One Vector is Not Enough: Entity-Augmented Distributed Semantics for Discourse Relations

Yangfeng Ji and **Jacob Eisenstein**
School of Interactive Computing
Georgia Institute of Technology
{jiyfeng, jacobe}@gatech.edu

Presenter:
Naoya Inoue
(Tohoku University)

Implicit discourse relation recognition

- Identify *implicit* discourse relation *(=not signaled by discourse connective)* between two discourse segments

(1) *Bob gave Tina the burger. She was hungry.*) REASON

- This work focuses on Penn Discourse Treebank (PDTB)-style structure [Prasad+ 2008]
  - rel(arg1, arg2)
  - c.f. Rhetorical Structure Theory (RST) [Mann & Thompson 1988]; etc.

Research questions

- How do we learn long-tailed bi-lexical relationship?
  - e.g., hungry -- {burger, onigiri, pizza, pasta, steak, …}
  - => Use *vector-based representation* of discourse segments
- How do we represent discourse segment as vector?
  - Recursive composition (e.g., Socher+ 2011)? チッチッチッ:

(1) *Bob gave Tina the burger.*
She was hungry.
) REASON (because)

(2) *Bob gave Tina the burger.*
***He** was hungry.*
) CONTRA-EXPECTATION (although)

  - Segment pairs are superficially similar, but have totally different (opposite) relation…

Idea: entity-centric vector rep.

- Vector of discourse segment pair =

Sentence vectors $\otimes$ Coreferent entity vector
(Previous work) (NEW!)

Discourse segments	(1) *Bob gave Tina the burger. She was hungry.*	(2) *Bob gave Tina the burger. **He** was hungry.*
Sentence vec.	vec(Bob gave Tina the burger) cec(She was hungry)	vec(Bob gave Tina the burger) vec(He was hungry)
Coref. entity vector	vec(Tina got the burger from Bob) vec(Tina was hungry)	vec(Bob gave Tina the burger) vec(Bob was hungry)

The overall framework

- **Given:** two discourse segments m, n
- **Output:** discourse relation y
- Decision function ψ is defined as follows:

$$\psi(y) = \overbrace{\boxed{(\boldsymbol{u}_0^{(m)})^\top \mathbf{A}_y \boldsymbol{u}_0^{(n)}}}^{\text{(a)}} + \overbrace{\boxed{\sum_{i,j \in \mathcal{A}(m,n)} (\boldsymbol{d}_i^{(m)})^\top \mathbf{B}_y \boldsymbol{d}_j^{(n)}}}^{\text{(b)}}$$

$$+ \underbrace{\boxed{\boldsymbol{\beta}_y^\top \boldsymbol{\phi}_{(m,n)} + b_y,}}_{\text{(c)}}$$

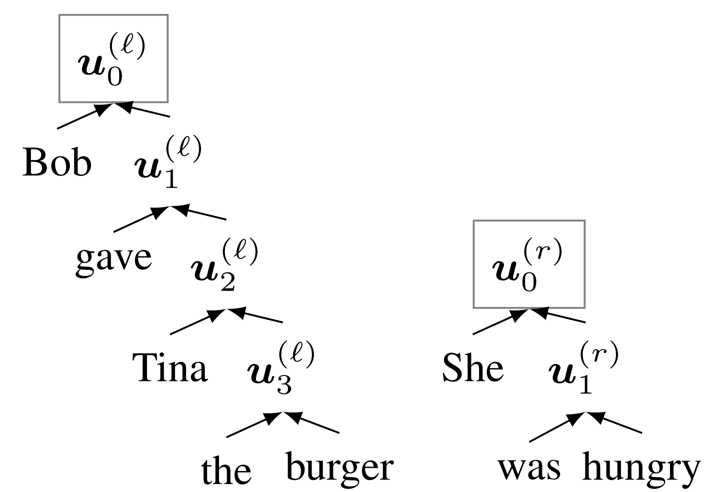
(a) … segment semantics: sentence vectors $u_o^{(m)}$ and $u_o^{(n)}$, parameter $\mathbf{A}_y$
(b) … coref. entity semantics: entity vectors $d_i^{(m)}$ and $d_j^{(n)}$, parameter $\mathbf{B}_y$
(c) … surface features: feature vector $\varphi_{(m,n)}$, parameter $\boldsymbol{\beta}y$

Segment semantics: upward comp.

- Follow Recursive Neural Network-based sentence composition approach [Socher+ 2011]
- Sentence (upward) vector u_{o} is recursively composed over parse tree

$$\boldsymbol{u}_i = \tanh\left(\mathbf{U}[\boldsymbol{u}_{\ell(i)}; \boldsymbol{u}_{r(i)}]\right),$$

$l(i)$: left child of i
$r(i)$: right child of I
$\mathbf{U}$: upward comp. matrix

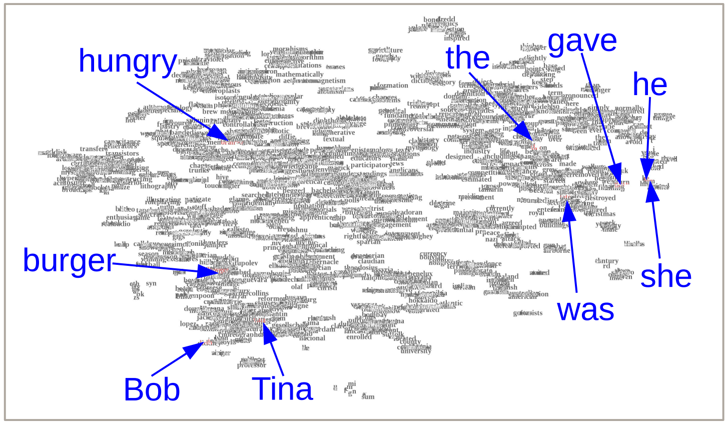


Are we done?

- 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.

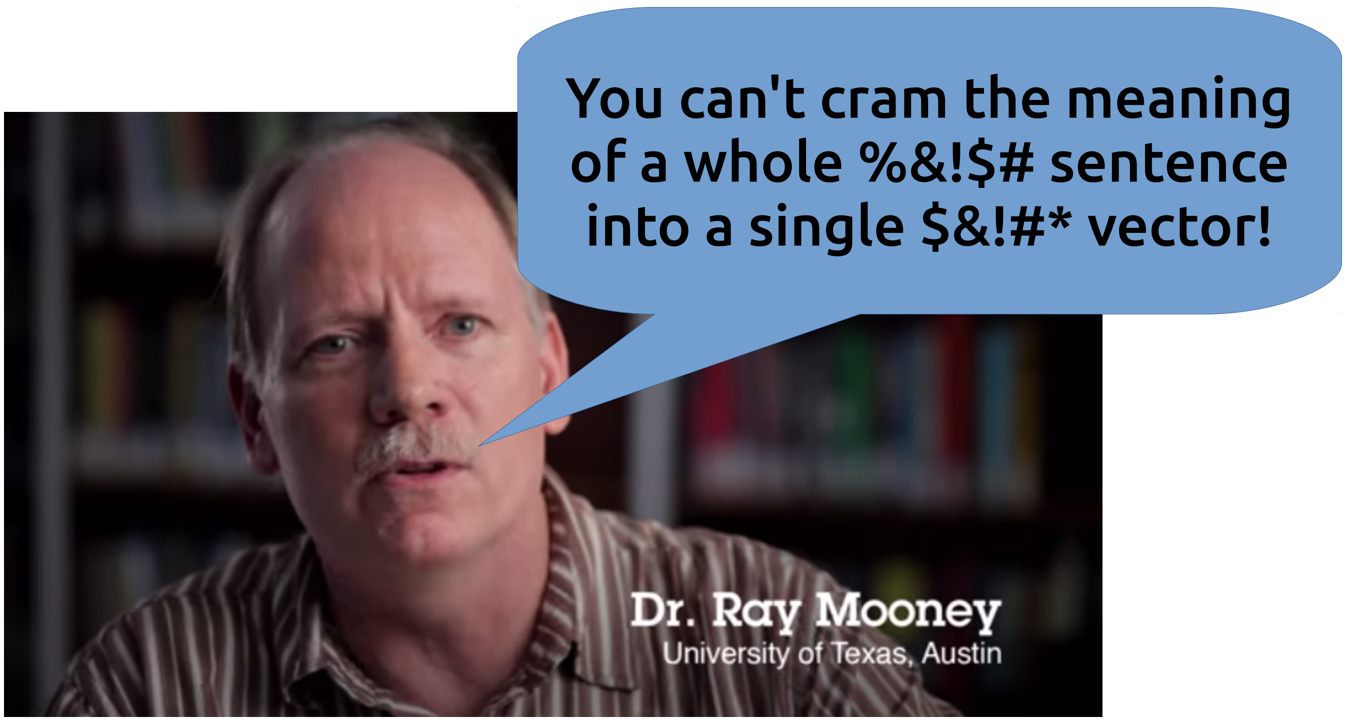


http://www.cc.gatech.edu/~jeisenst/papers/nips-2014-workshop-slides.pdf

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

(1) *Bob gave Tina the burger.*
She was hungry.

Look at things from Tina's perspective:

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

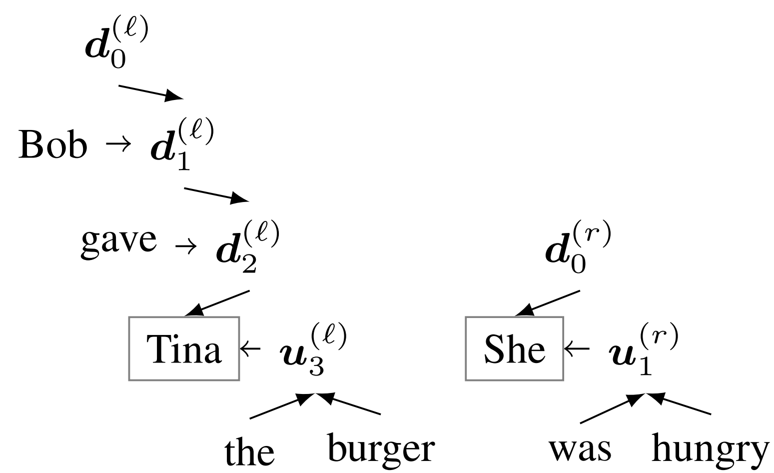
Let's represent these Tina-centric meanings with more vectors!

Entity semantics: downward comp.

- Tracking “roles” played by coreferent entities
- Entity (downward) vector d_i is recursively computed by up-down compositional algorithm based on its parent and sibling

$$\boldsymbol{d}_i = \tanh\left(\mathbf{V}[\boldsymbol{d}_{\rho(i)}; \boldsymbol{u}_{s(i)}]\right)$$

$\rho(i)$: parent of i
$s(i)$: sibling of i
$\mathbf{V}$: downward comp. matrix



Are there so many discourse segment pairs with coreferent entities in PDTB?

Dataset	Annotation	Training (%)	Test (%)
1. PDTB	Automatic	27.4	29.1
2. PDTB∩Onto	Automatic	26.2	32.3
3. PDTB∩Onto	Gold	40.9	49.3

Table 2: Proportion of relations with coreferent entities, according to automatic coreference resolution and gold coreference annotation.

(Coref resolver: Berkeley coreference system [Durrett & Klein 2013])

Learning framework

- Parameter reduction of $\mathbf{A}_y$, $\mathbf{B}_y$
  - $\mathbf{A}_y = \boldsymbol{a}_{y,1}\boldsymbol{a}_{y,2}^\top + \mathrm{diag}(\boldsymbol{a}_{y,3})$. ($|\mathcal{Y}|K^2$ => $|\mathcal{Y}|3K$)
- Large-margin learning framework
  - Learned parameters: $\theta = \theta_{class} \cup \theta_{comp}$
    - $\theta_{class} = \{\mathbf{A}_y, \mathbf{B}_y, \boldsymbol{\beta}_y, b_y\}$
    - $\theta_{comp} = \{\mathbf{U}, \mathbf{V}\}$
  - Objective function [Socher+ 2011]:
    - Minimize regularized hinge loss:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{y':y' \neq y^*} \max\Big(0, 1 - \psi(y^*) + \psi(y')\Big) + \lambda ||\boldsymbol{\theta}||_2^2$$

Experiment

- Dataset
  - Corpus: Penn Discourse Treebank [Prasad+ 2008]
  - Training: sections 2-20, testing: sections 21-22
  - Relations: second-level discourse relations (16 class)
- Learning
  - Learning rate: tuned with AdaGrad [Duchi+ 2011]
  - Initialization: θ_{class} => 0, θ_{comp} => random ([-sqrt(6/2K), sqrt(6/2K)])
- Word rep.
  - word2vec [Mikolov+ 2013]-based vectors trained on PDTB (not updated during learning)
- Parsers
  - Syntactic parser: Stanford parser [Klein & Manning 2003]
  - Coreference: Berkeley coreference system [Durrett & Klein 2013]

Results

Model	+Entity semantics	+Surface features	K	Accuracy(%)
Baseline models				
1. Most common class				26.03
2. Additive word representations			50	28.73
Prior work				
3. (Lin et al., 2009)		✓		$\boxed{40.2}$
Our work				
4. Surface features + Brown clusters		✓		40.66
5. DISCO2			50	36.98
6. DISCO2	✓		50	37.63
7. DISCO2		✓	50	$\boxed{43.75^*}$
8. DISCO2	✓	✓	50	$\boxed{44.59^*}$

* signficantly better than lines 3 and 4 with $p < 0.05$

(a)

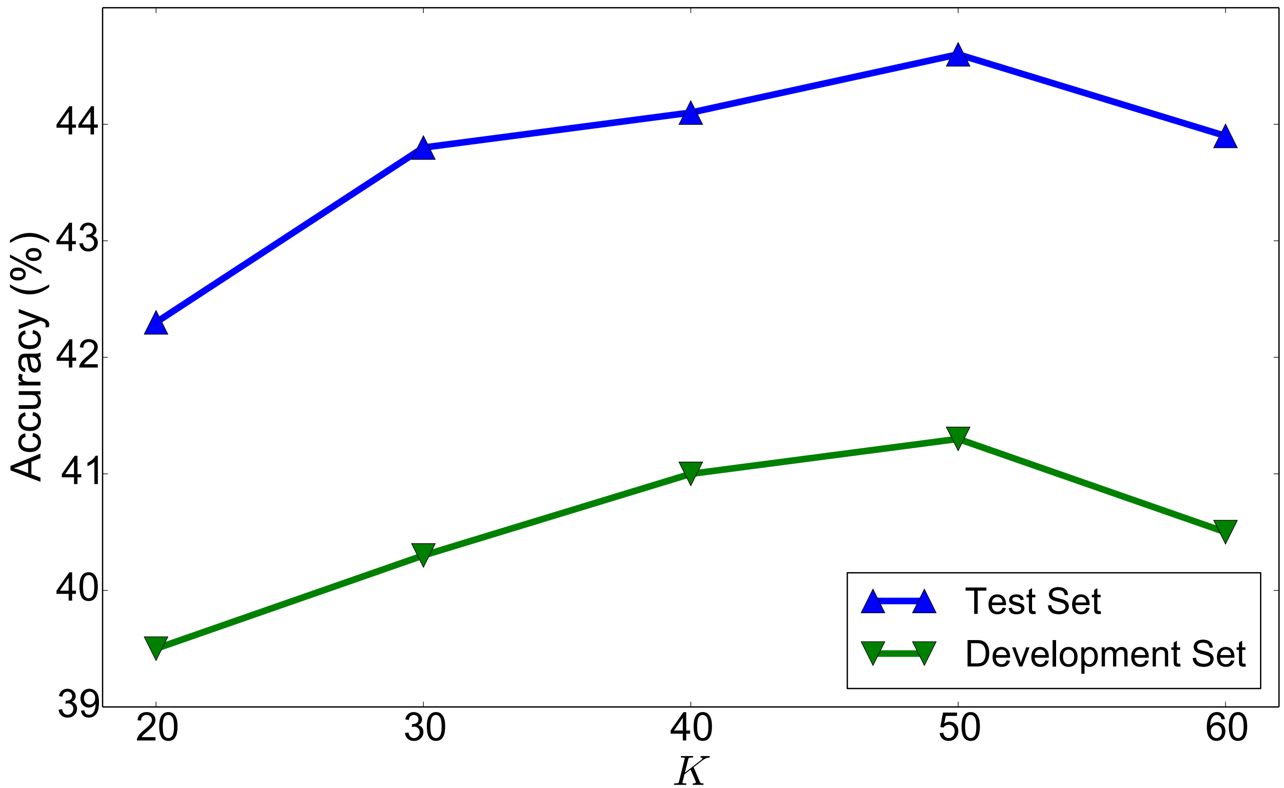
(b)

(a) DISCO2 outperforms state-of-the-art

(b) Coref. entity-centric vector helped

(considering all pairs of NPs: 42.14%)

Sensitivity of K



Improved 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.*

RESTATEMENT
(w/o ent. => CAUSE)

(4) **Arg 1**: *[Mr. Greenberg] got out just before the 1987 crash and, to [his] regret, never went back even as the market soared.*
Arg 2: *This time [he]'s ready to buy in "when the panic wears off."*

CONTRAST
(w/o ent. => CONJUNCTION)

(5) **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.*

CONTRAST
(w/o ent. => CONJUNCTION)

Conclusions

- Vector representation of discourse segment pair needs to be carefully designed

- One vector is not enough; adding entity-centric information leads to significant performance improvement