Distributed Semantics for Automatically Classifying Discourse Relations

Jacob Eisenstein

Georgia Institute of Technology

May 26, 2016

Predicting implicit discourse relations

(1) The more people you love, the weaker you are.
(?) You'll do things for them that you know you shouldn't do.
(?) You'll act the fool to make them happy, to keep them safe.
(?) Love no one but your children.
(?) On that front, a mother has no choice.

Predicting implicit discourse relations

(1) The more people you love, the weaker you are.
(For example,) You'll do things for them that you know you shouldn't do.
(In addition,) You'll act the fool to make them happy, to keep them safe.
(Therefore,) Love no one but your children.
On that front (AltLex), a mother has no choice.

Predicting implicit discourse relations

(1) The more people you love, the weaker you are.
(Expansion) You'll do things for them that you know you shouldn't do.
(Expansion) You'll act the fool to make them happy, to keep them safe.
(Contingency) Love no one but your children.
[Contingency] a mother has no choice.

Predicting implicit discourse relations

(1) The more people you love, the weaker you are.
(Expansion) You'll do things for them that you know you shouldn't do.
(Expansion) You'll act the fool to make them happy, to keep them safe.
(Contingency) Love no one but your children.
[Contingency] a mother has no choice.

Application: summarization

(2) The more people you love, the weaker you are.
(EXPANSION) You'll do things for them that you know you shouldn't do.
(EXPANSION) You'll act the fool to make them happy, to keep them safe.
(CONTINGENCY) Love no one but your children.
(CONTINGENCY) On that front, a mother has no choice.

Discourse structure guides the selection of extracts for summaries (Marcu, 1999; Louis et al., 2010; Hirao et al., 2013).

Application: summarization

(2) **The more people you love, the weaker you are.**
(EXPANSION) You'll do things for them that you know you shouldn't do.
(EXPANSION) You'll act the fool to make them happy, to keep them safe.
(CONTINGENCY) **Love no one but your children.**
(CONTINGENCY) On that front, a mother has no choice.

Discourse structure guides the selection of extracts for summaries (Marcu, 1999; Louis et al., 2010; Hirao et al., 2013).

Application: document classification

(3) The federal budget should be an honest blueprint for the spending priorities of the government.
However, this budget is dishonest.

Contrast relations can reverse the scope of sentiment polarity (Somasundaran et al., 2009; Yang & Cardie, 2014; Bhatia et al., 2015).

Why is predicting discourse relations hard?

Many discourse relations are fundamentally semantic (Hobbs, 1979):

(4) Love no one but your children.
On that front, a mother has no choice.

Typical solution (Lin et al., 2009; Rutherford & Xue, 2014) is bilexical features, e.g.,

$$\langle \text{love, choice} \rangle, \langle \text{children, mother} \rangle, \langle \text{no, no} \rangle, \ldots$$

But bilexical features are sparse and noisy, and discourse-annotated datasets are inherently small.

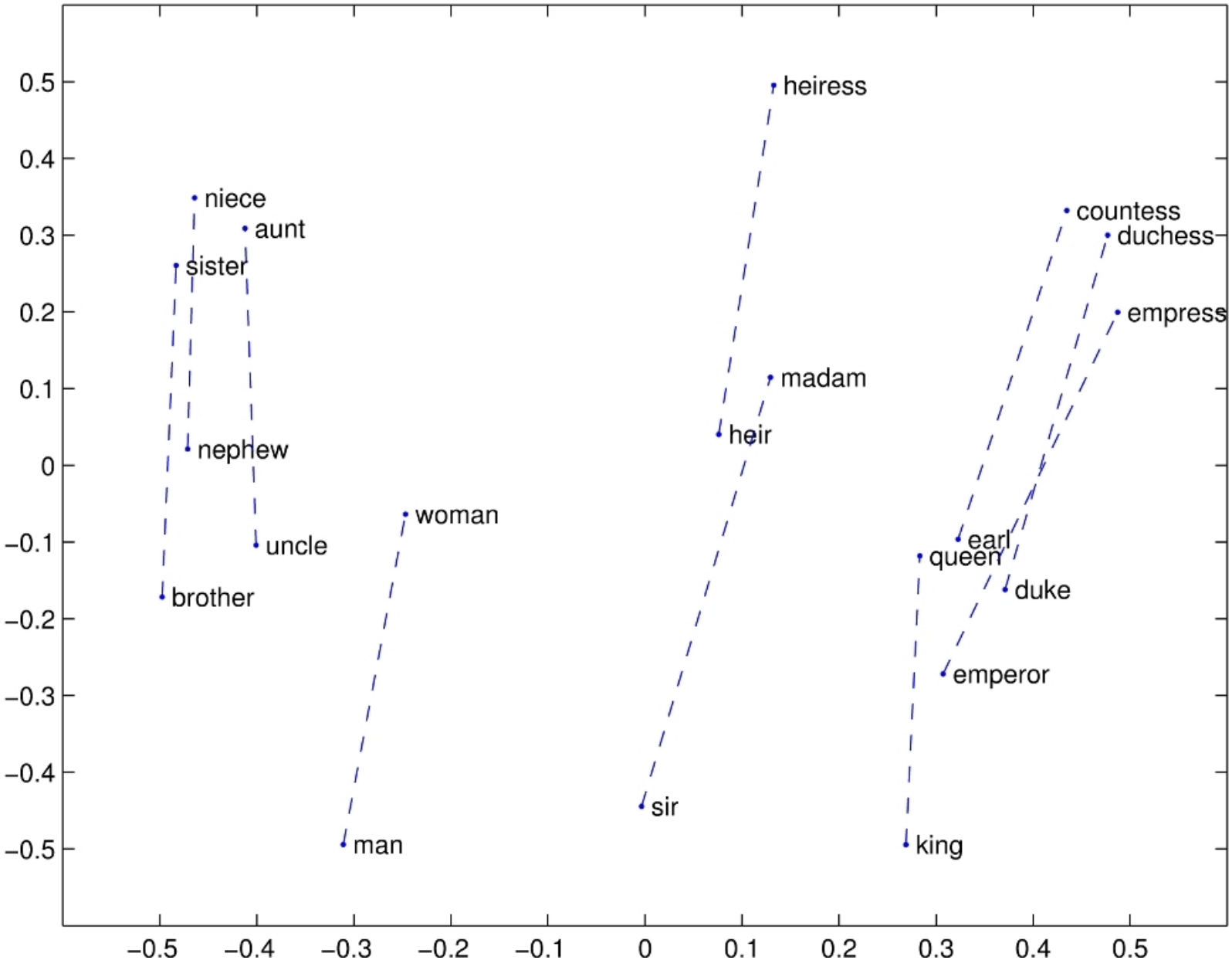
Can distributed semantics help?

Distributed semantics proposes to capture meaning in dense numerical vectors.

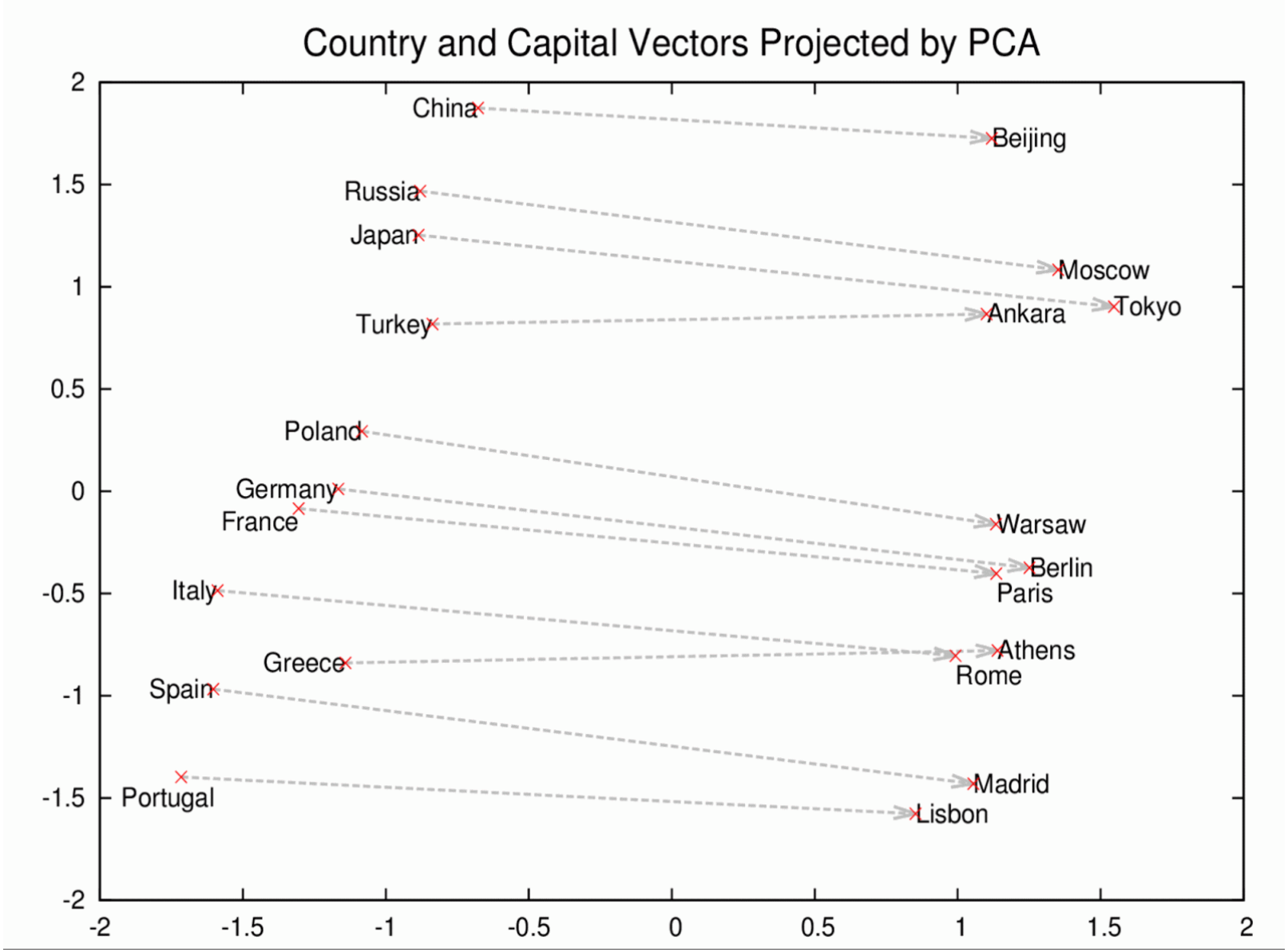
$$\begin{aligned}
\boldsymbol{u}_{\text{easy}} &= [0.1, -0.5, -0.4, \ldots] \\
\boldsymbol{u}_{\text{short}} &= [-0.6, 0.5, -1.0, \ldots] \\
\boldsymbol{u}_{\text{visit}} &= [-0.7, 1.0, 0.6, \ldots] \\
\boldsymbol{u}_{\text{distance}} &= [1.7, 1.9, -1.5, \ldots]
\end{aligned}$$

Basic idea of WORD2VEC *et al* is to induce word representations that predict distributional statistics (Mikolov et al., 2013; Levy & Goldberg, 2014).

Pretty picture 1

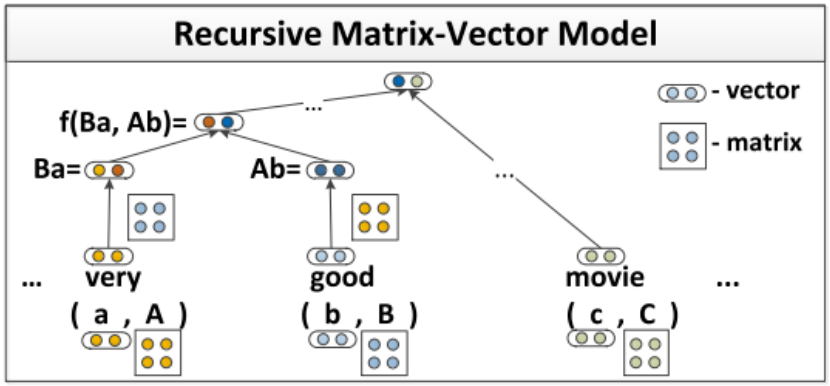


Pretty picture 2



Distributed semantics beyond the lexicon?

Can we build distributed representations of multi-word linguistic units?



Vector-semantic **composition** has been applied to:

- phrases (Baroni & Zamparelli, 2010; Mikolov et al., 2013);
- sentences (Socher et al., 2012, 2013);
- paragraphs and beyond (Kalchbrenner & Blunsom, 2013; Le & Mikolov, 2014).

Distributed semantics for discourse

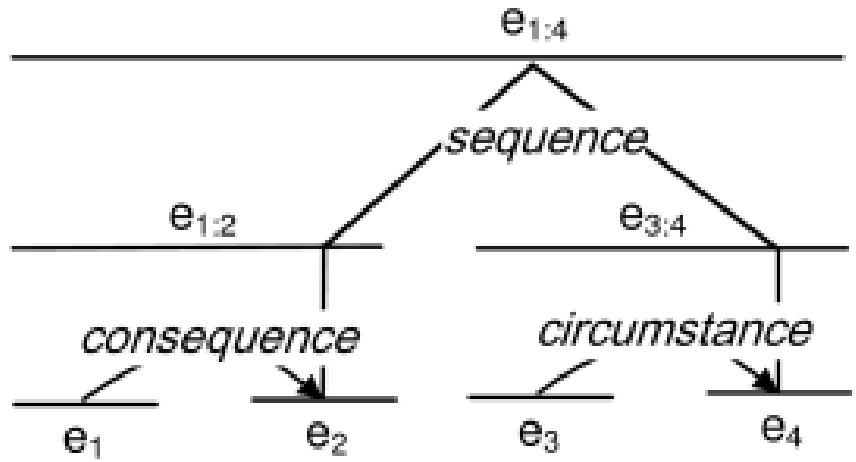
Key questions:

- What should distributed representations of discourse units look like?
- How should we learn them?
- How to apply distributed representations to discourse relation detection and parsing?

Project 1: RST Parsing

"Representation Learning for Text-level Discourse Parsing" (Ji & Eisenstein, 2014)

- **Goal**: rhetorical structure theory parsing

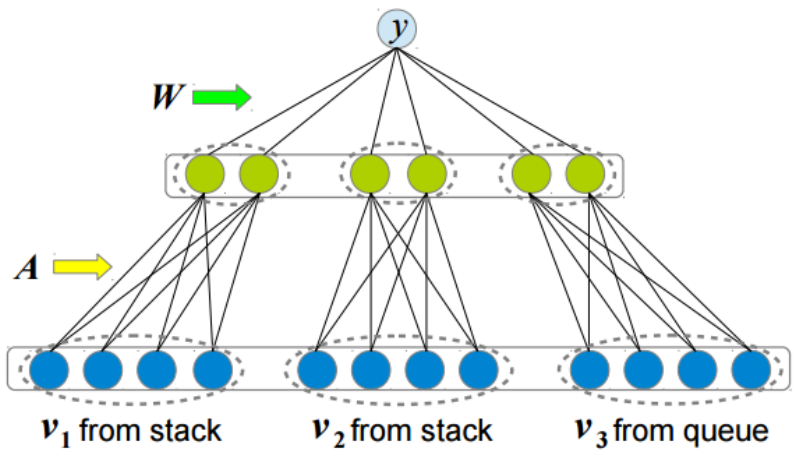


- **Algorithm**: shift-reduce (Marcu, 1996; Sagae, Sagae) with an SVM classifier.

Shift-reduce parsing for RST

At each point, the parser can:

- **shift** the next elementary discourse unit onto the stack;
- **reduce** the top two elements on the stack into a discourse relation.



These shift/reduce decisions are driven by a classifier, with access to the **distributed representation** of each discourse unit.

Building the distributed representations

- **Elementary discourse units**:

$$\boldsymbol{u}(\text{Love no one but your children}) = \boldsymbol{u}_{\text{love}} + \boldsymbol{u}_{\text{no}} + \dots$$

"Averaging pooling" of word representations (Blacoe & Lapata, 2012)

Building the distributed representations

- **Elementary discourse units**:

$$\boldsymbol{u}(\text{Love no one but your children}) = \boldsymbol{u}_{\text{love}} + \boldsymbol{u}_{\text{no}} + \dots$$

"Averaging pooling" of word representations (Blacoe & Lapata, 2012)

- **Higher-order discourse units** inherit the distributed representation of their nucelus (strong compositionality criterion).
- See Li et al. (2014) for more sophisticated composition via recursive neural networks.

RST Results

	Span	Nuclearity	Relation
Annotator agreement	88.7	77.7	65.8

RST Results

	Span	Nuclearity	Relation
Annotator agreement	88.7	77.7	65.8
HILDA (Hernault, Prendinger, duVerle & Ishizuka, Hernault et al.)	83.0	68.4	54.8
TSP (Joty et al., 2013)	82.7	68.4	55.7
"Basic features"	79.4	68.0	53.0

RST Results

	Span	Nuclearity	Relation
Annotator agreement	88.7	77.7	65.8
HILDA (Hernault, Prendinger, duVerle & Ishizuka, Hernault et al.)	83.0	68.4	54.8
TSP (Joty et al., 2013)	82.7	68.4	55.7
"Basic features"	79.4	68.0	53.0
Distributed			
Collobert & Weston	75.3	67.1	53.8
Non-neg. matrix factorization	78.6	67.7	54.8

Supervised distributed semantics

- Pre-trained word embeddings are no better than surface features.
- Let's learn the word representations jointly with the parser!
- Basically, a hidden-variable support vector machine. Iterate:
  1. solve SVM dual objective
  2. perform gradient update to word representations

RST Results

	Span	Nuclearity	Relation
Annotator agreement	88.7	77.7	65.8
HILDA (Hernault, Prendinger, duVerle & Ishizuka, Hernault et al.)	83.0	68.4	54.8
TSP (Joty et al., 2013)	82.7	68.4	55.7
"Basic features"	79.4	68.0	53.0
Distributed			
Collobert & Weston	75.3	67.1	53.8
Non-neg. matrix factorization	78.6	67.7	54.8

RST Results

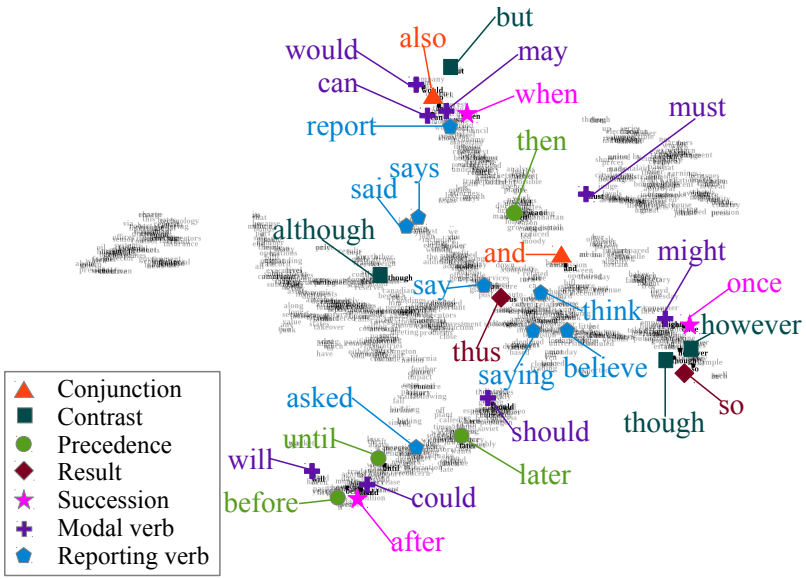
	Span	Nuclearity	Relation
Annotator agreement	88.7	77.7	65.8
HILDA (Hernault, Prendinger, duVerle & Ishizuka, Hernault et al.)	83.0	68.4	54.8
TSP (Joty et al., 2013)	82.7	68.4	55.7
"Basic features"	79.4	68.0	53.0
Distributed			
Collobert & Weston	75.3	67.1	53.8
Non-neg. matrix factorization	78.6	67.7	54.8
Distributed	80.9	69.4	59.0

RST Results

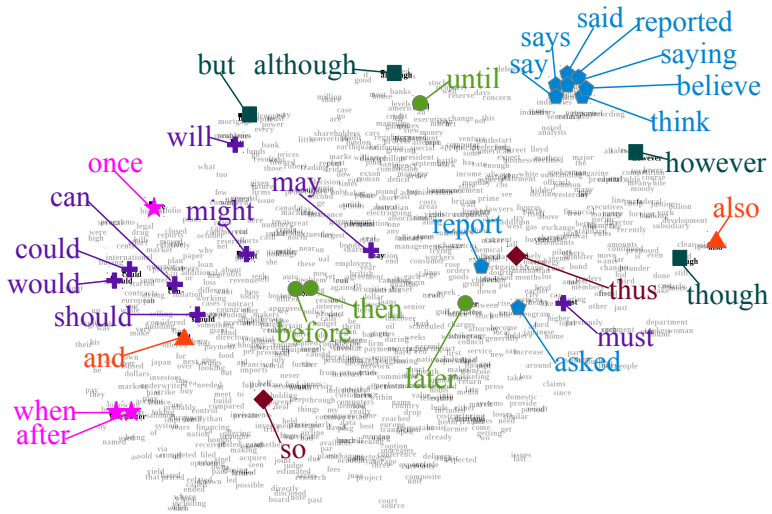
	Span	Nuclearity	Relation
Annotator agreement	88.7	77.7	65.8
HILDA (Hernault, Prendinger, duVerle & Ishizuka, Hernault et al.)	**83.0**	68.4	54.8
TSP (Joty et al., 2013)	82.7	68.4	55.7
"Basic features"	79.4	68.0	53.0
Distributed			
Collobert & Weston	75.3	67.1	53.8
Non-neg. matrix factorization	78.6	67.7	54.8
Distributed	80.9	69.4	59.0
+basic features	82.1	**71.1**	**61.6**

On discourse relations, the distributed representation cuts the gap between SOTA and inter-annotator agreement by 60%!

Representation learned



NMF, $K = 20$



Representation learning, $K = 20$

Project 2: PDTB Implicit Relations

"One Vector is not Enough: Entity-Augmented Distributed Semantics for Discourse Relations" (Ji & Eisenstein, 2015)

- **Goal**: PDTB implicit relation classification
- **Prior work**: augment bilexical features with Brown cluster features (Rutherford & Xue, 2014; Wang & Lan, 2015).

$$\langle \text{fun}, \text{buildings} \rangle, \langle \text{place}, \text{interesting} \rangle, \ldots$$

Project 2: PDTB Implicit Relations

"One Vector is not Enough: Entity-Augmented Distributed Semantics for Discourse Relations" (Ji & Eisenstein, 2015)

- **Goal**: PDTB implicit relation classification
- **Prior work**: augment bilexical features with Brown cluster features (Rutherford & Xue, 2014; Wang & Lan, 2015).

$$\langle 0010, 1011 \rangle, \langle 1010, 0001 \rangle, \ldots$$

Project 2: PDTB Implicit Relations

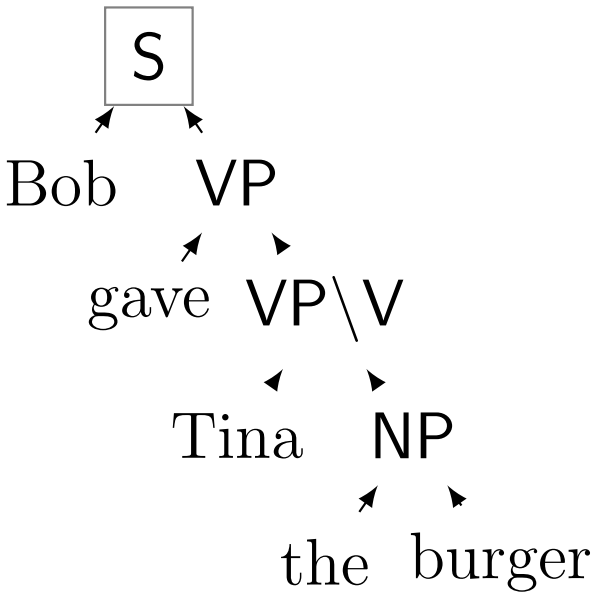
"One Vector is not Enough: Entity-Augmented Distributed Semantics for Discourse Relations" (Ji & Eisenstein, 2015)

- **Goal**: PDTB implicit relation classification
- **Prior work**: augment bilexical features with Brown cluster features (Rutherford & Xue, 2014; Wang & Lan, 2015).

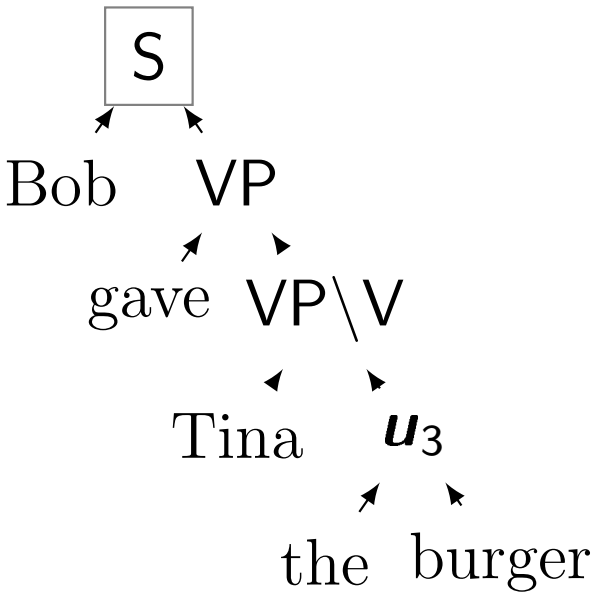
$$\langle 0010, 1011 \rangle, \langle 1010, 0001 \rangle, \ldots$$

- **Our approach**: construct meaning of discourse units through composition over the parse tree.

Vector-semantic composition

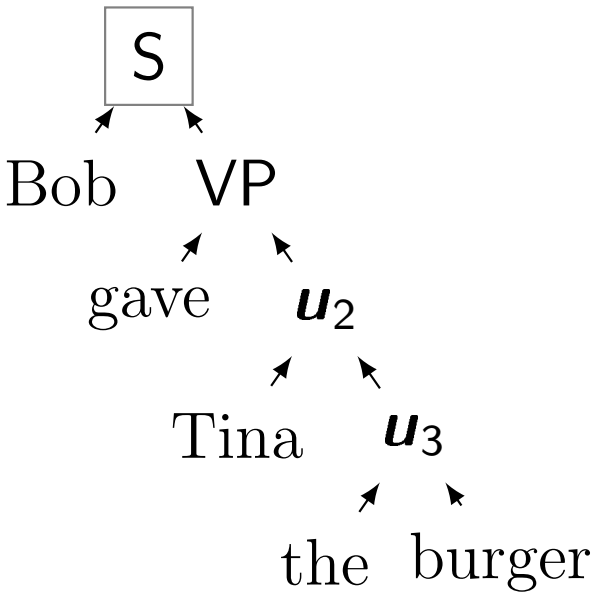


Vector-semantic composition



$$u_3 = \tanh\left(\mathbf{U}\left[u_{\text{the}}^\top \; u_{\text{burger}}^\top\right]^\top\right)$$

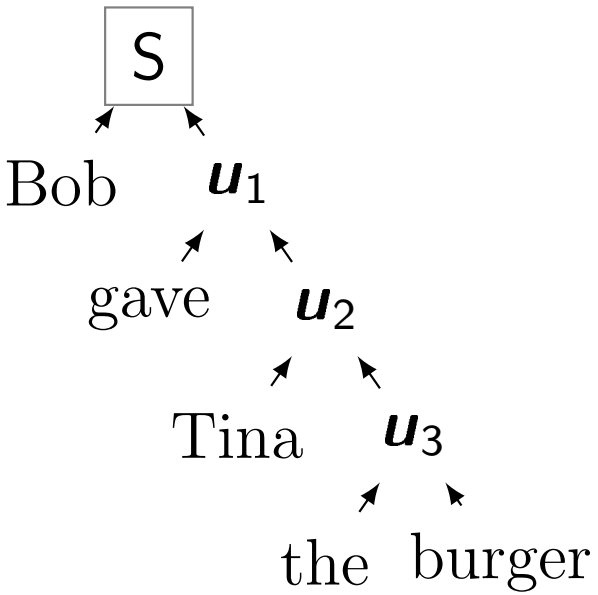
Vector-semantic composition



$$u_3 = \tanh\left(\mathbf{U}\left[u_{\text{the}}^\top\ u_{\text{burger}}^\top\right]^\top\right)$$

$$u_2 = \tanh\left(\mathbf{U}\left[u_{\text{Tina}}^\top\ u_3^\top\right]^\top\right)$$

Vector-semantic composition

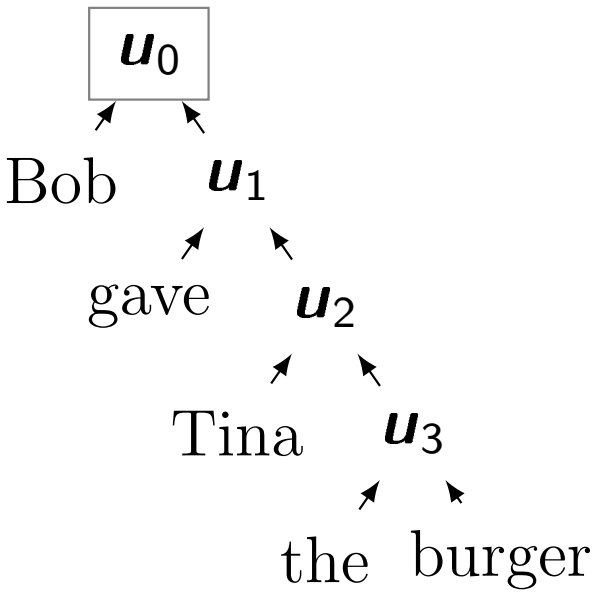


$$u_3 = \tanh\left(\mathbf{U}\left[u_{\text{the}}^\top\ u_{\text{burger}}^\top\right]^\top\right)$$

$$u_2 = \tanh\left(\mathbf{U}\left[u_{\text{Tina}}^\top\ u_3^\top\right]^\top\right)$$

$$u_1 = \tanh\left(\mathbf{U}\left[u_{\text{gave}}^\top\ u_2^\top\right]^\top\right)$$

Vector-semantic composition



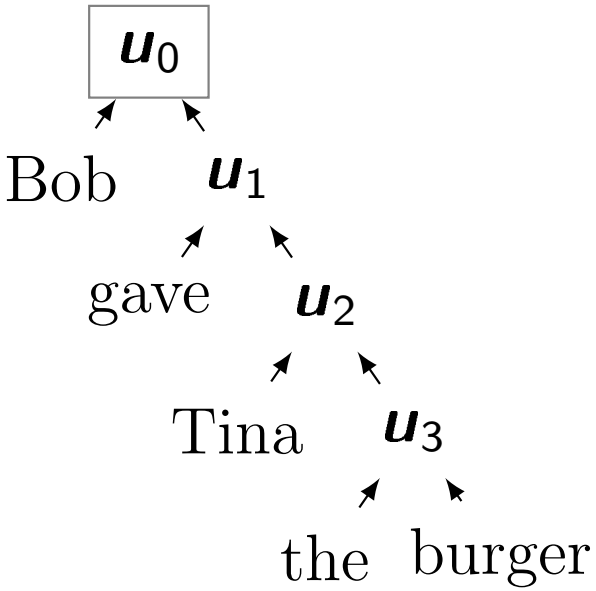
$$u_3 = \tanh\left(\mathbf{U}\left[u_{\text{the}}^\top\ u_{\text{burger}}^\top\right]^\top\right)$$

$$u_2 = \tanh\left(\mathbf{U}\left[u_{\text{Tina}}^\top\ u_3^\top\right]^\top\right)$$

$$u_1 = \tanh\left(\mathbf{U}\left[u_{\text{gave}}^\top\ u_2^\top\right]^\top\right)$$

$$u_0 = \tanh\left(\mathbf{U}\left[u_{\text{Bob}}^\top\ u_1^\top\right]^\top\right)$$

Vector-semantic composition



$$u_3 = \tanh\left(\mathbf{U}\left[u_{\text{the}}^\top\ u_{\text{burger}}^\top\right]^\top\right)$$
$$u_2 = \tanh\left(\mathbf{U}\left[u_{\text{Tina}}^\top\ u_3^\top\right]^\top\right)$$
$$u_1 = \tanh\left(\mathbf{U}\left[u_{\text{gave}}^\top\ u_2^\top\right]^\top\right)$$
$$u_0 = \tanh\left(\mathbf{U}\left[u_{\text{Bob}}^\top\ u_1^\top\right]^\top\right)$$

- DISCO2: **Dis**tributional **co**mpositional semantics for **disco**urse.
- Same architecure as Socher, Huang, Pennington, Ng & Manning (Socher et al.).

A bilinear model

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} \quad (\boldsymbol{u}^{(\ell)})^\top \mathbf{A}_y \boldsymbol{u}^{(r)} + b_y$$

- $\boldsymbol{u}^{(\ell)}$ is the representation of the left argument
- $\boldsymbol{u}^{(r)}$ is the representation of the right argument
- In practice, we set

$$\mathbf{A}_y = \boldsymbol{a}_{y,1} \boldsymbol{a}_{y,2}^\top + \text{diag}(\boldsymbol{a}_{y,3}).$$

Learning

- Word representations are fixed to WORD2VEC. Fine-tuning $\rightarrow$ bad overfitting in this model.
- We learn $\mathbf{U}, \mathbf{A}, b$ by backpropagating from a hinge loss on relation classification. (Second-level PDTB relations)

PDTB Results

Most common class	26.0
Additive word representations	**28.7**

PDTB Results

Most common class	26.0
Additive word representations	28.7
Lin et al. (2009)	40.2
SFM: Our reimplementation of Lin et al. (2009)	39.7
SFMB: Lin et al. (2009) + Brown clusters	**40.7**

PDTB Results

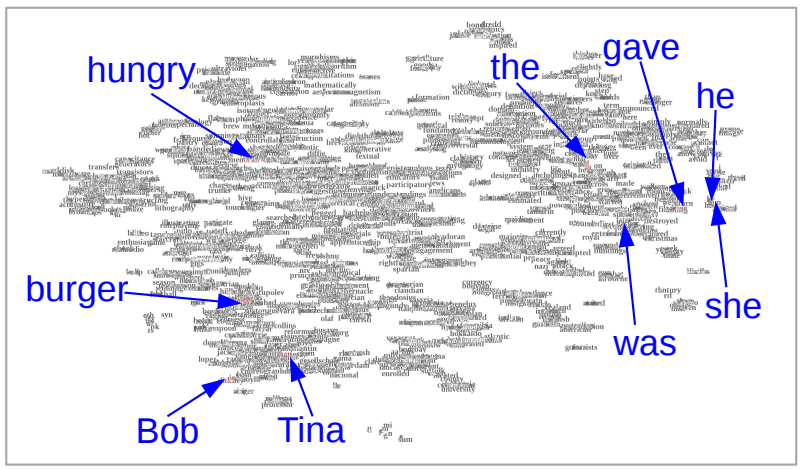
Most common class	26.0
Additive word representations	28.7
Lin et al. (2009)	40.2
SFM: Our reimplementation of Lin et al. (2009)	39.7
SFMB: Lin et al. (2009) + Brown clusters	40.7
Disco2	37.0
Disco2 + SFMB	**43.8**

Are we done?

- Bob gave Tina the burger.
- **She** was hungry.

- Bob gave Tina the burger.
- **He** was hungry.

The discourse relations are completely different.
The distributed representations are nearly identical.



One vector is not enough.

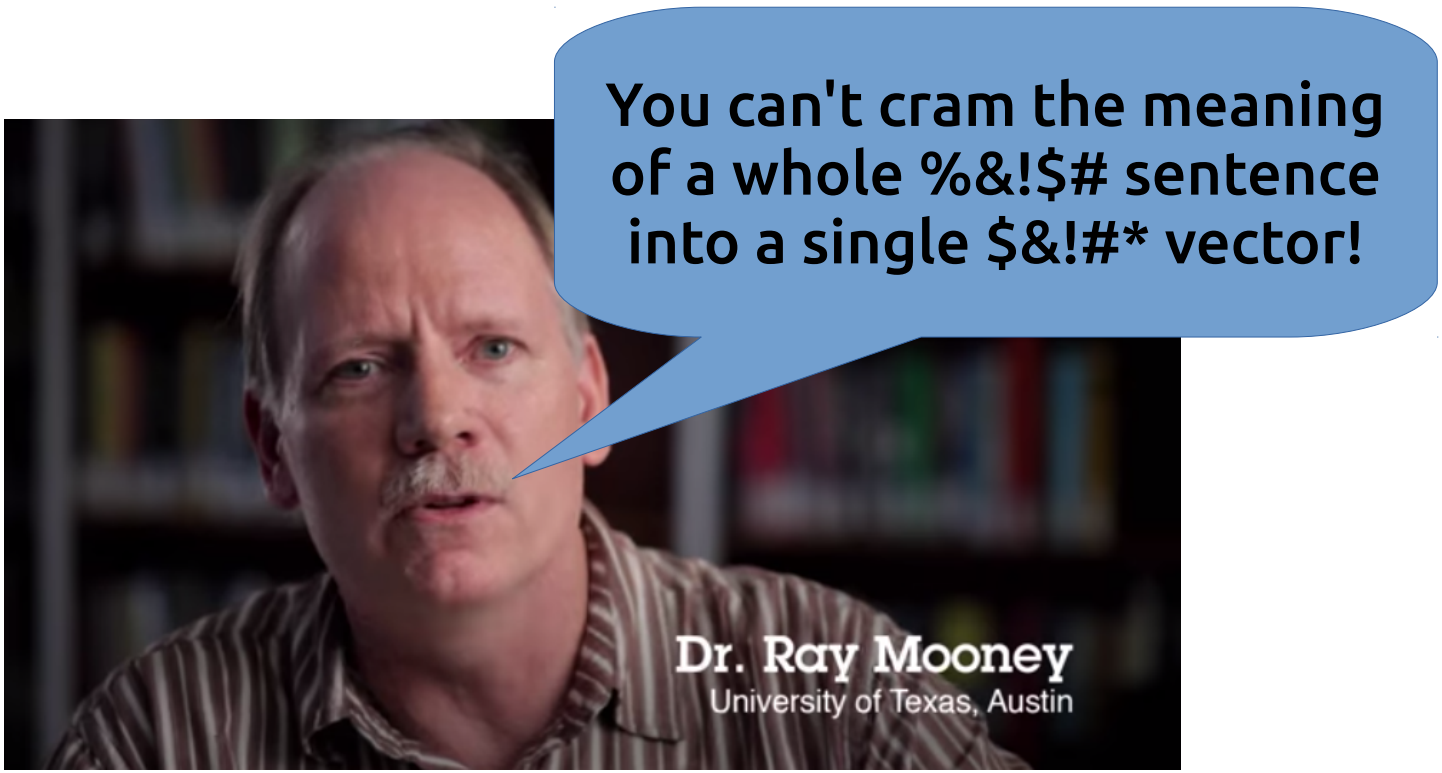
If we insist on representing each discourse argument as a single vector, we lose the ability to track references across the discourse.

Or to put it another way...

One vector is not enough.

If we insist on representing each discourse argument as a single vector, we lose the ability to track references across the discourse.

Or to put it another way...



Entity-augmented distributed semantics

Look at things from Tina's perspective:

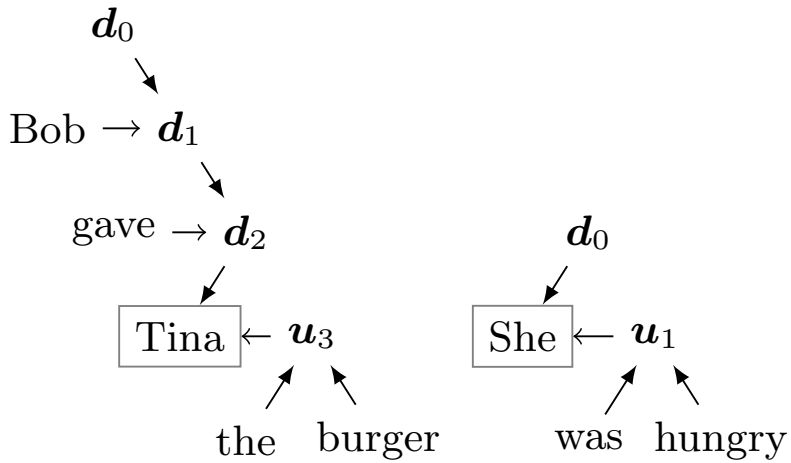
- $s1$: She got the burger from Bob
- $s2$: She was hungry

Let's represent these Tina-centric meanings with more vectors!

The downward pass

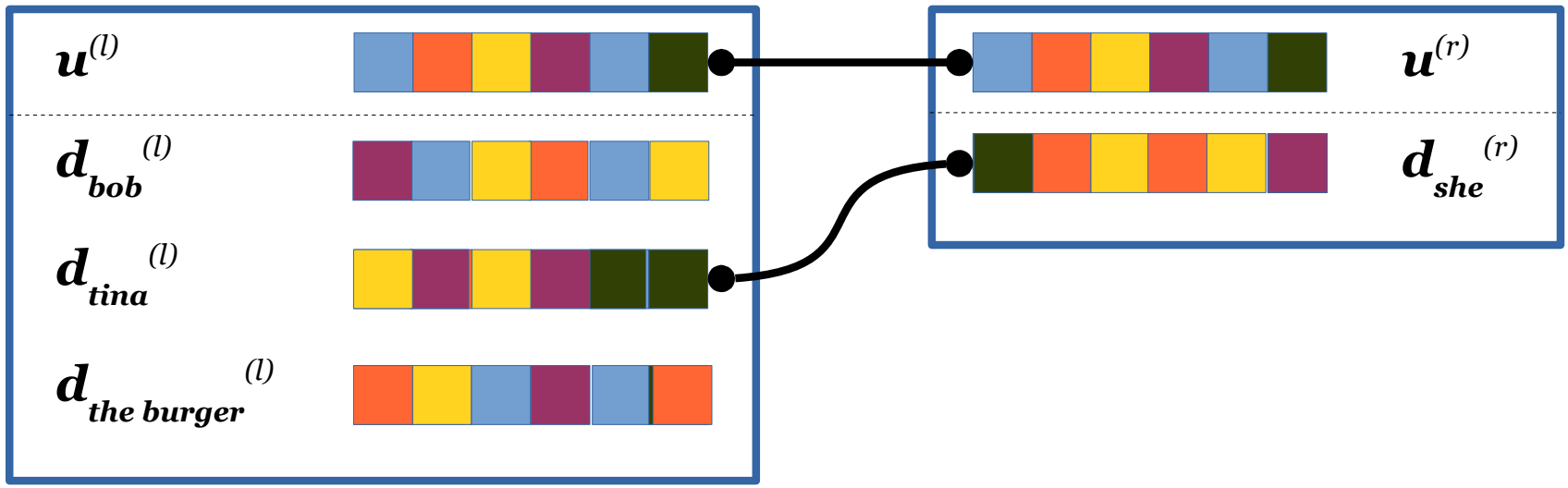
A **downward pass** computes a downward vector for each node in the parse.

$$\boldsymbol{d}_i = \tanh\left(\mathbf{V}\begin{bmatrix}\boldsymbol{d}_{\rho(i)} \\ \boldsymbol{u}_{s(i)}\end{bmatrix}\right)$$



This computation preserves the feedforward architecture.

A new bilinear model



$$\hat{y} = \arg\max_{y \in \mathcal{Y}} (\boldsymbol{u}^{(\ell)})^\top \mathbf{A}_y \boldsymbol{u}^{(r)} + \sum_{\langle i,j \rangle \in \mathcal{A}} (\boldsymbol{d}_i^{(\ell)})^\top \mathbf{B}_y \boldsymbol{d}_j^{(r)} + b_y$$

We now sum over coreferent mention pairs $\langle i, j \rangle \in \mathcal{A}$, obtained from the Berkeley coreference system.

PDTB Results

Most common class	26.0
Additive word representations	28.7
Lin et al. (2009)	40.2
SFM: Our reimplementation of Lin et al. (2009)	39.7
SFMB: Lin et al. (2009) + Brown clusters	40.7
Disco2	37.0
Disco2 + SFMB	**43.8**

PDTB Results

Most common class	26.0
Additive word representations	28.7
Lin et al. (2009)	40.2
SFM: Our reimplementation of Lin et al. (2009)	39.7
SFMB: Lin et al. (2009) + Brown clusters	40.7
Disco2	37.0
Disco2 + SFMB	43.8
Disco2 + SFMB + entity semantics	**44.6**

PDTB Results

Most common class	26.0
Additive word representations	28.7
Lin et al. (2009)	40.2
SFM: Our reimplementation of Lin et al. (2009)	39.7
SFMB: Lin et al. (2009) + Brown clusters	40.7
Disco2	37.0
Disco2 + SFMB	43.8
Disco2 + SFMB + entity semantics	**44.6**

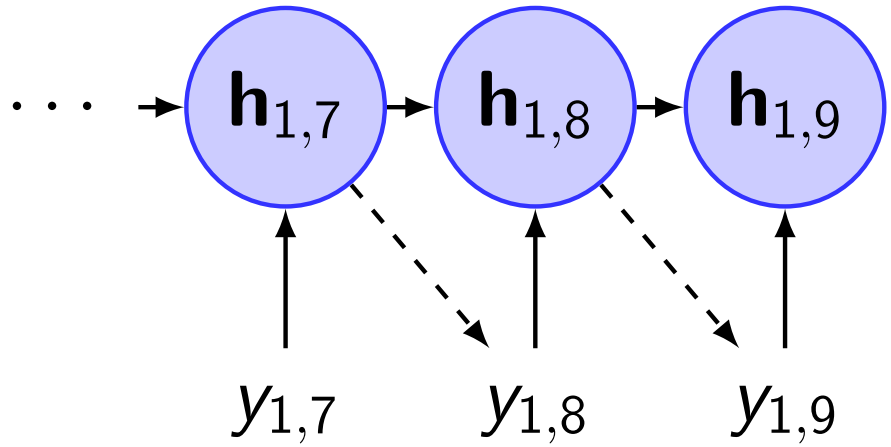
- Only 30% of PDTB relation pairs have coreferent mentions (according to Berkeley coref).
- On these examples, the improvement is 2.7%.

Project 3: A Generative Model

Benefits of **joint** probabilistic models of discourse and text:

- evaluate texts (summaries, translations, ...) in terms of discourse coherence;
- train from partial supervision;
- train from auxiliary supervision.

The Discourse Relation Language Model



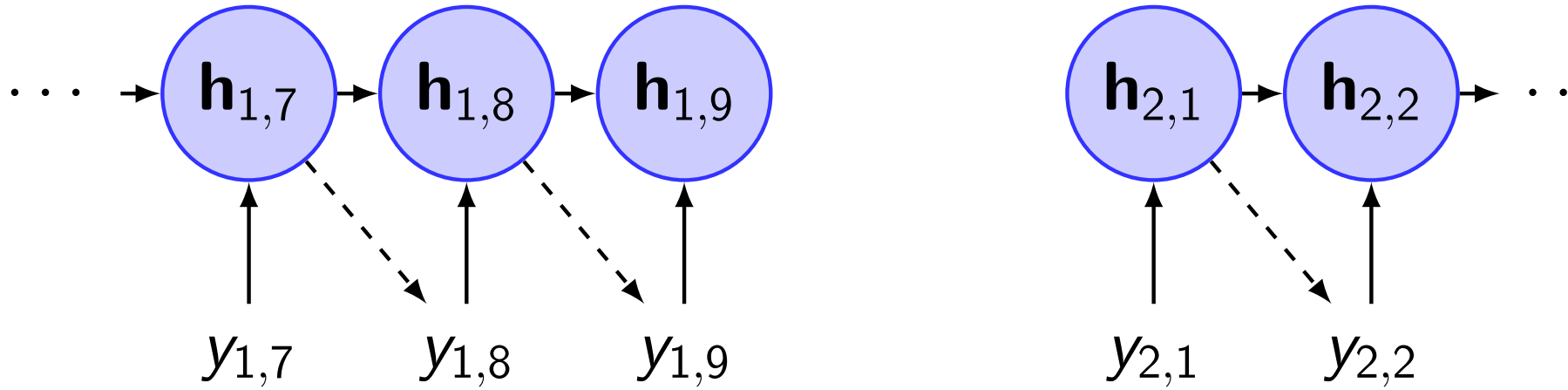
RNNLM (Mikolov et al 2010)

$$\mathbf{h}_{i,n} = f(E_{y_{i,n}} + U\mathbf{h}_{i,n-1}) \quad (1)$$

$$y_{i,n+1} \sim \text{SoftMax}(V\mathbf{h}_{i,n} \quad) \quad (2)$$

(3)

The Discourse Relation Language Model



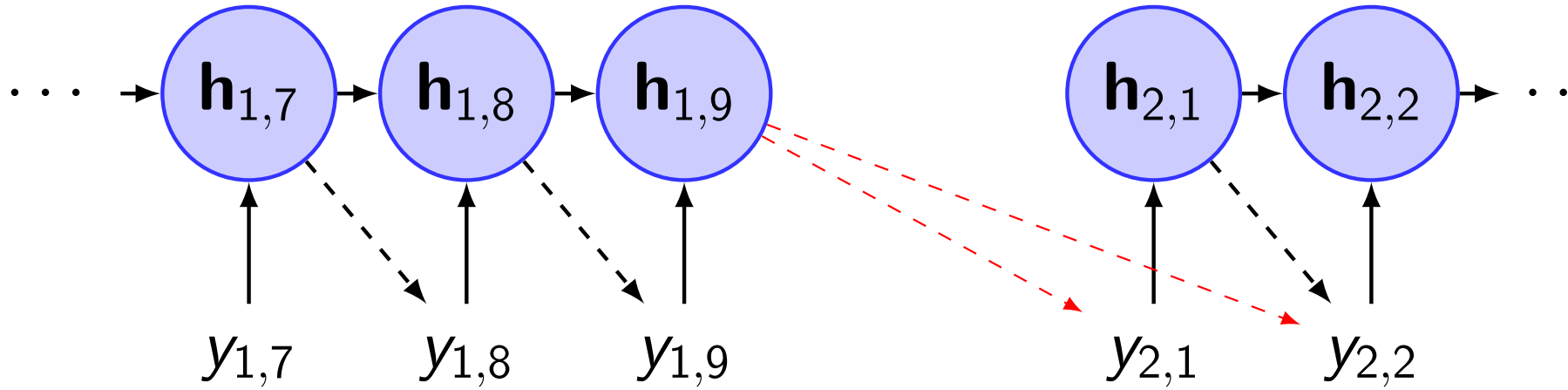
RNNLM (Mikolov et al 2010)

$$\mathbf{h}_{i,n} = f(E_{y_{i,n}} + U\mathbf{h}_{i,n-1}) \quad (1)$$

$$y_{i,n+1} \sim \text{SoftMax}(V\mathbf{h}_{i,n}) \quad (2)$$

(3)

The Discourse Relation Language Model



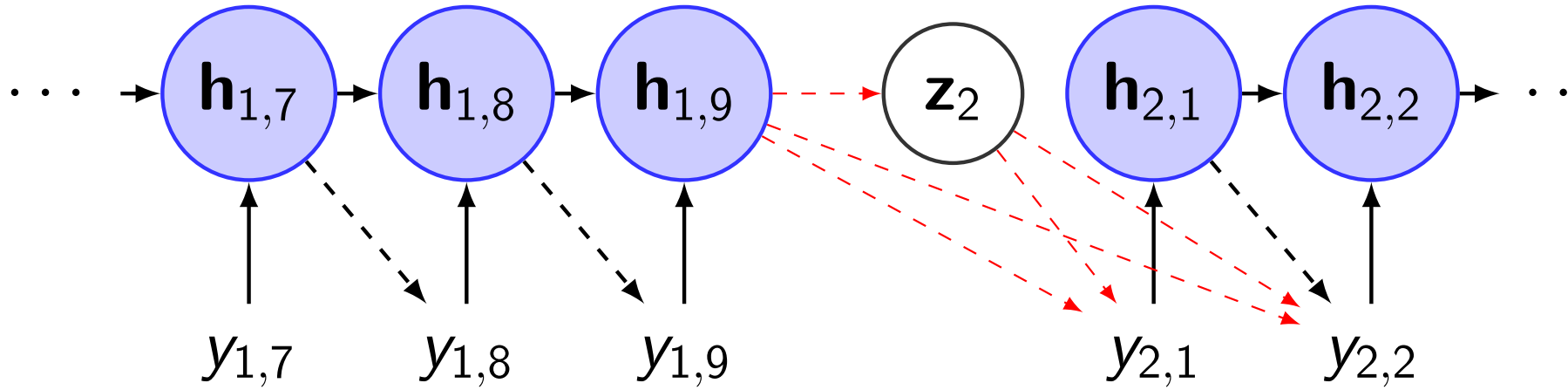
DCLM (Ji et al 2015)

$$\mathbf{h}_{i,n} = f(E_{y_{i,n}} + U\mathbf{h}_{i,n-1}) \quad (1)$$

$$y_{i,n+1} \sim \text{SoftMax}(V\mathbf{h}_{i,n} + U\mathbf{h}_{i-1,N_{i-1}}) \quad (2)$$

$$(3)$$

The Discourse Relation Language Model



D**R**LM (Ji et al 2016)

$$\mathbf{h}_{i,n} = f(E_{y_{i,n}} + U\mathbf{h}_{i,n-1}) \quad (1)$$

$$y_{i,n+1} \sim \text{SoftMax}(V^{(z_i)}\mathbf{h}_{i,n} + U^{(z_i)}\mathbf{h}_{i-1,N_{i-1}}) \quad (2)$$

$$z_i \sim \text{SoftMax}(\beta \cdot \mathbf{h}_{i-1,N_{i-1}} + \mathbf{b}) \quad (3)$$

One Model, Two Tasks

Discourse-driven language modeling

$$p(y_{i,n+1} \mid \boldsymbol{y}_{i,1:n}, \boldsymbol{y}_{i-1}) = \sum_{z_i} p(y_{i,n+1}, z_i \mid \boldsymbol{y}_{i,1:n}, \boldsymbol{y}_{i-1})$$

Discourse relation prediction

$$p(z_i \mid \boldsymbol{y}_i, \boldsymbol{y}_{i-1}) = \frac{p(z_i, \boldsymbol{y}_i \mid \boldsymbol{y}_{i-1})}{\sum_{z'} p(z', \boldsymbol{y}_i \mid \boldsymbol{y}_{i-1})}$$

One Model, Two Tasks

Discourse-driven language modeling

$$p(y_{i,n+1} \mid \boldsymbol{y}_{i,1:n}, \boldsymbol{y}_{i-1}) = \sum_{z_i} p(y_{i,n+1}, z_i \mid \boldsymbol{y}_{i,1:n}, \boldsymbol{y}_{i-1})$$

Discourse relation prediction

$$p(z_i \mid \boldsymbol{y}_i, \boldsymbol{y}_{i-1}) = \frac{p(z_i, \boldsymbol{y}_i \mid \boldsymbol{y}_{i-1})}{\sum_{z'} p(z', \boldsymbol{y}_i \mid \boldsymbol{y}_{i-1})}$$

PDTB Relation Prediction

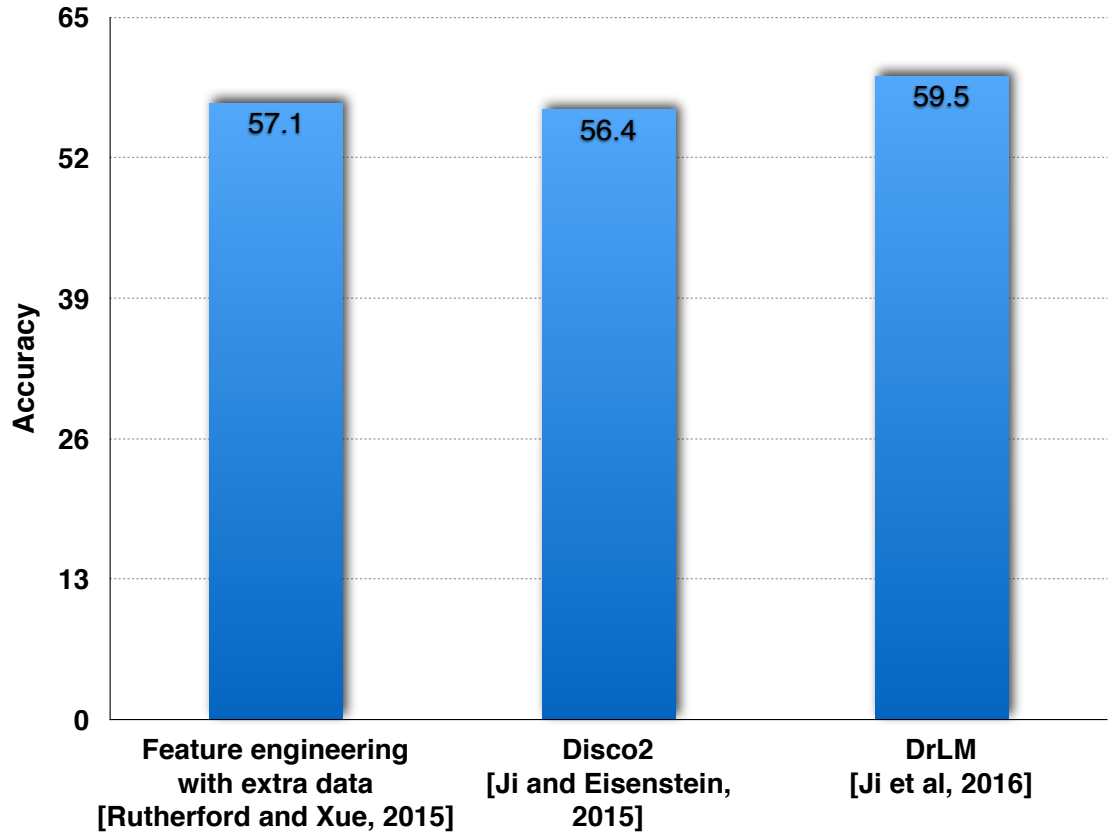


Figure: Evaluation on the first-level implicit discourse relation identification in the PDTB.

One Model, Two Tasks

Discourse-driven language modeling

$$p(y_{i,n+1} \mid \boldsymbol{y}_{i,1:n}, \boldsymbol{y}_{i-1}) = \sum_{z_i} p(y_{i,n+1}, z_i \mid \boldsymbol{y}_{i,1:n}, \boldsymbol{y}_{i-1})$$

Discourse relation prediction

$$p(z_i \mid \boldsymbol{y}_i, \boldsymbol{y}_{i-1}) = \frac{p(z_i, \boldsymbol{y}_i \mid \boldsymbol{y}_{i-1})}{\sum_{z'} p(z', \boldsymbol{y}_i \mid \boldsymbol{y}_{i-1})}$$

One Model, Two Tasks

Discourse-driven language modeling

$$p(y_{i,n+1} \mid \boldsymbol{y}_{i,1:n}, \boldsymbol{y}_{i-1}) = \sum_{z_i} p(y_{i,n+1}, z_i \mid \boldsymbol{y}_{i,1:n}, \boldsymbol{y}_{i-1})$$

Discourse relation prediction

$$p(z_i \mid \boldsymbol{y}_i, \boldsymbol{y}_{i-1}) = \frac{p(z_i, \boldsymbol{y}_i \mid \boldsymbol{y}_{i-1})}{\sum_{z'} p(z', \boldsymbol{y}_i \mid \boldsymbol{y}_{i-1})}$$

Discourse-driven Language Modeling

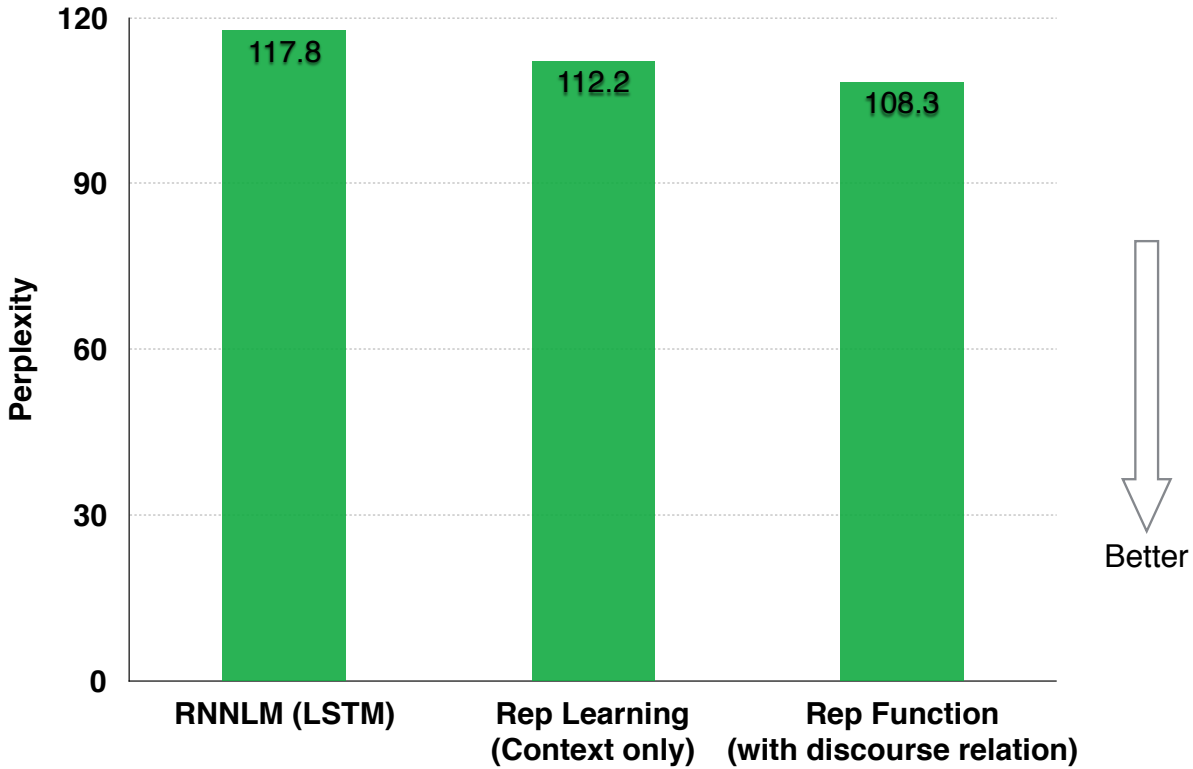


Figure: Language modeling on the PDTB

Dialogue act labeling

Sequential discourse structure on dialogues (Jurafsky et al., 1997)

$$\text{Utterance}_i \text{ — Dialog\_act — } \text{Utterance}_{i+1}$$

Dialogue act labeling

Sequential discourse structure on dialogues (Jurafsky et al., 1997)

$$\text{Utterance}_i \text{ — Dialog\_act — } \text{Utterance}_{i+1}$$

Speaker	Dialog Act	Utterance
A	YES-NO-QUESTION	So do you go to college right now?
B	YES-ANSWER	Yeah,
B	STATEMENT	it's my last year
...	...	...

Dialog act tagging

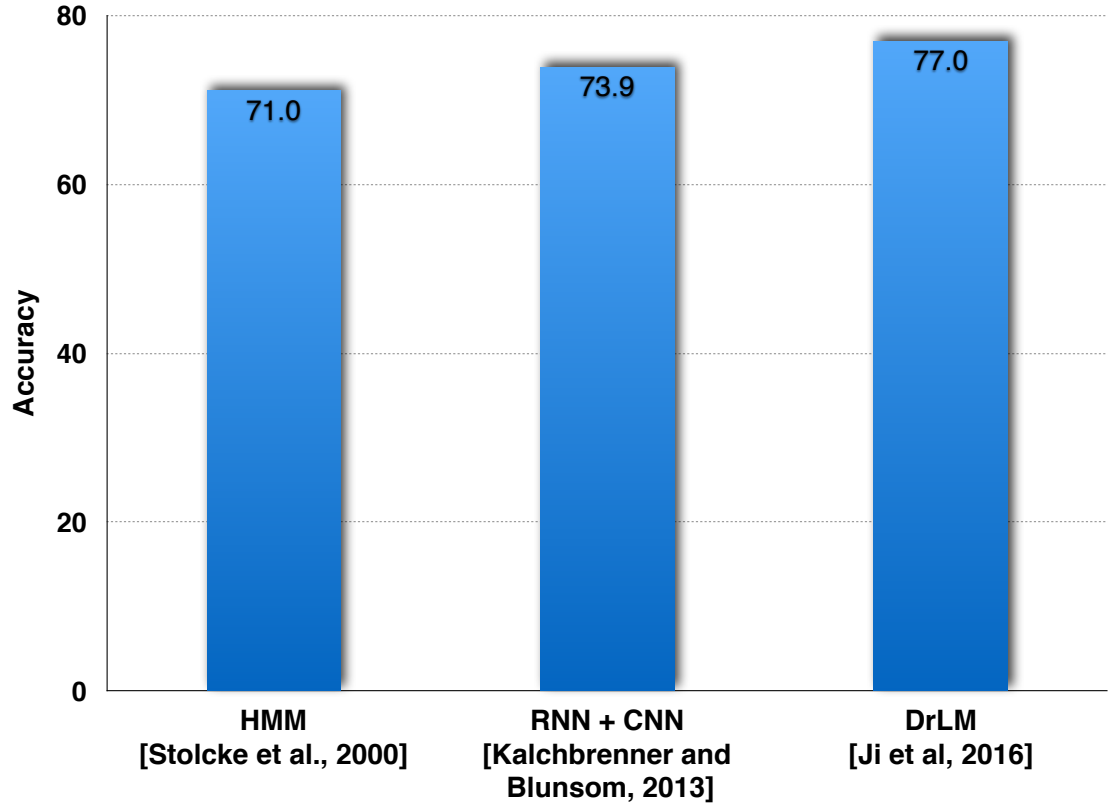


Figure: Dialog act tagging on the Switchboard Dialog Act Corpus (Stolcke et al., 2000).

Discourse-driven Language Modeling

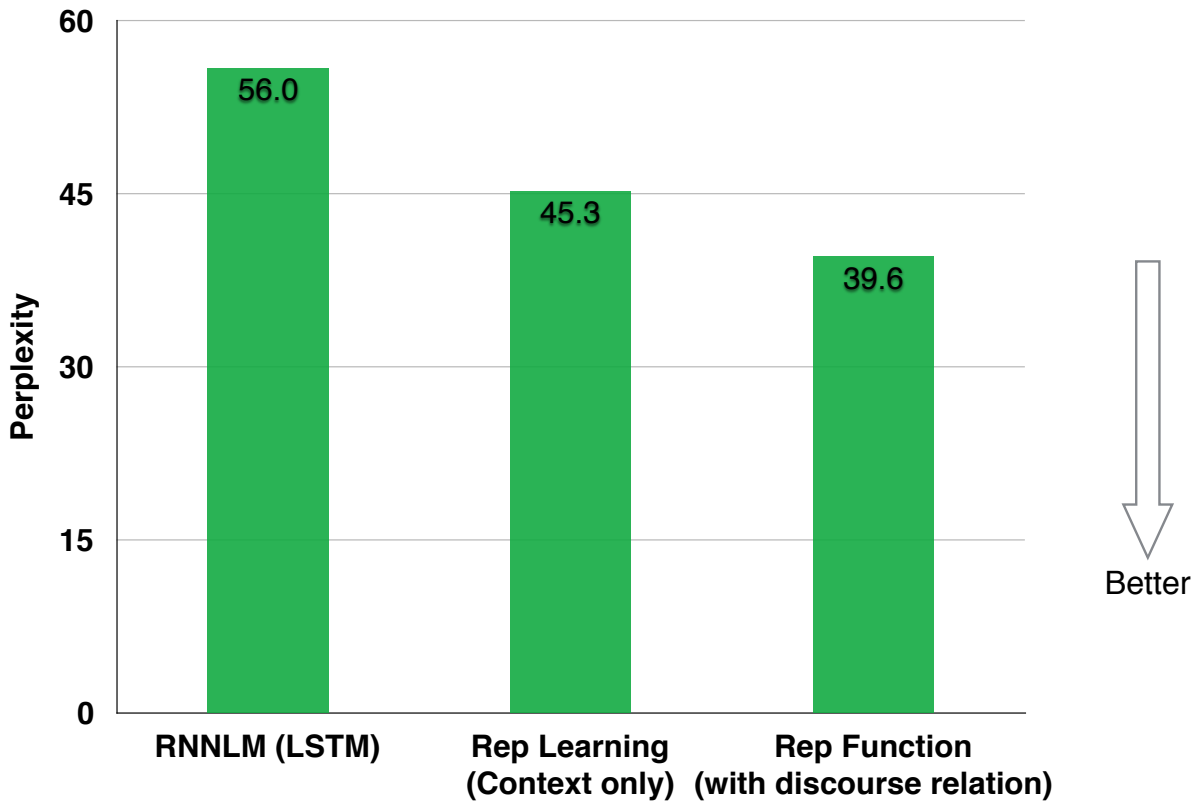
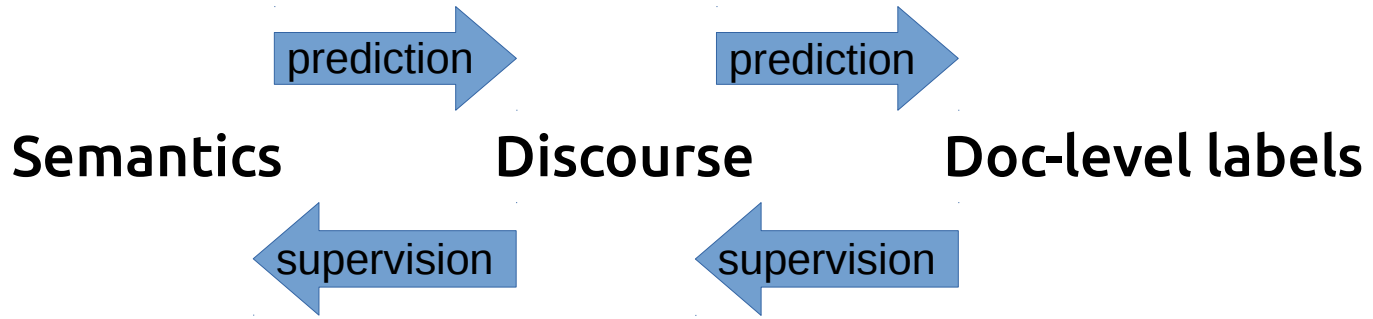


Figure: Language modeling on the Switchboard Dialog Act Corpus

Linking discourse and semantics



- Annotating semantics is hard! Maybe we should give up (Clarke et al., 2010; Artzi & Zettlemoyer, 2011; Berant et al., 2013).
- In comparison, annotating and predicting discourse relations is relatively easy.
- Can discourse structure be learned from downstream tasks?

Next steps

- Better distributed semantics for discourse arguments:
  - Attention mechanisms for word pairs
  - Encoder-decoders (Kiros et al., 2015)
  - LSTM sequence models with latent discourse attention (Li et al., 2015)

Next steps

- Better distributed semantics for discourse arguments:
  - Attention mechanisms for word pairs
  - Encoder-decoders (Kiros et al., 2015)
  - LSTM sequence models with latent discourse attention (Li et al., 2015)
- Discourse-coherent machine translation (long live JSALT 2015!)

Thanks!

Yangfeng Ji

Google Faculty Research Awards

The Computational Linguistics Lab at GT: Akanksha, Parminder Bhatia, Rahul Goel, Naman Goyal, Robert Guthrie, Umashanthi Pavalanathan, Ana L. Smith, Sandeep Soni, Ian Stewart, Patrick Violette, Yi Yang, Gongbo Zhang

References I

Artzi, Y. & Zettlemoyer, L. (2011). Bootstrapping semantic parsers from conversations. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, (pp. 421–432).

Baroni, M. & Zamparelli, R. (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, (pp. 1183–1193). Association for Computational Linguistics.

Berant, J., Chou, A., Frostig, R., & Liang, P. (2013). Semantic parsing on freebase from question-answer pairs. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, (pp. 1533–1544)., Seattle, WA.

Bhatia, P., Ji, Y., & Eisenstein, J. (2015). Better document-level sentiment analysis from rst discourse parsing. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, Lisbon.

Blacoe, W. & Lapata, M. (2012). A comparison of vector-based representations for semantic composition. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, (pp. 546–556).

Clarke, J., Goldwasser, D., Chang, M.-W., & Roth, D. (2010). Driving semantic parsing from the world's response. In *CONLL*, (pp. 18–27). Association for Computational Linguistics.

Hernault, H., Prendinger, H., duVerle, D. A., & Ishizuka, M. HILDA: A discourse parser using support vector machine classification. *Dialogue and Discourse*, *1*(3), 1–33.

Hirao, T., Yoshida, Y., Nishino, M., Yasuda, N., & Nagata, M. (2013). Single-document summarization as a tree knapsack problem. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, (pp. 1515–1520)., Seattle, WA.

Hobbs, J. R. (1979). Coherence and coreference. *Cognitive science*, *3*(1), 67–90.

Ji, Y., Cohn, T., Kong, L., Dyer, C., & Eisenstein, J. (2015). Document context language models. In *International Conference on Learning Representations, Poster Paper*, volume abs/1511.03962.

Ji, Y. & Eisenstein, J. (2014). Representation learning for text-level discourse parsing. In *Proceedings of the Association for Computational Linguistics (ACL)*, Baltimore, MD.

Ji, Y. & Eisenstein, J. (2015). One vector is not enough: Entity-augmented distributional semantics for discourse relations. *Transactions of the Association for Computational Linguistics (TACL)*.

References II

Ji, Y., Haffari, G., & Eisenstein, J. (2016). A latent variable recurrent neural network for discourse relation language models. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, San Diego, CA.

Joty, S., Carenini, G., Ng, R., & Mehdad, Y. (2013). Combining intra- and multi-sentential rhetorical parsing for document-level discourse analysis. In *Proceedings of the Association for Computational Linguistics (ACL)*, Sophia, Bulgaria.

Jurafsky, D., Bates, R., Coccaro, N., Martin, R., Meteer, M., Ries, K., Shriberg, E., Stolcke, A., Taylor, P., Ess-Dykema, V., et al. (1997). Automatic detection of discourse structure for speech recognition and understanding. In *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, (pp. 88–95). IEEE.

Kalchbrenner, N. & Blunsom, P. (2013). Recurrent convolutional neural networks for discourse compositionality. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, (pp. 119–126)., Sofia, Bulgaria. Association for Computational Linguistics.

Kiros, R., Zhu, Y., Salakhudinov, R., Zemel, R. S., Torralba, A., Urtasun, R., & Fidler, S. (2015). Skip-thought vectors. In *Neural Information Processing Systems (NIPS)*, Montréal.

Le, Q. & Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Levy, O. & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Neural Information Processing Systems (NIPS)*, Montréal.

Li, J., Li, R., & Hovy, E. (2014). Recursive deep models for discourse parsing. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

Li, J., Luong, M.-T., & Jurafsky, D. (2015). A hierarchical neural autoencoder for paragraphs and documents. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, Lisbon.

Lin, Z., Kan, M.-Y., & Ng, H. T. (2009). Recognizing implicit discourse relations in the penn discourse treebank. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, (pp. 343–351)., Singapore.

Louis, A., Joshi, A., & Nenkova, A. (2010). Discourse indicators for content selection in summarization. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, (pp. 147–156). Association for Computational Linguistics.

References III

Marcu, D. (1996). Building up rhetorical structure trees. In *Proceedings of the National Conference on Artificial Intelligence*, (pp. 1069–1074).

Marcu, D. (1999). Discourse trees are good indicators of importance in text. *Advances in automatic text summarization*, 123–136.

Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In *INTERSPEECH*, (pp. 1045–1048).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, (pp. 3111–3119).

Polajnar, T., Rimell, L., & Clark, S. (2015). An exploration of discourse-based sentence spaces for compositional distributional semantics. In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, (pp. 1–11)., Lisbon, Portugal. Association for Computational Linguistics.

Rutherford, A. T. & Xue, N. (2014). Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*.

Sagae, K. Analysis of discourse structure with syntactic dependencies and data-driven shift-reduce parsing. In *Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09)*, (pp. 81–84)., Paris, France. Association for Computational Linguistics.

Socher, R., Huang, E. H., Pennington, J., Ng, A. Y., & Manning, C. D. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Advances in Neural Information Processing Systems (NIPS)*.

Socher, R., Huval, B., Manning, C. D., & Ng, A. Y. (2012). Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, (pp. 1201–1211). Association for Computational Linguistics.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, Seattle, WA.

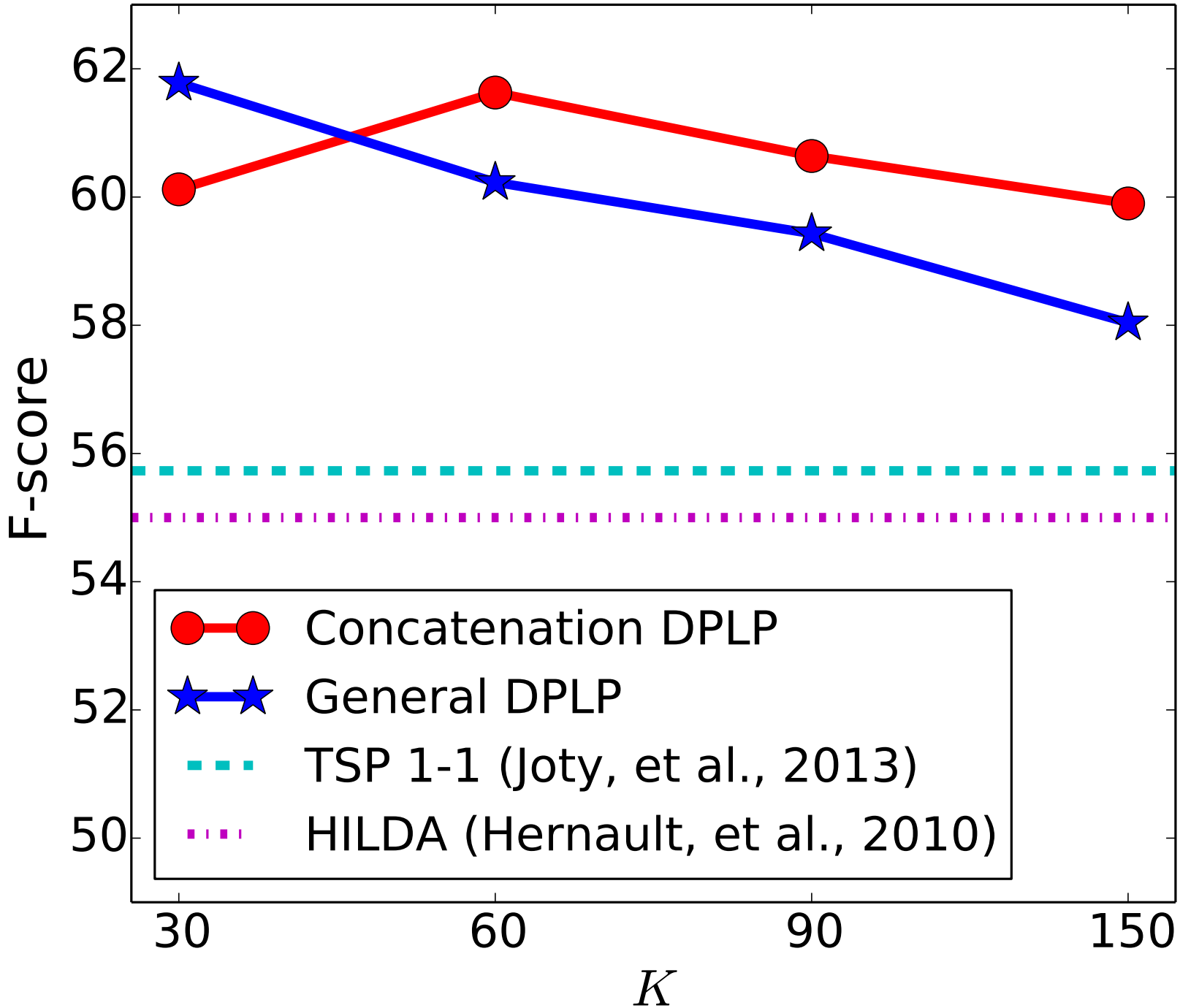
References IV

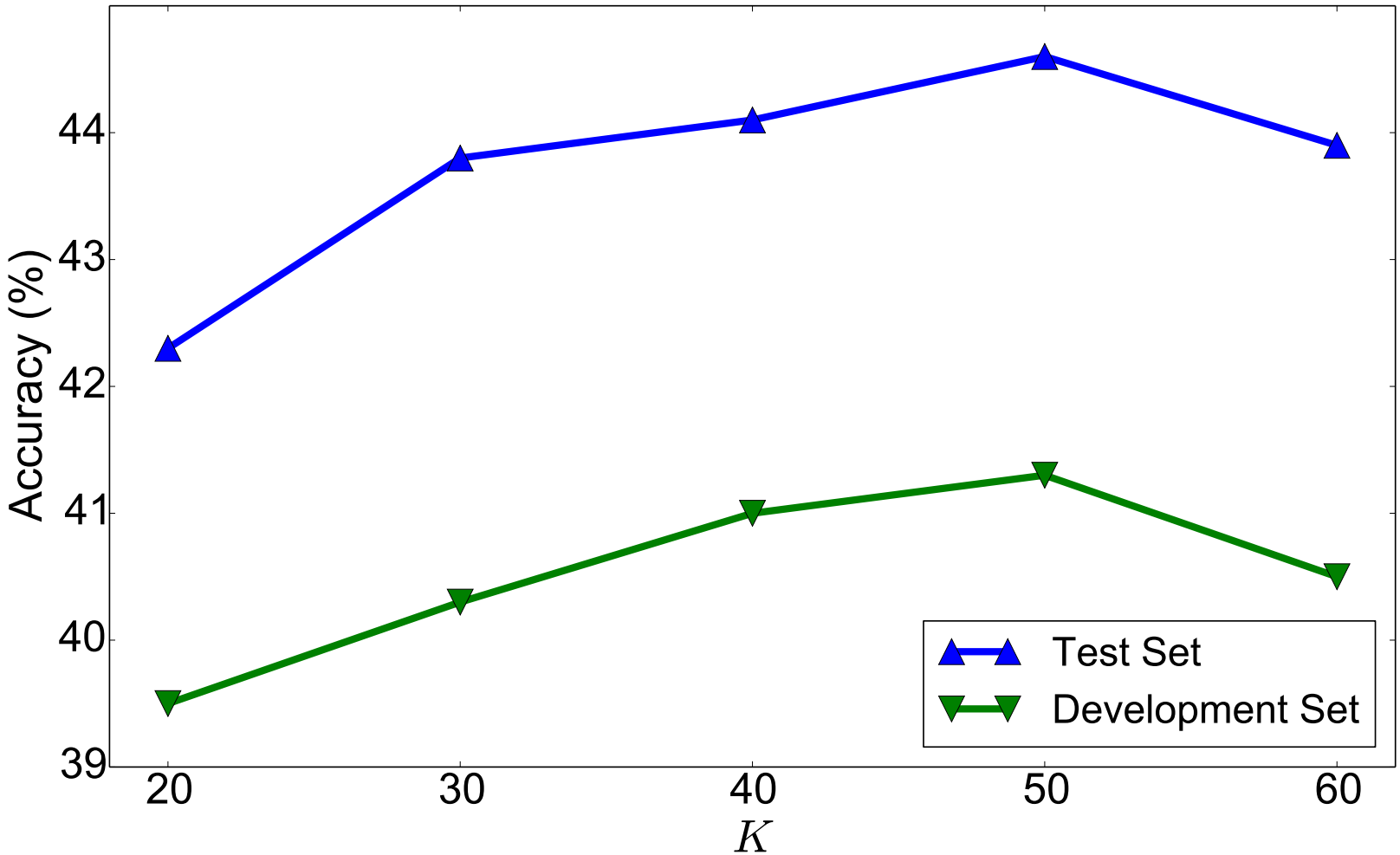
Somasundaran, S., Namata, G., Wiebe, J., & Getoor, L. (2009). Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, Singapore.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van Ess-Dykema, C., & Meteer, M. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, *26*(3), 339–373.

Wang, J. & Lan, M. (2015). A refined end-to-end discourse parser. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, (pp. 17–24)., Beijing, China. Association for Computational Linguistics.

Yang, B. & Cardie, C. (2014). Context-aware learning for sentence-level sentiment analysis with posterior regularization. In *Proceedings of the Association for Computational Linguistics (ACL)*, Baltimore, MD.


62
60
58
56
54
52
50
F-score
30
60
90
150
K
Concatenation DPLP
General DPLP
TSP 1-1 (Joty, et al., 2013)
HILDA (Hernault, et al., 2010)


Accuracy (%)
44
43
42
41
40
39
20
30
40
50
60
K
Test Set
Development Set

Examples

(5) **Arg 1**: The drop in profit reflected, in part, continued softness in financial advertising at [The Wall Street Journal] and Barron's magazine.
Arg 2: Ad linage at [the Journal] fell 6.1% in the third quarter.

- Correct: RESTATEMENT
- Without coreference: CAUSE

Examples

(6) **Arg 1**: Half of [*them*]$_1$ are really scared and want to sell but [*I*]$_2$'m trying to talk them out of it.
Arg 2: If [*they*]$_1$ all were bullish, [*I*]$_2$'d really be upset.

- Correct: CONTRAST
- Without coreference: CONJUNCTION