

Distributed Semantics for Automatically Classifying Discourse Relations

Jacob Eisenstein

Georgia Institute of Technology

October 1, 2015

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?

Coherence relations are fundamentally semantic (Hobbs, 1979):

- (2) 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.,

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

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.

$$\mathbf{u}_{\text{easy}} = [0.1, -0.5, -0.4, \dots]$$

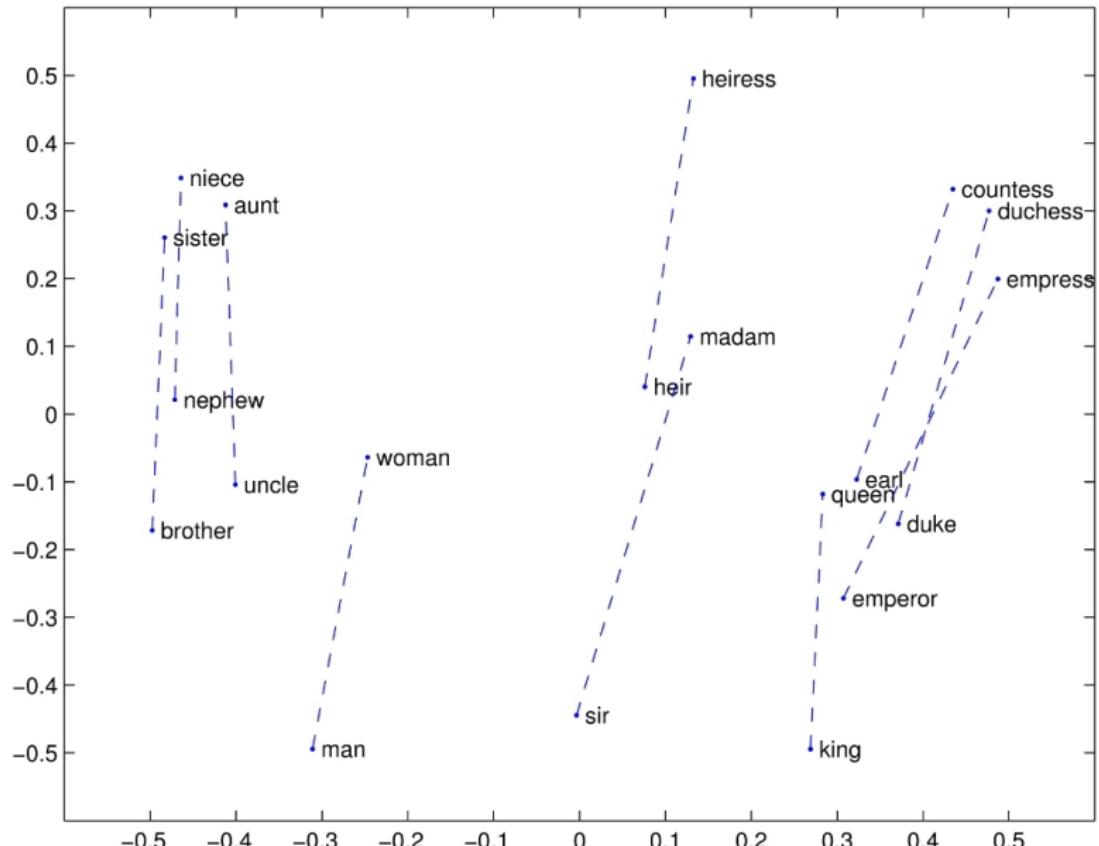
$$\mathbf{u}_{\text{short}} = [-0.6, 0.5, -1.0, \dots]$$

$$\mathbf{u}_{\text{visit}} = [-0.7, 1.0, 0.6, \dots]$$

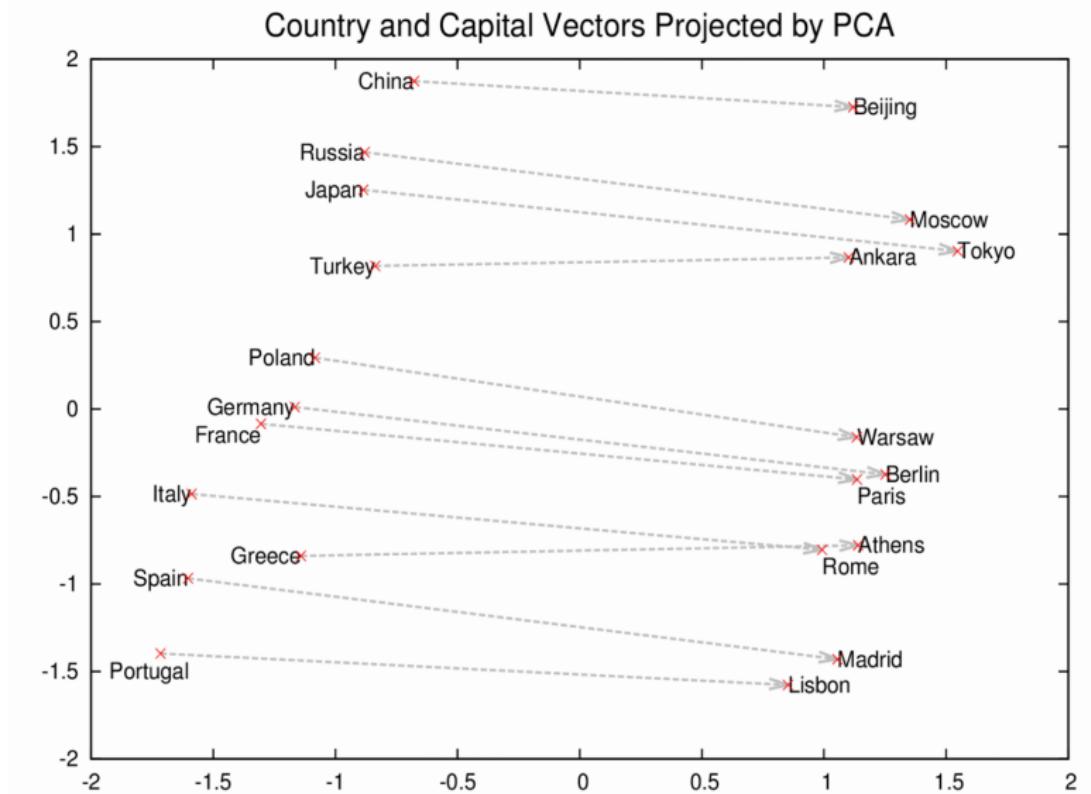
$$\mathbf{u}_{\text{distance}} = [1.7, 1.9, -1.5, \dots]$$

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

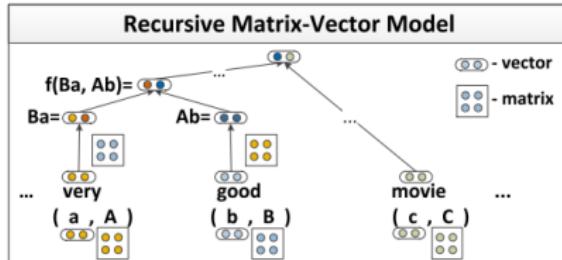


Pretty picture 2



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

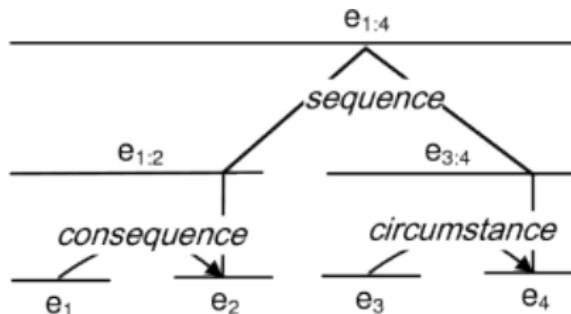
Key questions:

- ▶ 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

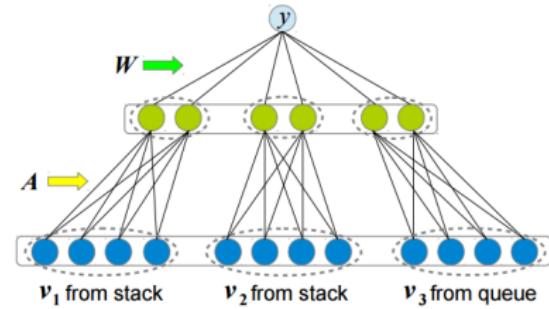


- ▶ **Algorithm:** shift-reduce (Marcu, 1996; Sagae, 2009) with an SVM classifier.

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:**

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

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

Building the distributed representations

- ▶ **Elementary discourse units:**

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

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

- ▶ **Higher-order discourse units** inherit the distributed representation of their nucleus (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
<i>Distributed</i>			
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
<i>Distributed</i>			
Collobert & Weston	75.3	67.1	53.8
Non-neg. matrix factorization	78.6	67.7	54.8

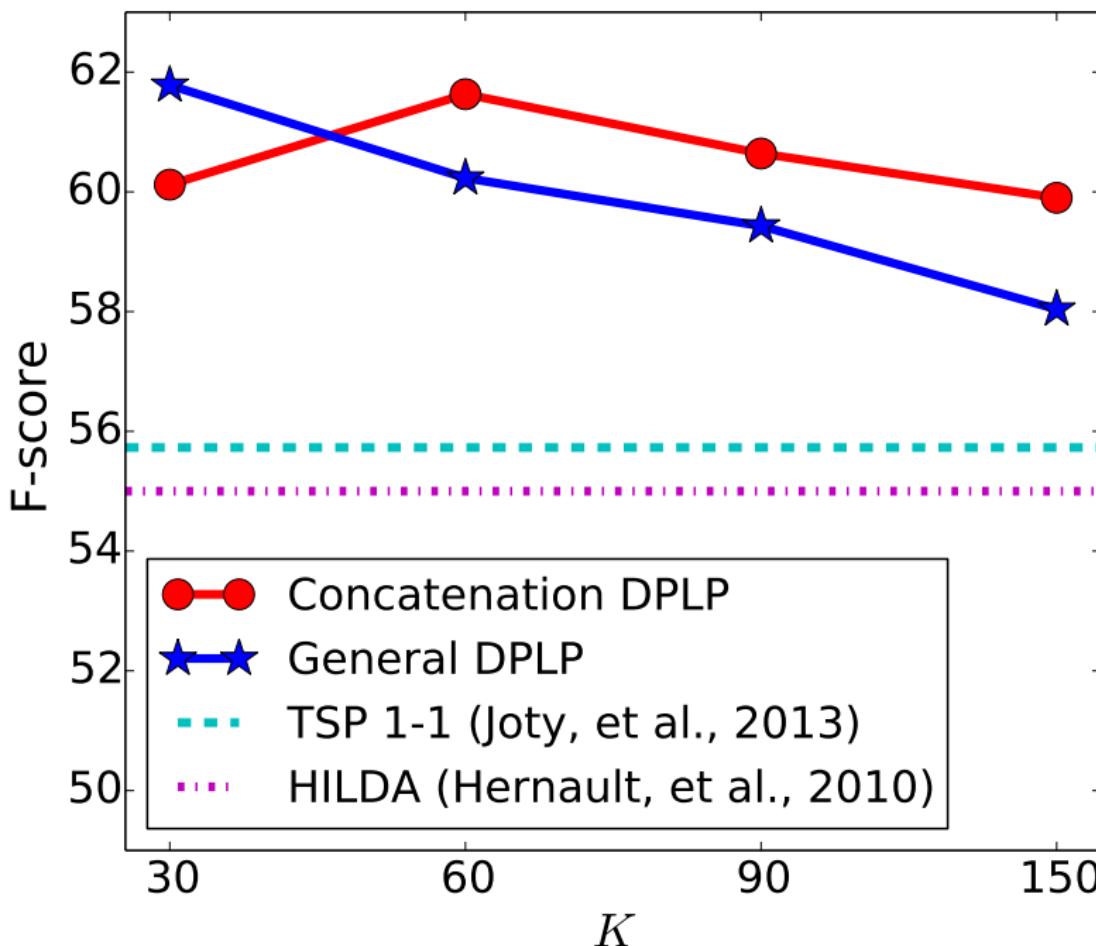
RST Results

	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
<i>Distributed</i>			
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

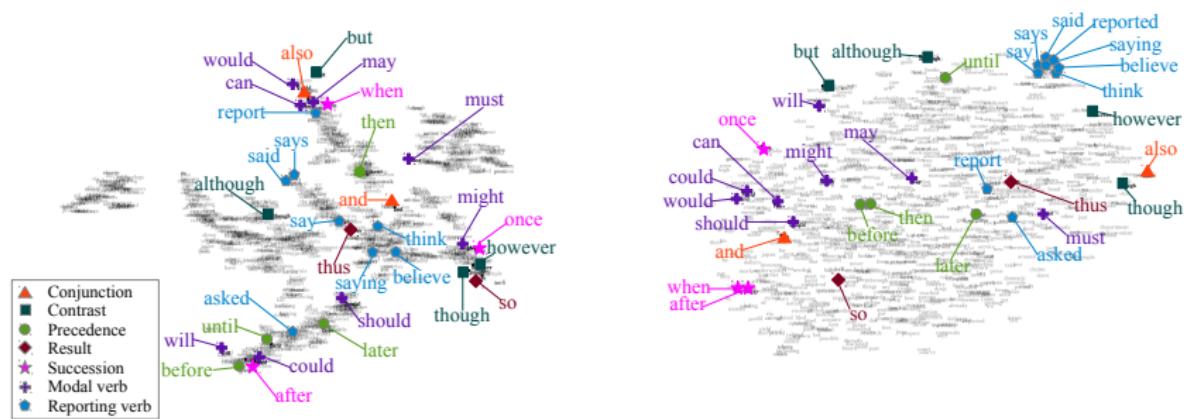
RST Results

	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
<i>Distributed</i>			
Collobert & Weston	75.3	67.1	53.8
Non-neg. matrix factorization	78.6	67.7	54.8
<i>Distributed</i>			
+basic features	80.9	69.4	59.0
	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).

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

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).

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

Project 2: PDTB Implicit Relations

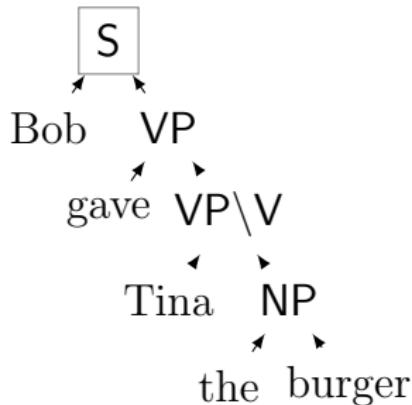
“One Vector is not Enough: Entity-Augmented Distributed Semantics for Discourse Relations” (Ji & Eisenstein, 2015)

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

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

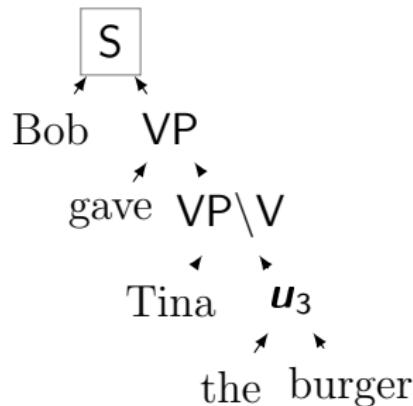
- ▶ **Our approach:** construct meaning of discourse units through composition over the parse tree.

Vector-semantic composition

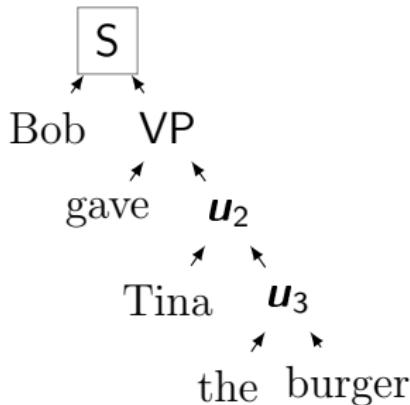


Vector-semantic composition

$$\mathbf{u}_3 = \tanh \left(\mathbf{U} \begin{bmatrix} \mathbf{u}_{\text{the}}^\top & \mathbf{u}_{\text{burger}}^\top \end{bmatrix}^\top \right)$$



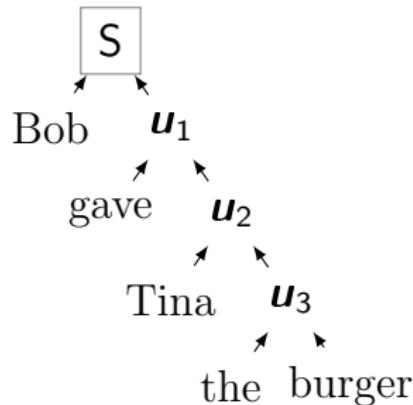
Vector-semantic composition



$$\mathbf{u}_3 = \tanh \left(\mathbf{U} \begin{bmatrix} \mathbf{u}_{\text{the}}^\top & \mathbf{u}_{\text{burger}}^\top \end{bmatrix}^\top \right)$$

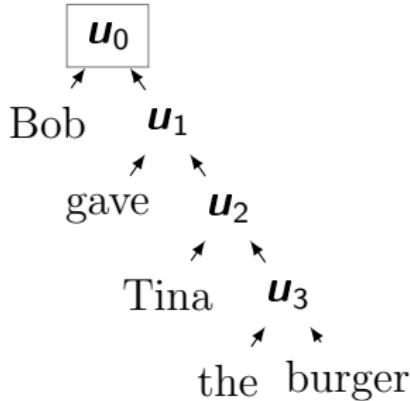
$$\mathbf{u}_2 = \tanh \left(\mathbf{U} \begin{bmatrix} \mathbf{u}_{\text{Tina}}^\top & \mathbf{u}_3^\top \end{bmatrix}^\top \right)$$

Vector-semantic composition



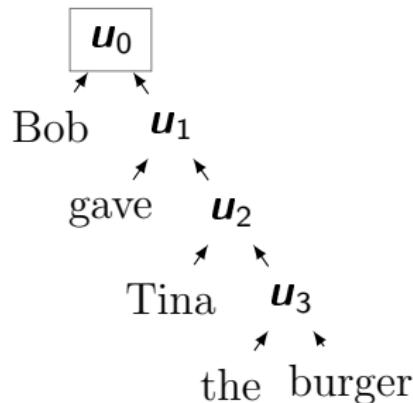
$$\begin{aligned} \mathbf{u}_3 &= \tanh \left(\mathbf{U} \left[\mathbf{u}_{\text{the}}^\top \mathbf{u}_{\text{burger}}^\top \right]^\top \right) \\ \mathbf{u}_2 &= \tanh \left(\mathbf{U} \left[\mathbf{u}_{\text{Tina}}^\top \mathbf{u}_3^\top \right]^\top \right) \\ \mathbf{u}_1 &= \tanh \left(\mathbf{U} \left[\mathbf{u}_{\text{gave}}^\top \mathbf{u}_2^\top \right]^\top \right) \end{aligned}$$

Vector-semantic composition



$$\begin{aligned}\mathbf{u}_3 &= \tanh \left(\mathbf{U} \left[\mathbf{u}_{\text{the}}^\top \mathbf{u}_{\text{burger}}^\top \right]^\top \right) \\ \mathbf{u}_2 &= \tanh \left(\mathbf{U} \left[\mathbf{u}_{\text{Tina}}^\top \mathbf{u}_3^\top \right]^\top \right) \\ \mathbf{u}_1 &= \tanh \left(\mathbf{U} \left[\mathbf{u}_{\text{gave}}^\top \mathbf{u}_2^\top \right]^\top \right) \\ \mathbf{u}_0 &= \tanh \left(\mathbf{U} \left[\mathbf{u}_{\text{Bob}}^\top \mathbf{u}_1^\top \right]^\top \right)\end{aligned}$$

Vector-semantic composition



$$\begin{aligned}\mathbf{u}_3 &= \tanh \left(\mathbf{U} \left[\mathbf{u}_{\text{the}}^\top \mathbf{u}_{\text{burger}}^\top \right]^\top \right) \\ \mathbf{u}_2 &= \tanh \left(\mathbf{U} \left[\mathbf{u}_{\text{Tina}}^\top \mathbf{u}_3^\top \right]^\top \right) \\ \mathbf{u}_1 &= \tanh \left(\mathbf{U} \left[\mathbf{u}_{\text{gave}}^\top \mathbf{u}_2^\top \right]^\top \right) \\ \mathbf{u}_0 &= \tanh \left(\mathbf{U} \left[\mathbf{u}_{\text{Bob}}^\top \mathbf{u}_1^\top \right]^\top \right)\end{aligned}$$

- DISCO2: **D**istributional **c**ompositional semantics for **d**iscourse.
- Same architecture as Socher et al. (2011).

A bilinear model

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

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

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

Learning

- ▶ Word representations are fixed to WORD2VEC.
Fine-tuning → 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 reimplemention of Lin et al. (2009)	39.7
SFMB: Lin et al. (2009) + Brown clusters	40.7

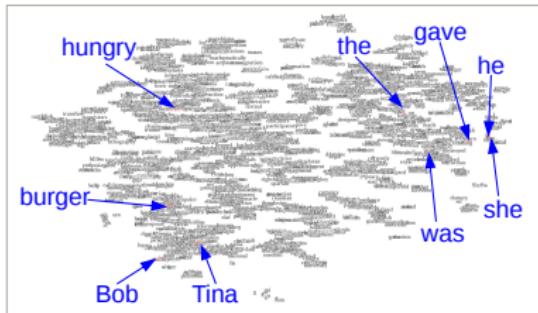
PDTB Results

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.

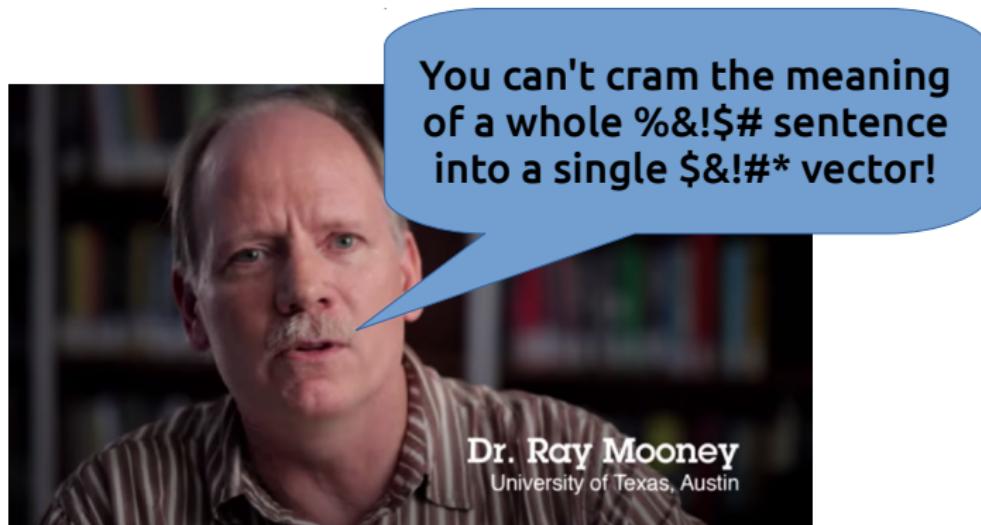
If we insist on representing each discourse argument as a single vector, we lose the ability to track references across the discourse.

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:

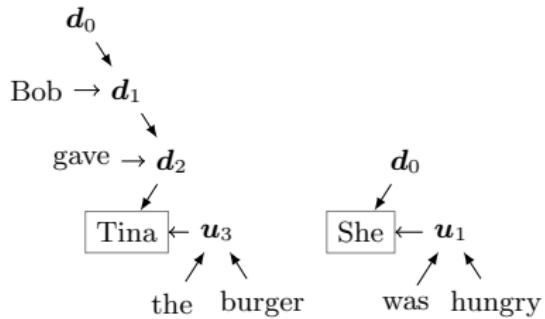
- ▶ s_1 : She got the burger from Bob
- ▶ s_2 : She was hungry

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

The downward pass

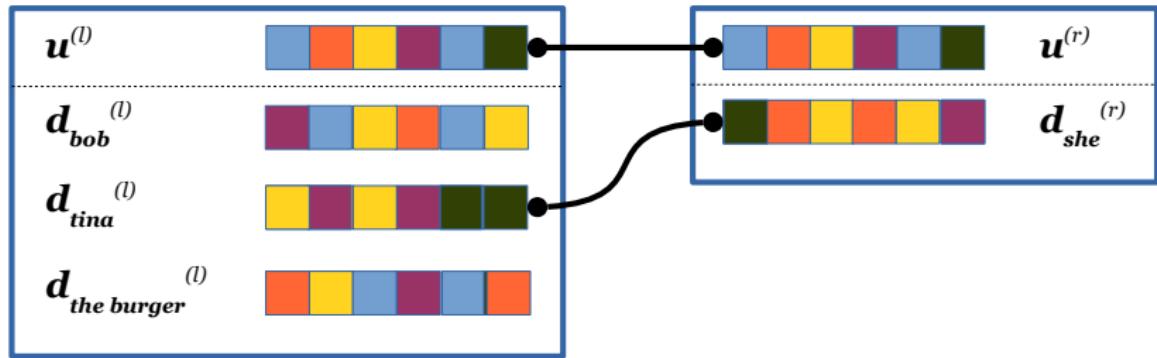
A **downward pass** computes a downward vector for each node in the parse.

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



This computation preserves the feedforward architecture.

A new bilinear model



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

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

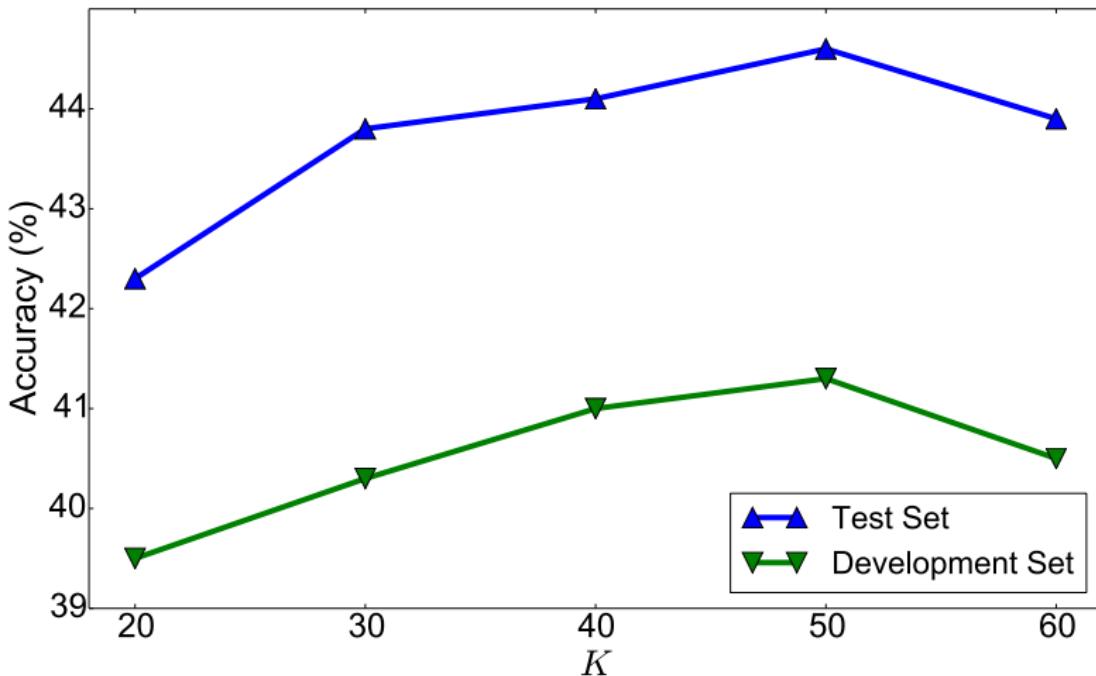
PDTB Results

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 reimplemention 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%.



Examples

(3) **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

(4) **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

Project 3: Learning with less supervision

“Closing the Gap: Domain Adaptation from Explicit to Implicit Discourse Relations” (Ji et al., 2015)

- ▶ **Goal:** learn to classify implicit discourse relations from explicitly-labeled examples.
Will this work?

Project 3: Learning with less supervision

“Closing the Gap: Domain Adaptation from Explicit to Implicit Discourse Relations” (Ji et al., 2015)

- ▶ **Goal:** learn to classify implicit discourse relations from explicitly-labeled examples.
Will this work?



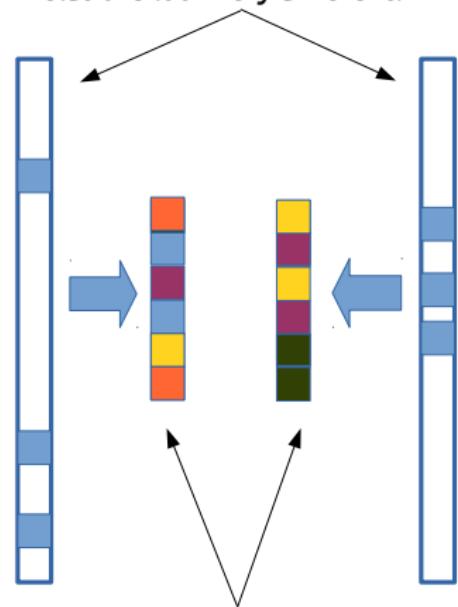
Asr & Demberg (2012, 2013): the decision to provide explicit discourse connectives reflects information content.

- ▶ **Approach:** treat explicit and implicit relations as different *domains* (Braud & Denis, 2014)

Challenges for domain adaptation

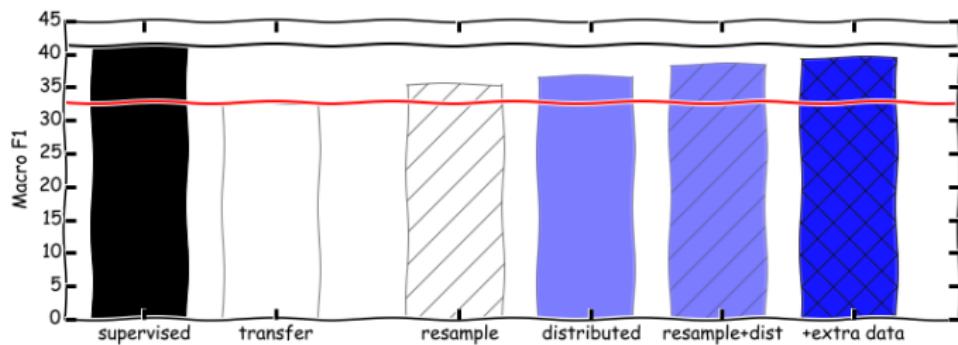
- ▶ **Mismatched feature representation** (Ben-David et al., 2007).
Solution: project instances into a dense, distributed representation.
- ▶ **Mismatched label distribution** (Jiang & Zhai, 2007). Solution: resample instances based on target relation distribution.

Original, sparse representation;
explicit and implicit discourse
relations look very different.



Induced distributed representation;
transfer from explicit relations is
more effective here.

Results: 80% reduction in transfer loss

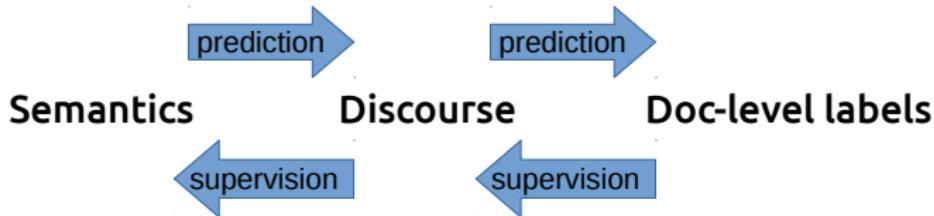


Training data	+DistSem	+Resamp.	TEMP.	COMP.	Exp.	CONT.	Avg F ₁
1. Implicit			24.2	28.9	68.8	43.5	41.3
2. Explicit	No	No	17.3	23.9	53.5	36.2	32.7
3. Explicit	No	Yes	17.7	25.4	62.7	36.0	35.4
4. Explicit	Yes	No	18.7	25.9	63.4	38.9	36.7
5. Explicit	Yes	Yes	19.3	25.7	68.1	41.4	38.6
6. +CNN	Yes	Yes	20.4	26.3	68.9	42.3	39.5

Lessons learned

- ▶ **Density**: Bengio et al. (2013) argues that dense distributed representations are more compact, thus better for learning.
- ▶ **Supervision**: learn distributed semantic 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.
- ▶ Can discourse structure be learned from downstream tasks?

Next steps

- ▶ Better distributed semantics for discourse arguments:
 - ▶ Encoder-decoders (Kiros et al., 2015)
 - ▶ LSTM sequence models with latent discourse attention (Li et al., 2015)
 - ▶ Tensor composition (Polajnar et al., 2015)

Next steps

- ▶ Better distributed semantics for discourse arguments:
 - ▶ Encoder-decoders (Kiros et al., 2015)
 - ▶ LSTM sequence models with latent discourse attention (Li et al., 2015)
 - ▶ Tensor composition (Polajnar et al., 2015)
- ▶ Applications to models of argumentation, social media, and political discourse.

Thanks!



Yangfeng Ji



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).
- Asr, F. T. & Demberg, V. (2012). Implicitness of discourse relations. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, (pp. 2669–2684)., Mumbai, India.
- Asr, F. T. & Demberg, V. (2013). On the information conveyed by discourse markers. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics*, (pp. 84–93).
- 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.
- Ben-David, S., Blitzer, J., Crammer, K., Pereira, F., et al. (2007). Analysis of representations for domain adaptation. In *Neural Information Processing Systems (NIPS)*, Vancouver.
- 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).
- Braud, C. & Denis, P. (2014). Combining natural and artificial examples to improve implicit discourse relation identification. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.
- 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.
Jacob Eisenstein: Distributed Semantics for Automatically Classifying Discourse Relations

References II

- 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)*.
- Jiang, J. & Zhai, C. (2007). Instance weighting for domain adaptation in nlp. In *Proceedings of the Association for Computational Linguistics (ACL)*, Prague.
- 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.
- 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 *NIPS*.
- 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)*.
- 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. & Echihabi, A. (2003). An unsupervised approach to recognizing discourse relations. In *Proceedings of the Association for Computational Linguistics (ACL)*, (pp. 368–375).
- 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).
- 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.
- Polajnar, T., Rimell, L., & Clark, S. (2015). An exploration of discourse-based sentence spaces for compositional distributional semantics. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.
- 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)*.
- 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)*.

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.
- Sporleder, C. & Lascarides, A. (2008). Using automatically labelled examples to classify rhetorical relations: An assessment. *Natural Language Engineering*, 14(3), 369–416.
- 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.