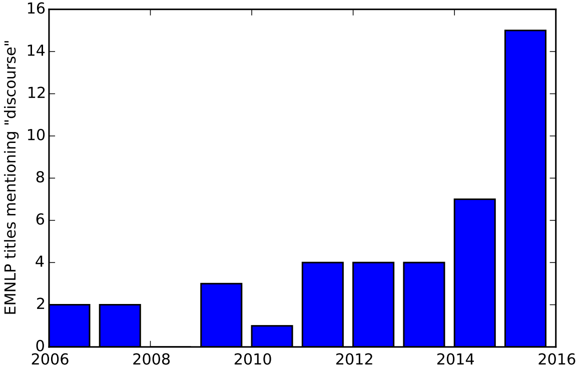
From Distributed Semantics to Discourse, and Back

Jacob Eisenstein

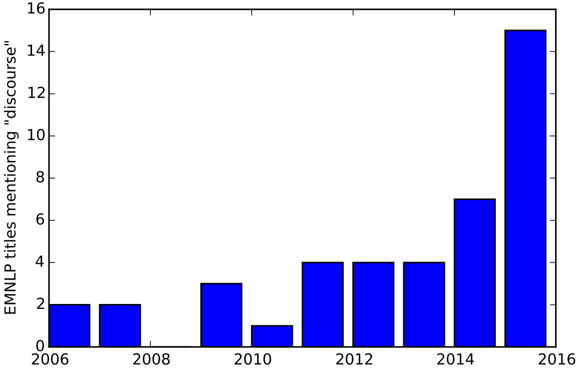
Georgia Institute of Technology

September 17, 2015

Discourse is blowing up!



Discourse is blowing up!



- Thanks for all the data!
- This talk is about what we can do with it.

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.

Applications to sentiment analysis (Somasundaran et al., 2009; Yang & Cardie, 2014), readability prediction (Pitler & Nenkova, 2008), summarization (Louis et al., 2010), . . .

Why is predicting discourse relations hard?

Discourse relations are fundamentally semantic (Hobbs, 1979):

(2) Lisbon is a fun place to visit. (Because) there are old buildings and interesting food.

- Typical solution is bilexical features, e.g., $\langle$fun,buildings$\rangle$, $\langle$place,interesting$\rangle$, ... (Lin et al., 2009; Rutherford & Xue, 2014)
- But bilexical features are sparse and noisy, and discourse-annotated datasets are small.

Can distributed semantics help?

Distributed semantics proposes to capture meaning in dense numerical vectors. 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, 2009) with an SVM classifier.

Building the Distributed Representation

- **Elementary discourse units**:

  $$\boldsymbol{u}(\text{Lisbon is a fun place to visit}) = \boldsymbol{u}_{\text{Lisbon}} + \boldsymbol{u}_{\text{is}} + \ldots$$

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

Building the Distributed Representation

- **Elementary discourse units**:

  $\boldsymbol{u}(\text{Lisbon is a fun place to visit}) = \boldsymbol{u}_{\text{Lisbon}} + \boldsymbol{u}_{\text{is}} + \ldots$

  "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 et al., 2010)	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 et al., 2010)	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 et al., 2010)	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

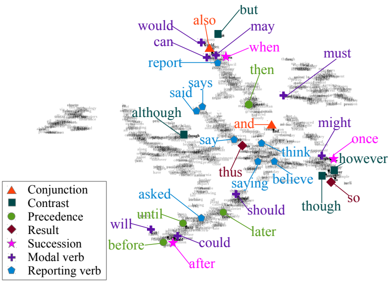
	Span	Nuclearity	Relation
Annotator agreement	88.7	77.7	65.8
HILDA (Hernault et al., 2010)	83.0	68.4	54.8
TSP (Joty et al., 2013)	82.7	68.4	55.7
“Basic features”	79.4	68.0	53.0
Distributed			
Collobert & Weston	75.3	67.1	53.8
Non-neg. matrix factorization	78.6	67.7	54.8
Distributed	80.9	69.4	59.0

RST Results

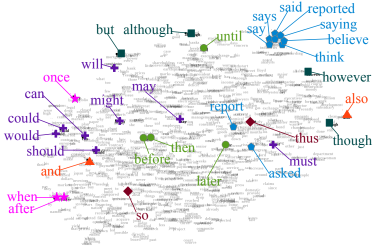
	Span	Nuclearity	Relation
Annotator agreement	88.7	77.7	65.8
HILDA (Hernault et al., 2010)	**83.0**	68.4	54.8
TSP (Joty et al., 2013)	82.7	68.4	55.7
"Basic features"	79.4	68.0	53.0
Distributed			
Collobert & Weston	75.3	67.1	53.8
Non-neg. matrix factorization	78.6	67.7	54.8
Distributed	80.9	69.4	59.0
+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

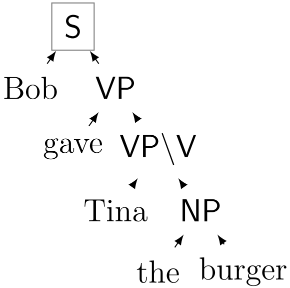
"One Vector is not Enough: Entity-Augmented Distributed Semantics for Discourse Relations" (Ji & Eisenstein, 2015)

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

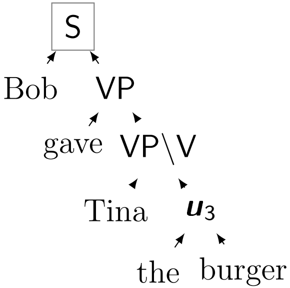
$$\langle 0010, 1011 \rangle, \langle 1010, 0001 \rangle, \ldots$$

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

Vector-semantic composition

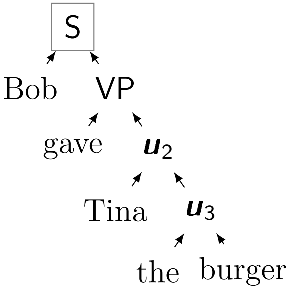


Vector-semantic composition



$$\boldsymbol{u}_3 = \tanh\left(\mathbf{U}\left[\boldsymbol{u}_{\text{the}}^\top\ \boldsymbol{u}_{\text{burger}}^\top\right]^\top\right)$$

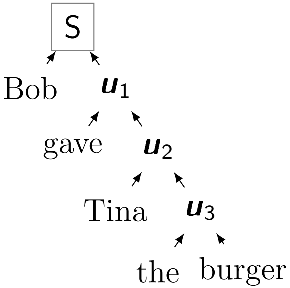
Vector-semantic composition



$$\boldsymbol{u}_3 = \tanh\left(\mathbf{U}\left[\boldsymbol{u}_{\text{the}}^\top\ \boldsymbol{u}_{\text{burger}}^\top\right]^\top\right)$$

$$\boldsymbol{u}_2 = \tanh\left(\mathbf{U}\left[\boldsymbol{u}_{\text{Tina}}^\top\ \boldsymbol{u}_3^\top\right]^\top\right)$$

Vector-semantic composition

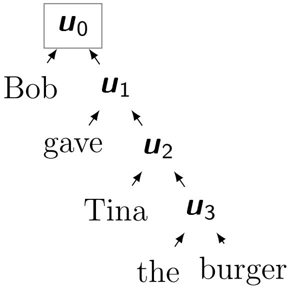


$$\boldsymbol{u}_3 = \tanh\left(\mathbf{U}\left[\boldsymbol{u}_{\text{the}}^\top\ \boldsymbol{u}_{\text{burger}}^\top\right]^\top\right)$$

$$\boldsymbol{u}_2 = \tanh\left(\mathbf{U}\left[\boldsymbol{u}_{\text{Tina}}^\top\ \boldsymbol{u}_3^\top\right]^\top\right)$$

$$\boldsymbol{u}_1 = \tanh\left(\mathbf{U}\left[\boldsymbol{u}_{\text{gave}}^\top\ \boldsymbol{u}_2^\top\right]^\top\right)$$

Vector-semantic composition



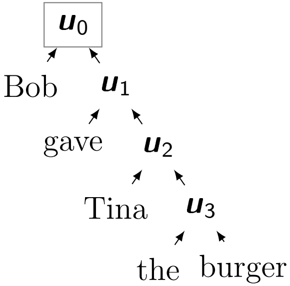
$$\boldsymbol{u}_3 = \tanh\left(\mathbf{U}\left[\boldsymbol{u}_{\text{the}}^\top\ \boldsymbol{u}_{\text{burger}}^\top\right]^\top\right)$$

$$\boldsymbol{u}_2 = \tanh\left(\mathbf{U}\left[\boldsymbol{u}_{\text{Tina}}^\top\ \boldsymbol{u}_3^\top\right]^\top\right)$$

$$\boldsymbol{u}_1 = \tanh\left(\mathbf{U}\left[\boldsymbol{u}_{\text{gave}}^\top\ \boldsymbol{u}_2^\top\right]^\top\right)$$

$$\boldsymbol{u}_0 = \tanh\left(\mathbf{U}\left[\boldsymbol{u}_{\text{Bob}}^\top\ \boldsymbol{u}_1^\top\right]^\top\right)$$

Vector-semantic composition



$$\boldsymbol{u}_3 = \tanh\left(\mathbf{U}\left[\boldsymbol{u}_{\text{the}}^\top\ \boldsymbol{u}_{\text{burger}}^\top\right]^\top\right)$$

$$\boldsymbol{u}_2 = \tanh\left(\mathbf{U}\left[\boldsymbol{u}_{\text{Tina}}^\top\ \boldsymbol{u}_3^\top\right]^\top\right)$$

$$\boldsymbol{u}_1 = \tanh\left(\mathbf{U}\left[\boldsymbol{u}_{\text{gave}}^\top\ \boldsymbol{u}_2^\top\right]^\top\right)$$

$$\boldsymbol{u}_0 = \tanh\left(\mathbf{U}\left[\boldsymbol{u}_{\text{Bob}}^\top\ \boldsymbol{u}_1^\top\right]^\top\right)$$

- DISCO2: **Dis**tributional **co**mpositional semantics for **disco**urse.
- Same architecure as Socher et al. (2011).

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

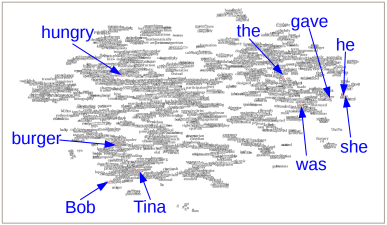
Most common class	26.0
Additive word representations	28.7
Lin et al. (2009)	40.2
SFM: Our reimplementation of Lin et al. (2009)	39.7
SFMB: Lin et al. (2009) + Brown clusters	40.7
Disco2	37.0
Disco2 + SFMB	**43.8**

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.

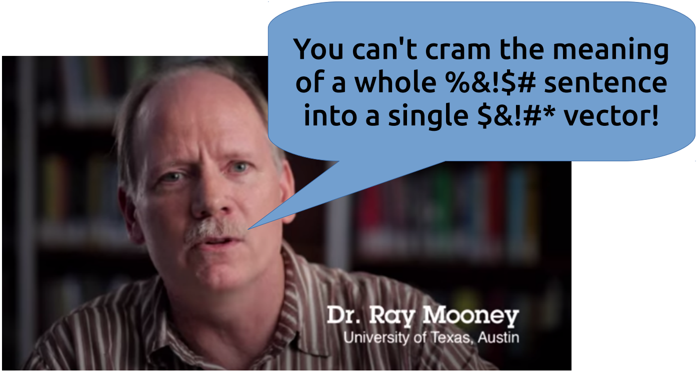
If we insist on representing each discourse argument as a single vector, we lose the ability to track references across the discourse.

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:

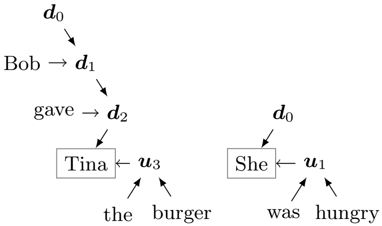
- $s1$: She got the burger from Bob
- $s2$: She was hungry

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

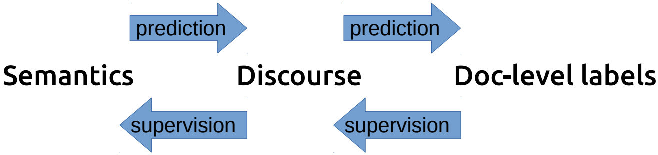
Most common class	26.0
Additive word representations	28.7
Lin et al. (2009)	40.2
SFM: Our reimplementation of Lin et al. (2009)	39.7
SFMB: Lin et al. (2009) + Brown clusters	40.7
Disco2	37.0
Disco2 + SFMB	43.8
Disco2 + SFMB + entity semantics	**44.6**

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

Lessons learned

- **Density**: Bengio et al. (2013) argues that dense distributed representations are more compact, thus better for learning.
- **Supervision**: learn distributed representations from discourse annotations.
- **Structured distributed representations** have advantages of both symbolic and distributed semantics.

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.
- Or, discourse structure can be learned from distant supervision (Ji et al., 2015).

Thanks!

Yangfeng Ji
(graduating soon!)

Google
Faculty Research Awards

The Computational Linguistics Lab at GT: Parminder Bhatia, Rahul Goel, Naman Goyal, 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).

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828.

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).

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 *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, (pp. 18–27). Association for Computational Linguistics.

Hernault, H., Prendinger, H., duVerle, D. A., & Ishizuka, M. (2010). HILDA: A Discourse Parser Using Support Vector Machine Classification. *Dialogue and Discourse*, *1*(3), 1–33.

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

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)*, *3*, 329–344.

Ji, Y., Zhang, G., & Eisenstein, J. (2015). Closing the gap: Domain adaptation from explicit to implicit discourse relations. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

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.

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

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.

References II

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.

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

Pitler, E. & Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, (pp. 186–195)., Honolulu, HI.

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)*, Stroudsburg, Pennsylvania. Association for Computational Linguistics.

Sagae, K. (2009). 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. (2011). Dynamic Pooling And Unfolding Recursive Autoencoders For Paraphrase Detection. In *Advances in Neural Information Processing Systems (NIPS)*.

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.

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.