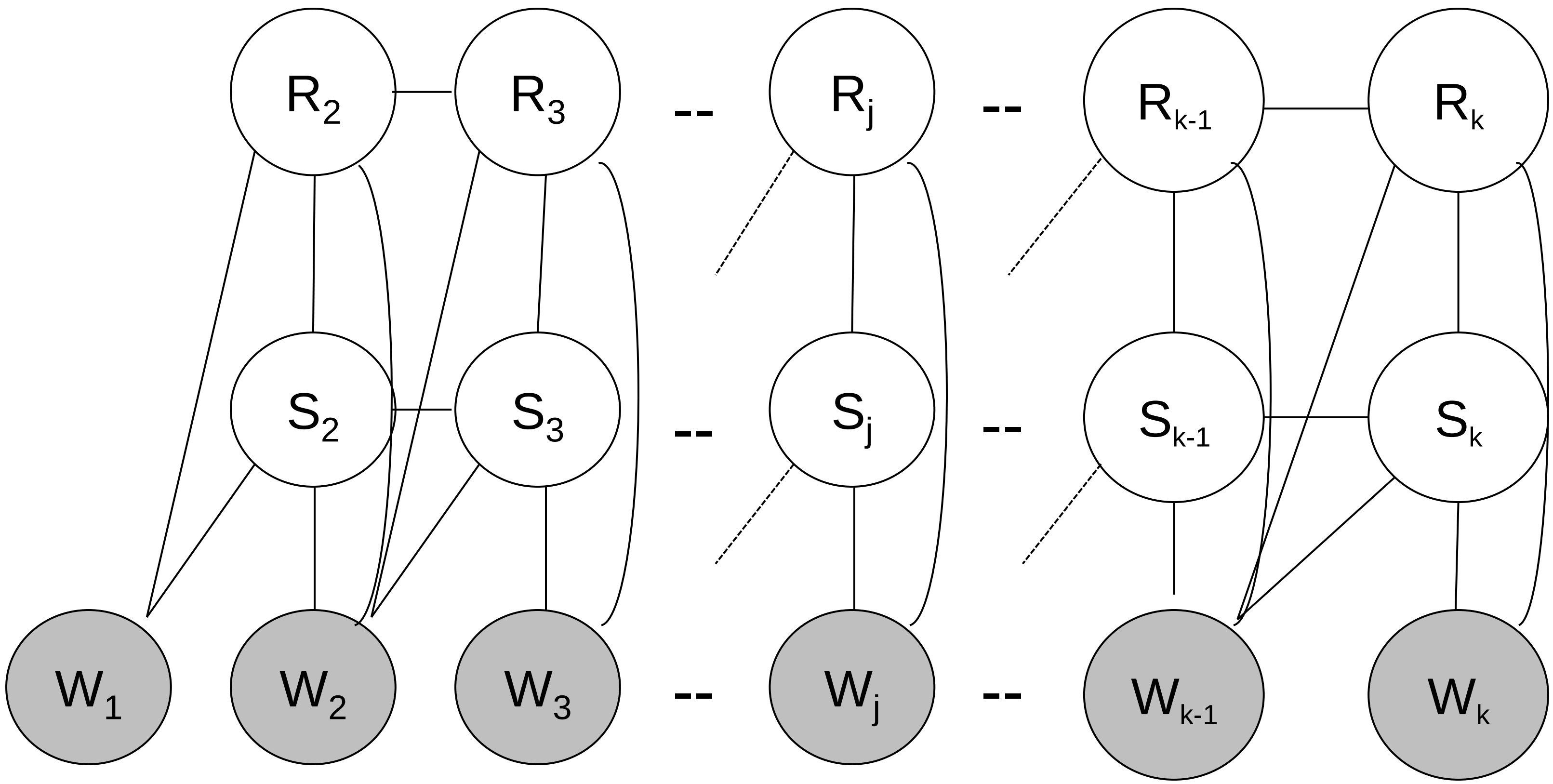


# Requirements for Our Parsing Model

- ✓ 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

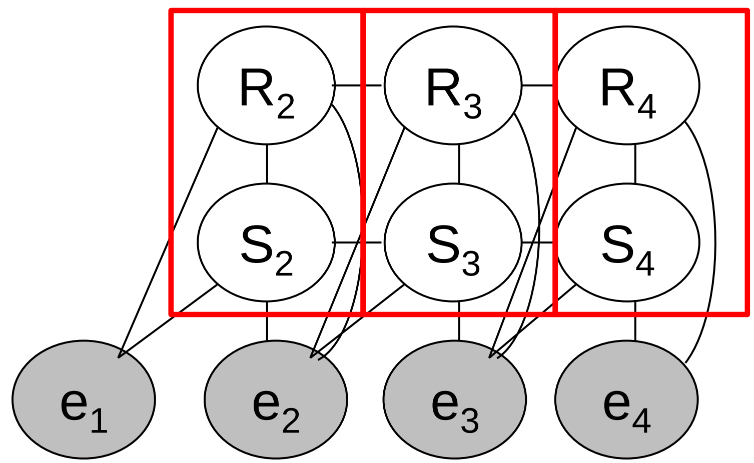
**Dynamic Conditional Random Field (DCRF) [Sutton et al, 2007]**

**Models sequential dependencies**

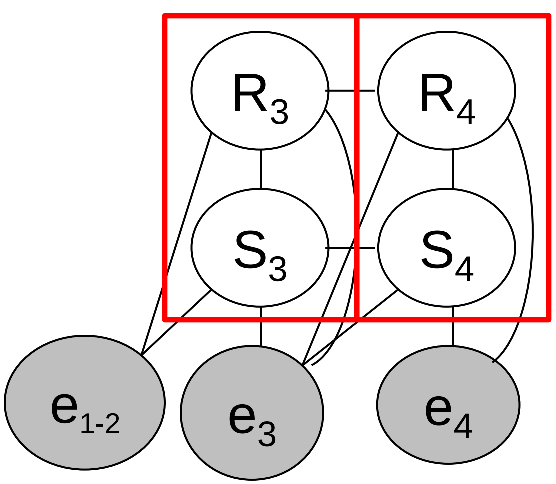
# Obtaining probabilities

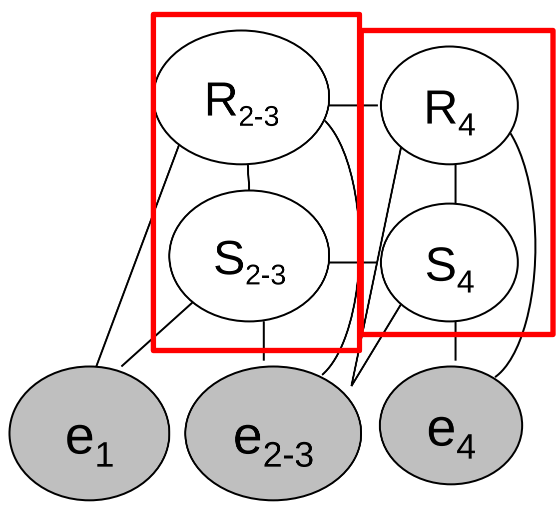
Apply DCRF recursively at different levels and compute posterior marginals of relation-structure pairs

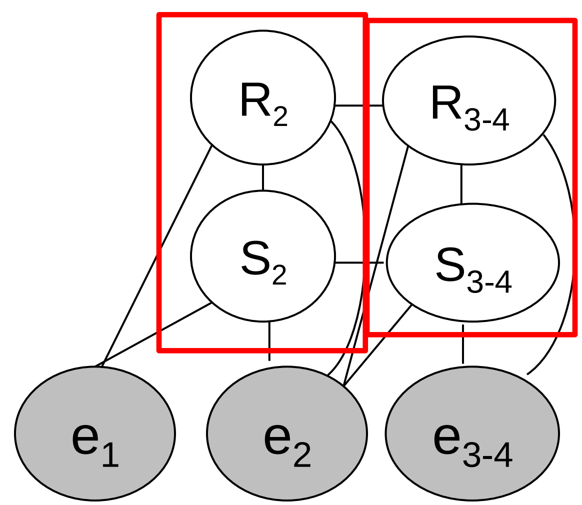
Level 1




Level 2



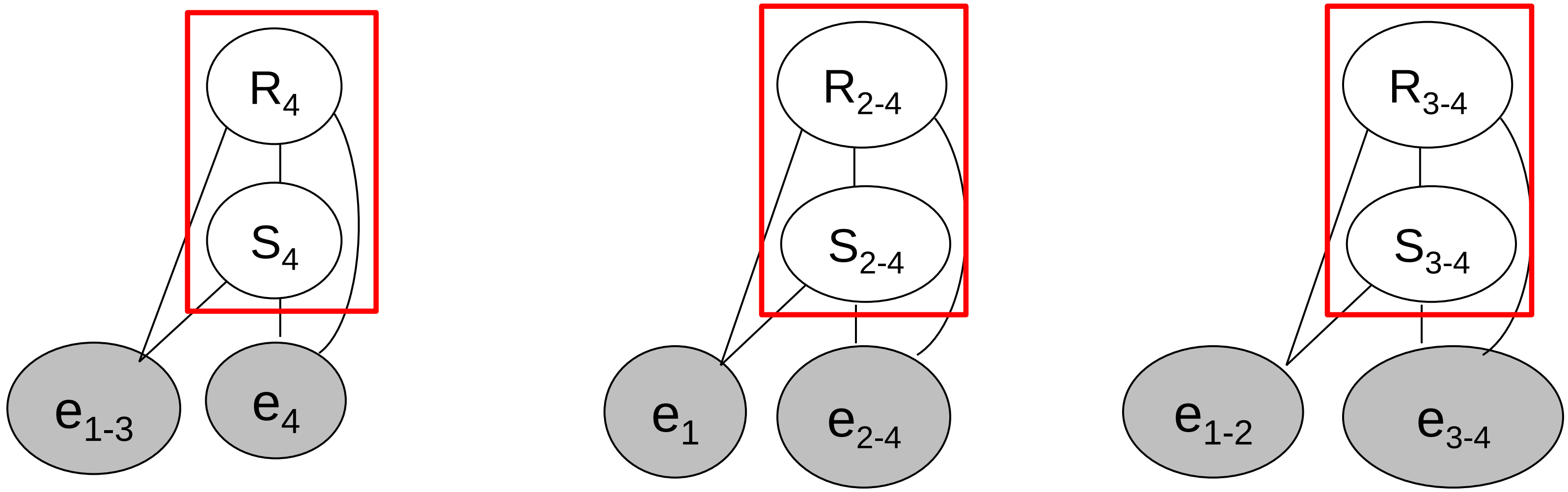



# Obtaining probabilities

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

Level 3



# Features Used in Parsing Model

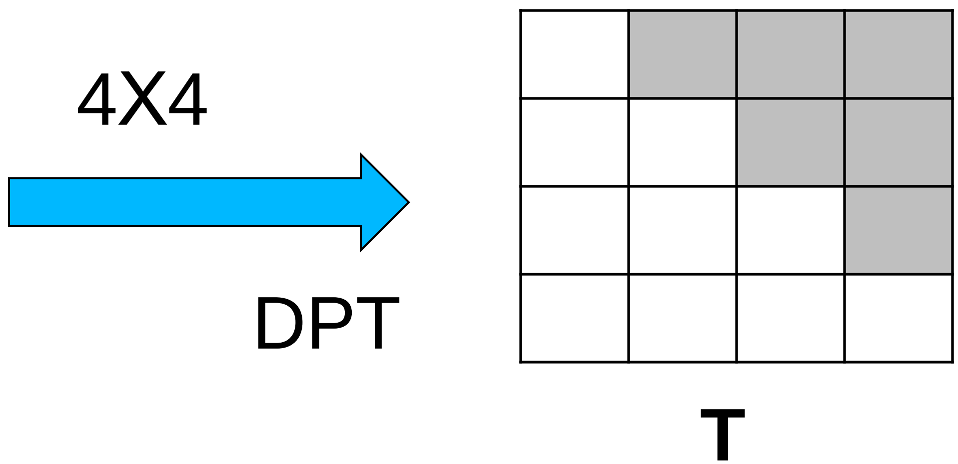
| |
|---|
| **8 organizational features** |
| Relative number of EDUs in *span 1* and *span 2*. |
| Relative number of tokens in *span 1* and *span 2*. |
| Distances of span 1 in EDUs to the *beginning* and to the *end*. |
| Distances of span 2 in EDUs to the *beginning* and to the *end*. |
| **8 N-gram 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




$$T(i, j) = P(R[i, m, j])$$

$$m = \text{argmax}_{i \leq k \leq j} P(R[i, k, j])$$

R ranges over set of relations

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)**

- 385 news articles

  -Train: 347 (7673 sentences)

  -Test: 38 (991 sentences)

**Relations**

- 18 relations
- 39 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
- 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?