Distributed Representations beyond the Sentence Level

Jacob Eisenstein

Georgia Institute of Technology

February 2, 2018

How does language shape and reflect the social world?

- Language variation and change
  (Eisenstein et al., 2014; Goel et al., 2016)
- Social meaning
  (Krishnan & Eisenstein, 2015; Pavalanathan et al., 2017)

How to represent linguistic structure and meaning?

- Compositionality in subword representations
  (Bhatia et al., 2016; Pinter et al., 2017)
- Semantic representations for discourse structure
  (Ji & Eisenstein, 2015; Ji et al., 2016)

How does language shape and reflect the social world?

- Language variation and change
  (Eisenstein et al., 2014; Goel et al., 2016)
- Social meaning
  (Krishnan & Eisenstein, 2015; Pavalanathan et al., 2017)

How to represent linguistic structure and meaning?

- Compositionality in subword representations
  (Bhatia et al., 2016; Pinter et al., 2017)
- Semantic representations for discourse structure
  (Ji & Eisenstein, 2015; Ji et al., 2016)

How does language shape and reflect the social world?

- Language variation and change (Eisenstein et al., 2014; Goel et al., 2016)
- Social meaning (Krishnan & Eisenstein, 2015; Pavalanathan et al., 2017)

How to represent linguistic structure and meaning?

- Compositionality in subword representations (Bhatia et al., 2016; Pinter et al., 2017)
- Semantic representations for discourse structure (Ji & Eisenstein, 2015; Ji et al., 2016)

“No computation without representation”

-Fodor (1981)

“No computation without representation”

-Fodor (1981)

- Word embeddings have transformed natural language processing. Natural next step is to try to “embed all the things.”
- **Task-neutral sentence embeddings** could dramatically facilitate learning in settings where labeled data is limited (Conneau et al., 2017).
- Sentence embeddings are a bridge to more holistic understanding of stories, arguments, dialogues, and negotiations.

Embeddings versus formal semantics

The early bird gets the worm.

$$\forall x.(\textsc{early}(x) \wedge \textsc{bird}(x))$$
$$\Rightarrow(\exists y.\textsc{worm}(y) \wedge \textsc{gets}(x, y))$$

(Blackburn & Bos, 2005; Zettlemoyer & Collins, 2005; Liang et al., 2013)

Logical semantic representations:

- expressive
- hard to learn
- what's the right level of abstraction for predicates?

Embeddings versus formal semantics

The early bird gets the worm.

$$\forall x.(\textsc{early}(x) \land \textsc{bird}(x)) \Rightarrow (\exists y.\textsc{worm}(y) \land \textsc{gets}(x, y))$$

(Blackburn & Bos, 2005; Zettlemoyer & Collins, 2005; Liang et al., 2013)

Logical semantic representations:

- expressive
- hard to learn
- what's the right level of abstraction for predicates?

There's a small amount of work bridging the gap between formal and distributed representations (Lewis & Steedman, 2013).

Questions for sentence embeddings

- Are sentence embeddings expressive enough to support the inferences we want to make?
- How should we learn to build sentence embeddings from smaller units of text?
- How should sentence embeddings combine to create meaning across larger units of text?

Multi-sentence discourse structure

These questions are naturally framed within the context of **discourse structure**.

- Centering (Grosz et al., 1995)
- Rhetorical structure theory (Mann & Thompson, 1988)
- Penn Discourse Treebank (Prasad et al., 2008)

Lots of theory, but basic question is the same:
How are adjacent units of text related?

How are adjacent sentences related?

How are adjacent sentences related?

(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.

Why I like this question

1. Discourse puts semantics in context.

The intuition behind "natural language inference" tasks like SNLI:

(2) Barack Obama was born in Hawaii.
Barack Obama was born in the United States.

(3) The early bird gets the worm.
Big Bird woke up early but couldn't find any breakfast.

The intuition behind “natural language inference” tasks like SNLI:

(2) Barack Obama was born in Hawaii.
Barack Obama was born in the United States.

(3) The early bird gets the worm.
Big Bird woke up early but couldn’t find any breakfast.

The formal semantic analyses of all four of these sentences lack free variables.

The reality:

(4) A man inspects the uniform of a figure in some East Asian country.
The man is sleeping. (Bowman et al., 2015)

The reality:

(4) A man inspects the uniform of a figure in some East Asian country.
The man is sleeping. (Bowman et al., 2015)

- Entailment and contradiction are meaningful only for sentences whose semantic analysis lacks free variables.
- SNLI relies on implicit pragmatic intuitions about image captions that do not generalize to arbitrary text “in the wild” (Bowman et al., 2015).

Why I like this question

1. Discourse puts semantics in context.

Why I like this question

1. Discourse puts semantics in context.
2. Discourse relations are fine-grained.

- TEMPORAL
  - Asynchronous
  - Synchronous
    - precedence
    - succession
- CONTINGENCY
  - Cause
    - reason
    - result
  - *Pragmatic Cause*
    - *justification*
  - Condition
    - hypothetical
    - general
    - unreal present
    - unreal past
    - factual present
    - factual past
  - *Pragmatic Condition*
    - *relevance*
    - *implicit assertion*
- COMPARISON
  - Contrast
    - juxtaposition
    - opposition
  - *Pragmatic Contrast*
  - Concession
    - expectation
    - contra-expectation
  - *Pragmatic Concession*
- EXPANSION
  - Conjunction
  - Instantiation
  - Restatement
    - specification
    - equivalence
    - generalization
  - Alternative
    - conjunctive
    - disjunctive
    - chosen alternative
  - Exception
  - List

Why I like this question

1. Discourse puts semantics in context.
2. Discourse relations are fine-grained.

Why I like this question

1. Discourse puts semantics in context.
2. Discourse relations are fine-grained.
3. Lots of applications:
    - How to summarize this article? (Louis et al., 2010)
    - What causal relationships are the author asserting? (Hidey & McKeown, 2016)
    - Is this assertion hypothetical, counterfactual or real? (Son et al., 2017)
    - What is the appropriate translation of this text? (Hardmeier, 2012)
    - What is the appropriate response in this dialogue? (Kalchbrenner & Blunsom, 2013)

The Penn Discourse Treebank Approach

- Hierarchy of a few dozen **discourse relations**.
- Each relation is anchored in a set of discourse connectors, which may be implicit.
- Annotations available in English (Prasad et al., 2008), Chinese (Zhou & Xue, 2012), Arabic (Al-Saif & Markert, 2010), Czech (Poláková et al., 2013), ...

How are adjacent sentences related?

(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.

How are adjacent sentences related?

(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.

How are adjacent sentences related?

(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 discourse relations

The usual recipe?

1. Encode each sentence.
2. Train a network that combines these representations to predict a discourse relation.

Predicting discourse relations

The usual recipe?

1. Encode each sentence.
2. Train a network that combines these representations to predict a discourse relation.

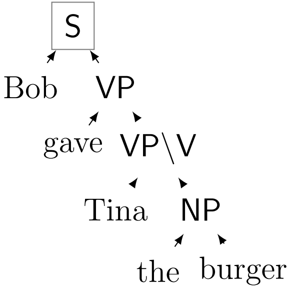
Design decisions:

- **How to represent each sentence?**
- How to train the classifier?

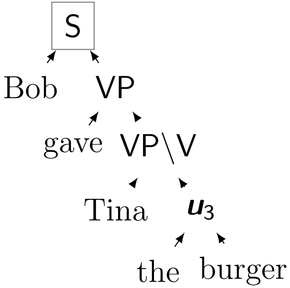
Encoders

- Convolution (Kim, 2014; Kalchbrenner et al., 2014)
- Recurrence (Kiros et al., 2015; Conneau et al., 2017)
- Recursion (Socher et al., 2013)

Vector-semantic composition

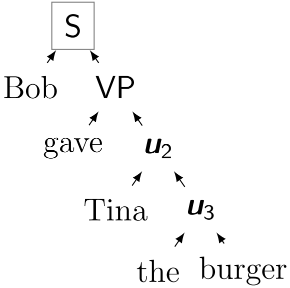


Vector-semantic composition



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

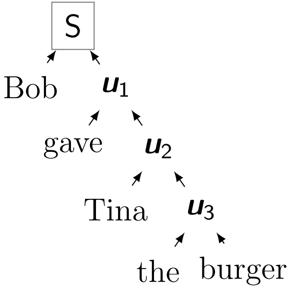
Vector-semantic composition



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

$$\boldsymbol{u}_2 = \tanh\left(\mathbf{U}\left[\boldsymbol{u}_{\text{Tina}}^{\top}\ \boldsymbol{u}_{3}^{\top}\right]^{\top}\right)$$

Vector-semantic composition

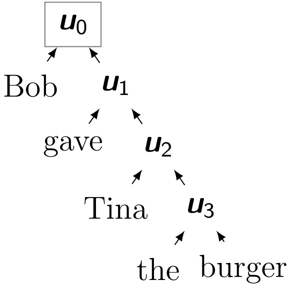


$$\boldsymbol{u}_3 = \tanh\left(\mathbf{U}\left[\boldsymbol{u}_{\text{the}}^\top \; \boldsymbol{u}_{\text{burger}}^\top\right]^\top\right)$$

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

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

Vector-semantic composition



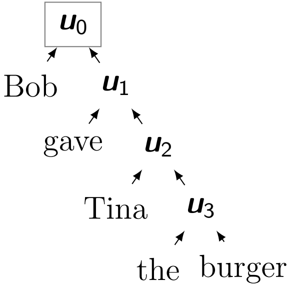
$$\boldsymbol{u}_3 = \tanh\left(\mathbf{U}\left[\boldsymbol{u}_{\text{the}}^\top\ \boldsymbol{u}_{\text{burger}}^\top\right]^\top\right)$$

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

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

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

Vector-semantic composition



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

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

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

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

- DISCO2: **Dis**tributional **co**mpositional semantics for **disco**urse.

Relation prediction

$$\hat{y} = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \qquad (\boldsymbol{u}^{(\ell)})^\top \mathbf{A}_y \boldsymbol{u}^{(r)} + b_y$$

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

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

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

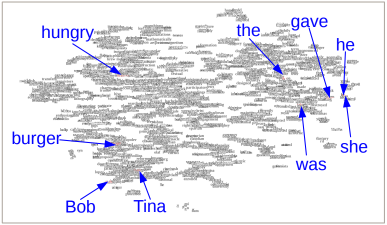
Most common class	26.0
Additive word representations	28.7
Lin et al. (2009)	40.2
SFM: Our reimplementation of Lin et al. (2009)	39.7
SFMB: Lin et al. (2009) + Brown clusters	40.7
Disco2	37.0
Disco2 + SFMB	**43.8**

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.



Encoders

- Convolution (Kim, 2014; Kalchbrenner et al., 2014)
- Recurrence (Kiros et al., 2015; Conneau et al., 2017)
- Recursion (Socher et al., 2013)

Encoders

- Convolution (Kim, 2014; Kalchbrenner et al., 2014)
- Recurrence (Kiros et al., 2015; Conneau et al., 2017)
- Recursion (Socher et al., 2013)

None of these methods are capable of disambiguating the two Bob/Tina cases.

One vector is not enough.

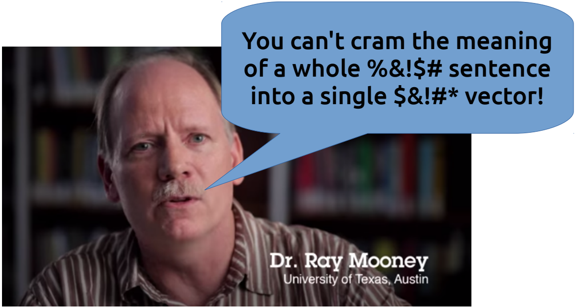
If we insist on representing each discourse argument as a single vector, we lose the ability to track references across the discourse.

As Ray Mooney puts it:

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.

As Ray Mooney puts it:



Entity-augmented distributed semantics

Look at things from Tina's perspective:

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

From Bob's perspective:

- $s1$: He gave Tina the burger.

Entity-augmented distributed semantics

Look at things from Tina's perspective:

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

From Bob's perspective:

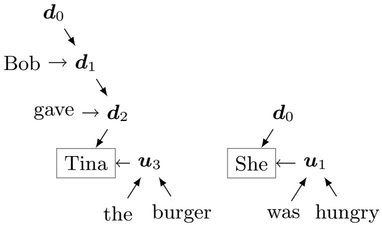
- $s1$: He gave Tina the burger.

A minimal concession to formal semantics (Groenendijk & Stokhof, 1991): Represent these Tina-centric and Bob-centric meanings with more vectors.

Computing entity-centric meanings

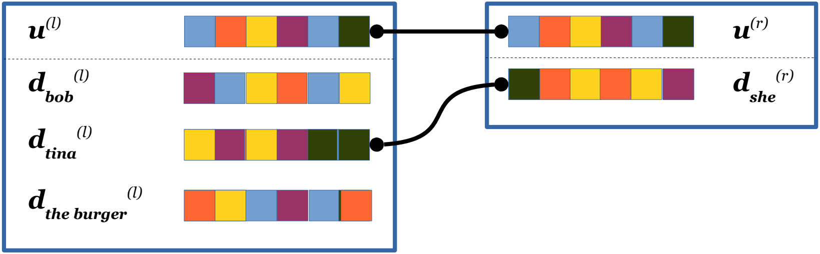
A **downward pass** computes a downward vector for each node in the parse.

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



This computation preserves the feedforward architecture.

A new bilinear model



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

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%.

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

Predicting discourse relations

The usual recipe?

1. Encode each sentence.
2. Train a network that combines these representations to predict a discourse relation.

Design decisions:

- **How to represent each sentence?**
- How to train the classifier?

Predicting discourse relations

The usual recipe?

1. Encode each sentence.
2. Train a network that combines these representations to predict a discourse relation.

Design decisions:

- How to represent each sentence?
- **How to train the classifier?**

How much data?

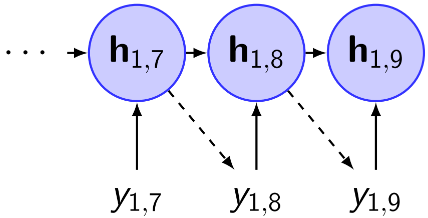
Task	instances
predict next word	10^9 (Chelba et al., 2013)
dependency parsing	10^6 (Nivre et al., 2016)
NL inference	10^5 (Bowman et al., 2015)
discourse relations	10^4 (Prasad et al., 2008)

How much data?

Task	**instances**
predict next word	10^9 (Chelba et al., 2013)
dependency parsing	10^6 (Nivre et al., 2016)
NL inference	10^5 (Bowman et al., 2015)
discourse relations	10^4 (Prasad et al., 2008)

- Annotating discourse relations in full documents is inherently more expensive.
- Can we learn from unlabeled or partially labeled data in a generative model?

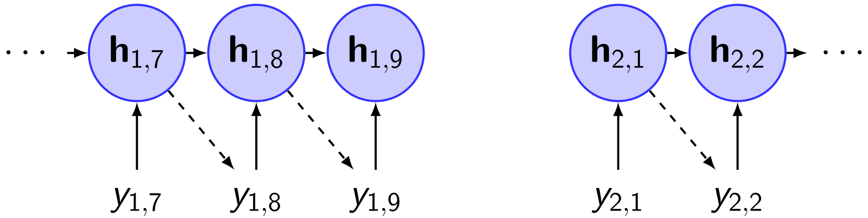
The Discourse Relation Language Model



RNNLM (Mikolov et al 2010)

$$\mathbf{h}_{i,n} = f(E_{y_{i,n}}, U\mathbf{h}_{i,n-1})$$
$$y_{i,n+1} \sim \text{SoftMax}(V\mathbf{h}_{i,n} \qquad)$$

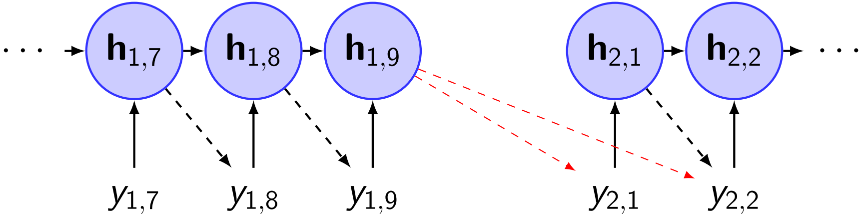
The Discourse Relation Language Model



RNNLM (Mikolov et al 2010)

$$\mathbf{h}_{i,n} = f(E_{y_{i,n}}, U\mathbf{h}_{i,n-1})$$
$$y_{i,n+1} \sim \text{SoftMax}(V\mathbf{h}_{i,n} \qquad)$$

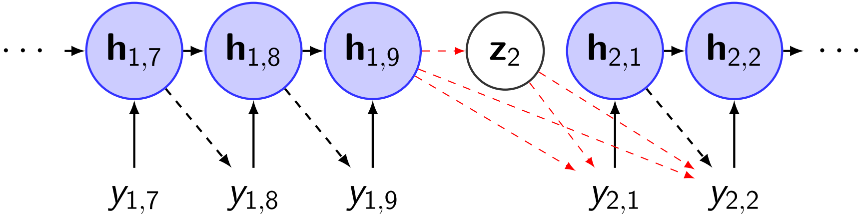
The Discourse Relation Language Model



DCLM (Ji et al 2015)

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

The Discourse Relation Language Model



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

$$\begin{aligned}
\mathbf{h}_{i,n} =& f(E_{y_{i,n}}, U\mathbf{h}_{i,n-1}) \\
y_{i,n+1} \sim& \text{SoftMax}(V^{(z_i)}\mathbf{h}_{i,n} + U^{(z_i)}\mathbf{h}_{i-1,N_{i-1}}) \\
z_i \sim& \text{SoftMax}(\beta \cdot \mathbf{h}_{i-1,N_{i-1}} + \mathbf{b})
\end{aligned}$$

One model, two tasks

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

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

One model, two tasks

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

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

Relation prediction

PDTB Discourse relations (first level)	ACC
Feature-based (Rutherford & Xue, 2015)	57.1
Disco2 (Ji & Eisenstein, 2015)	56.4
DrLM (Ji et al., 2016)	59.5

Dialog act labeling

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

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

Dialog act labeling

Sequential discourse structure on dialogues

(Jurafsky et al., 1997)

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

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`
...	...	...

Relation prediction

PDTB Discourse relations (first level)	ACC
Feature-based (Rutherford & Xue, 2015)	57.1
Disco2 (Ji & Eisenstein, 2015)	56.4
DrLM (Ji et al., 2016)	59.5

Relation prediction

PDTB Discourse relations (first level)	ACC
Feature-based (Rutherford & Xue, 2015)	57.1
Disco2 (Ji & Eisenstein, 2015)	56.4
DrLM (Ji et al., 2016)	59.5
Switchboard dialog acts	
HMM (Stolcke et al., 2000)	71.0
RNN+CNN (Kalchbrenner & Blunsom, 2013)	73.9
DrLM (Ji et al., 2016)	77.0

One model, two tasks

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

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

One model, two tasks

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

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

Language modeling

PDTB (news text)	PPLX
LSTM	117.8
DCLM (Ji et al., 2015)	112.2
DrLM (Ji et al., 2016)	108.3

Language modeling

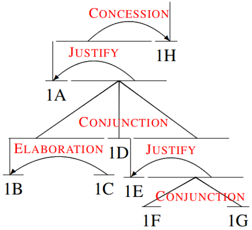
PDTB (news text)	PPLX
LSTM	117.8
DCLM (Ji et al., 2015)	112.2
DrLM (Ji et al., 2016)	108.3
Switchboard (phone transcripts)	
LSTM	56.0
DCLM	45.3
DrLM (Ji et al., 2016)	39.6

Next steps: short term

- Incorporate entity representation into the generative model (Weston et al., 2014; Vinyals et al., 2015).
- Train on unlabeled data
  - marginalize out the discourse relations (Doucet et al., 2000)
  - exploit discourse connectors (Ji et al., 2015)

Next steps: longer term

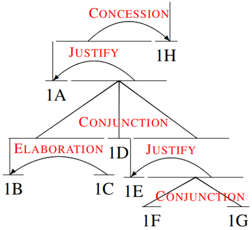
- Holistic models of document structure
  (Bhatia et al., 2015; Liu & Lapata, 2017)



Next steps: longer term

- Holistic models of document structure (Bhatia et al., 2015; Liu & Lapata, 2017)
- Non-entity references

  (7) They said I was too ugly for showbiz. And unfortunately <u>that</u> was true.

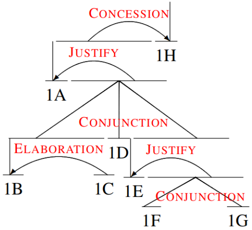


Next steps: longer term

- Holistic models of document structure (Bhatia et al., 2015; Liu & Lapata, 2017)
- Non-entity references

  (7) They said I was too ugly for showbiz. And unfortunately <u>that</u> was true.

- Language and knowledge



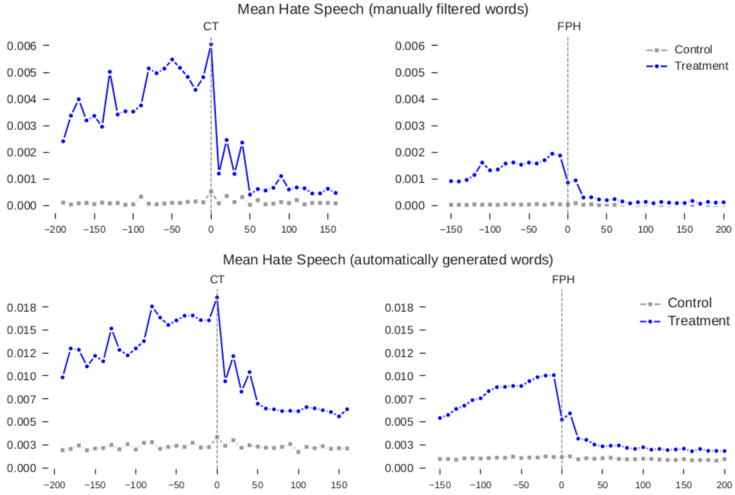
How does language shape and reflect the social world?

- Language variation and change (Eisenstein et al., 2014; Goel et al., 2016)
- Social meaning (Krishnan & Eisenstein, 2015; Pavalanathan et al., 2017)

How to represent linguistic structure and meaning?

- Compositionality in subword representations (Bhatia et al., 2016; Pinter et al., 2017)
- Semantic representations for discourse structure (Ji & Eisenstein, 2015; Ji et al., 2016)

Tracking hate speech on reddit



A day after the paper came out

46.8k

Reddit's bans of r/coontown and r/fatpeoplehate worked--many accounts of frequent posters on those subs were abandoned, and those who stayed reduced their use of hate speech ▸ comp.social.gatech.edu

5 days ago by asbruckman Professor | Interactive Computing x2

6874 comments share save hide give gold report

[–] Hey-Grandan2 349 points 5 days ago

What excacly qualifies for hate speech?

permalink embed save report give gold reply

[–] eegilbert **Author of Article** 652 points 5 days ago

One of the authors here. There was an unsupervised computational process used, documented on pages 6 and 7, and then a supervised human annotation step. Both lexicons are used throughout the rest of work.

permalink embed save parent report give gold reply

[+] ***Comment removed*** *5 days ago* (58 children)*

[–] Laminar_flo 92 points 5 days ago

Ok, adding to that, how did you ensure that the manual filtering process was ideological neutral and not just a reflection of the political sensitivities of the person filtering?

permalink save parent report give gold reply

[–] qwenjwenfljnanq 11 points 5 days ago

But then how did you differentiate between hate speech and people talking *about* hate speech?

permalink save parent report give gold reply

[–] Mode1961 -14 points 5 days ago

> number of words that indicate hate speech

Who choose those words.

permalink save parent report give gold reply

“No computation without representation”

“No computation without representation”

- The success of vector embeddings motivates two orthogonal research directions:
  - push their applications as far as possible;
  - ask “in principle” questions about what vector representations can and cannot do.
- A challenge task: understanding sentence meaning in a discourse context.

Acknowledgments

- **Colleagues**: Yangfeng Ji, Parminder Bhatia, Trevor Cohn, Chris Dyer, Robert Guthrie, Reza Haffari, Lingpeng Kong, Yuval Pinter, Gongbo Zhang
- **Sponsors**: Google and the National Science Foundation

References I

Al-Saif, A. & Markert, K. (2010). The leeds arabic discourse treebank: Annotating discourse connectives for arabic. In *LREC*.

Bhatia, P., Guthrie, R., & Eisenstein, J. (2016). Morphological priors for probabilistic neural word embeddings. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

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)*.

Blackburn, P. & Bos, J. (2005). *Representation and inference for natural language: A first course in computational semantics*. CSLI.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, (pp. 632–642).

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., & Robinson, T. (2013). One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, (pp. 681–691).

Doucet, A., Godsill, S., & Andrieu, C. (2000). On sequential monte carlo sampling methods for bayesian filtering. *Statistics and computing*, *10*(3), 197–208.

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2014). Diffusion of lexical change in social media. *PLoS ONE*, *9*.

Fodor, J. A. (1981). The mind-body problem. *Scientific American*, *244*(1), 114–123.

Goel, R., Soni, S., Goyal, N., Paparrizos, J., Wallach, H., Diaz, F., & Eisenstein, J. (2016). The social dynamics of language change in online networks. In *The International Conference on Social Informatics (SocInfo)*.

Groenendijk, J. & Stokhof, M. (1991). Dynamic predicate logic. *Linguistics and philosophy*, *14*(1), 39–100.

Grosz, B. J., Weinstein, S., & Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, *21*(2), 203–225.

References II

Hardmeier, C. (2012). Discourse in statistical machine translation. a survey and a case study. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (11).

Hidey, C. & McKeown, K. (2016). Identifying causal relations using parallel wikipedia articles. In *Proceedings of the Association for Computational Linguistics (ACL)*.

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

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)*.

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)*.

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.

Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A convolutional neural network for modelling sentences. In *Proceedings of the Association for Computational Linguistics (ACL)*, (pp. 655–665).

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, (pp. 1746–1751).

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)*.

References III

Krishnan, V. & Eisenstein, J. (2015). "You're Mr. Lebowski, I'm The Dude": Inducing address term formality in signed social networks. In *NAACL*.

Lewis, M. & Steedman, M. (2013). Combined distributional and logical semantics. *Transactions of the Association for Computational Linguistics*, *1*, 179–192.

Liang, P., Jordan, M. I., & Klein, D. (2013). Learning dependency-based compositional semantics. *Computational Linguistics*, *39*(2), 389–446.

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).

Liu, Y. & Lapata, M. (2017). Learning structured text representations. *Transactions of the ACL*.

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.

Mann, W. C. & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, *8*(3), 243–281.

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

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., & Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. In Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., & Piperidis, S. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Pavalanathan, U., Fitzpatrick, J., Kiesling, S. F., & Eisenstein, J. (2017). A multidimensional lexicon for interpersonal stancetaking. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Pinter, Y., Guthrie, R., & Eisenstein, J. (2017). Mimicking word embeddings using subword RNNs. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

References IV

Poláková, L., Mírovský, J., Nedoluzhko, A., Jínová, P., Zikánová, v., & Hajičová, E. (2013). Introducing the prague discourse treebank 1.0. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, (pp. 91–99)., Nagoya, Japan. Asian Federation of Natural Language Processing.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., & Webber, B. (2008). The Penn Discourse Treebank 2.0. In *Proceedings of LREC*.

Rutherford, A. & Xue, N. (2015). Improving the inference of implicit discourse relations via classifying explicit discourse connectives. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, (pp. 799–808).

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)*.

Son, Y., Buffone, A., Raso, J., Larche, A., Janocko, A., Zembroski, K., Schwartz, H. A., & Ungar, L. (2017). Recognizing counterfactual thinking in social media texts. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van Ess-Dykema, C., & Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, *26*(3), 339–373.

Vinyals, O., Fortunato, M., & Jaitly, N. (2015). Pointer networks. In *Neural Information Processing Systems (NIPS)*, (pp. 2692–2700).

Weston, J., Chopra, S., & Bordes, A. (2014). Memory networks. *arXiv preprint arXiv:1410.3916*.

Zettlemoyer, L. S. & Collins, M. (2005). Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *Proceedings of UAI*.

Zhou, Y. & Xue, N. (2012). Pdtb-style discourse annotation of chinese text. In *Proceedings of the Association for Computational Linguistics (ACL)*, (pp. 69–77).