Distributed Semantics for Automatically Classifying Discourse Relations

Jacob Eisenstein

Georgia Institute of Technology

May 26, 2016
(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.
(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.
(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.
(1) The more people you love, the weaker you are.
(Expansion) You’ll do things for them that you know you shouldn’t do.
(Expansion) You’ll act the fool to make them happy, to keep them safe.
(Contingency) Love no one but your children.
[Contingency] a mother has no choice.
Application: summarization

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

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

(2) The more people you love, the weaker you are.

(Expansion) You’ll do things for them that you know you shouldn’t do.

(Expansion) You’ll act the fool to make them happy, to keep them safe.

(Contingency) Love no one but your children.

(Contingency) On that front, a mother has no choice.

Discourse structure guides the selection of extracts for summaries (Marcu, 1999; Louis et al., 2010; Hirao et al., 2013).
The federal budget should be an honest blueprint for the spending priorities of the government. However, this budget is dishonest.

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

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

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

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

〈love, choice〉, 〈children, mother〉, 〈no, no〉, . . .

But bilexical features are sparse and noisy, and discourse-annotated datasets are inherently small.
Can distributed semantics help?

**Distributed semantics** proposes to capture meaning in dense numerical vectors.

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

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

Jacob Eisenstein: Distributed Semantics for Automatically Classifying Discourse Relations
Country and Capital Vectors Projected by PCA

Jacob Eisenstein: Distributed Semantics for Automatically Classifying Discourse Relations
Distributed semantics beyond the lexicon?
Can we build distributed representations of multi-word linguistic units?

Vector-semantic composition has been applied to:

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

Key questions:

▶ What should distributed representations of discourse units look like?
▶ How should we learn them?
▶ How to apply distributed representations to discourse relation detection and parsing?
Project 1: RST Parsing

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

- **Goal**: rhetorical structure theory parsing

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

At each point, the parser can:

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

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

▶ Elementary discourse units:

\[ u( \text{Love no one but your children} ) = u_{\text{love}} + u_{\text{no}} + \ldots \]

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

- **Elementary discourse units:**

  \[ u(\text{Love no one but your children}) = u_{\text{love}} + u_{\text{no}} + \ldots \]

  “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

<table>
<thead>
<tr>
<th></th>
<th>Span</th>
<th>Nuclearity</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotator agreement</td>
<td>88.7</td>
<td>77.7</td>
<td>65.8</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Span</th>
<th>Nuclearity</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotator agreement</td>
<td>88.7</td>
<td>77.7</td>
<td>65.8</td>
</tr>
<tr>
<td>HILDA (Hernault, Prendinger, duVerle &amp; Ishizuka, Hernault et al.)</td>
<td>83.0</td>
<td>68.4</td>
<td>54.8</td>
</tr>
<tr>
<td>TSP (Joty et al., 2013)</td>
<td>82.7</td>
<td>68.4</td>
<td>55.7</td>
</tr>
<tr>
<td>“Basic features”</td>
<td>79.4</td>
<td>68.0</td>
<td>53.0</td>
</tr>
</tbody>
</table>

On discourse relations, the distributed representation cuts the gap between SOTA and inter-annotator agreement by 60%!
<table>
<thead>
<tr>
<th>Method</th>
<th>Span</th>
<th>Nuclearity</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotator agreement</td>
<td>88.7</td>
<td>77.7</td>
<td>65.8</td>
</tr>
<tr>
<td>HILDA (Hernault, Prendinger, duVerle &amp; Ishizuka, Hernault et al.)</td>
<td>83.0</td>
<td>68.4</td>
<td>54.8</td>
</tr>
<tr>
<td>TSP (Joty et al., 2013)</td>
<td>82.7</td>
<td>68.4</td>
<td>55.7</td>
</tr>
<tr>
<td>“Basic features”</td>
<td>79.4</td>
<td>68.0</td>
<td>53.0</td>
</tr>
<tr>
<td>Distributed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Collobert &amp; Weston</td>
<td>75.3</td>
<td>67.1</td>
<td>53.8</td>
</tr>
<tr>
<td>Non-neg. matrix factorization</td>
<td>78.6</td>
<td>67.7</td>
<td>54.8</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Span</th>
<th>Nuclearity</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotator agreement</td>
<td>88.7</td>
<td>77.7</td>
<td>65.8</td>
</tr>
<tr>
<td>HILDA (Hernault, Prendinger, duVerle &amp; Ishizuka, Hernault et al.)</td>
<td>83.0</td>
<td>68.4</td>
<td>54.8</td>
</tr>
<tr>
<td>TSP (Joty et al., 2013)</td>
<td>82.7</td>
<td>68.4</td>
<td>55.7</td>
</tr>
<tr>
<td>“Basic features”</td>
<td>79.4</td>
<td>68.0</td>
<td>53.0</td>
</tr>
</tbody>
</table>

**Distributed**

<table>
<thead>
<tr>
<th></th>
<th>Span</th>
<th>Nuclearity</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collobert &amp; Weston</td>
<td>75.3</td>
<td>67.1</td>
<td>53.8</td>
</tr>
<tr>
<td>Non-neg. matrix factorization</td>
<td>78.6</td>
<td>67.7</td>
<td>54.8</td>
</tr>
</tbody>
</table>
## RST Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Span</th>
<th>Nuclearity</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotator agreement</td>
<td>88.7</td>
<td>77.7</td>
<td>65.8</td>
</tr>
<tr>
<td>HILDA (Hernault, Prendinger, duVerle &amp; Ishizuka, Hernault et al.)</td>
<td>83.0</td>
<td>68.4</td>
<td>54.8</td>
</tr>
<tr>
<td>TSP (Joty et al., 2013)</td>
<td>82.7</td>
<td>68.4</td>
<td>55.7</td>
</tr>
<tr>
<td>“Basic features”</td>
<td>79.4</td>
<td>68.0</td>
<td>53.0</td>
</tr>
</tbody>
</table>

**Distributed**

<table>
<thead>
<tr>
<th>Method</th>
<th>Span</th>
<th>Nuclearity</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collobert &amp; Weston</td>
<td>75.3</td>
<td>67.1</td>
<td>53.8</td>
</tr>
<tr>
<td>Non-neg. matrix factorization</td>
<td>78.6</td>
<td>67.7</td>
<td>54.8</td>
</tr>
<tr>
<td>Distributed</td>
<td>80.9</td>
<td>69.4</td>
<td>59.0</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Span</th>
<th>Nuclearity</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotator agreement</td>
<td>88.7</td>
<td>77.7</td>
<td>65.8</td>
</tr>
<tr>
<td>HILDA (Hernault, Prendinger, duVerle &amp; Ishizuka, Hernault et al.)</td>
<td>83.0</td>
<td>68.4</td>
<td>54.8</td>
</tr>
<tr>
<td>TSP (Joty et al., 2013)</td>
<td>82.7</td>
<td>68.4</td>
<td>55.7</td>
</tr>
<tr>
<td>“Basic features”</td>
<td>79.4</td>
<td>68.0</td>
<td>53.0</td>
</tr>
<tr>
<td><strong>Distributed</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Collobert &amp; Weston</td>
<td>75.3</td>
<td>67.1</td>
<td>53.8</td>
</tr>
<tr>
<td>Non-neg. matrix factorization</td>
<td>78.6</td>
<td>67.7</td>
<td>54.8</td>
</tr>
<tr>
<td>Distributed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+basic features</td>
<td>82.1</td>
<td><strong>71.1</strong></td>
<td><strong>61.6</strong></td>
</tr>
</tbody>
</table>

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

Jacob Eisenstein: Distributed Semantics for Automatically Classifying Discourse Relations
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).

\[ \langle \text{fun, buildings} \rangle, \langle \text{place, interesting} \rangle, \ldots \]
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.

Jacob Eisenstein: Distributed Semantics for Automatically Classifying Discourse Relations
Vector-semantic composition

Bob gave Tina the burger

u_3 = \tanh(U[u^\top \text{burger}]^\top)

u_2 = \tanh(U[u^\top \text{Tina}]^\top)

u_1 = \tanh(U[u^\top \text{gave}]^\top)

u_0 = \tanh(U[u^\top \text{Bob}]^\top)

▶ Disco2: Distributional compositional semantics for discourse.
▶ Same architecture as Socher, Huang, Pennington, Ng & Manning (Socher et al.).

Jacob Eisenstein: Distributed Semantics for Automatically Classifying Discourse Relations
vector-semantic composition

$u_3 = \tanh \left( U \begin{bmatrix} u_{the}^\top & u_{burger}^\top \end{bmatrix}^\top \right)$
Vector-semantic composition

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

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

Bob \hspace{1cm} gave \hspace{1cm} Tina

\[ u_2 \]

\[ \text{the} \hspace{1cm} \text{burger} \]

Jacob Eisenstein: Distributed Semantics for Automatically Classifying Discourse Relations
Vector-semantic composition

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

$u_2 = \tanh \left( U \left[ u_{\text{Tina}}^\top u_{\text{3}}^\top \right]^\top \right)$

$u_1 = \tanh \left( U \left[ u_{\text{gave}}^\top u_{\text{2}}^\top \right]^\top \right)$
Vector-semantic composition

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

\[ u_2 = \tanh \left( U \left[ u_{\text{Tina}} \ u_{\text{3}} \right]^\top \right) \]

\[ u_1 = \tanh \left( U \left[ u_{\text{gave}} \ u_{\text{2}} \right]^\top \right) \]

\[ u_0 = \tanh \left( U \left[ u_{\text{Bob}} \ u_{\text{1}} \right]^\top \right) \]
**Vector-semantic composition**

\[
\begin{align*}
\mathbf{u}_0 &= \text{tanh} \left( \mathbf{U} \left[ \mathbf{u}^\top_{\text{Bob}} \mathbf{u}^\top_1 \right]^\top \right) \\
\mathbf{u}_1 &= \text{tanh} \left( \mathbf{U} \left[ \mathbf{u}^\top_{\text{gave}} \mathbf{u}^\top_2 \right]^\top \right) \\
\mathbf{u}_2 &= \text{tanh} \left( \mathbf{U} \left[ \mathbf{u}^\top_{\text{Tina}} \mathbf{u}^\top_3 \right]^\top \right) \\
\mathbf{u}_3 &= \text{tanh} \left( \mathbf{U} \left[ \mathbf{u}^\top_{\text{the burger}} \mathbf{u}^\top_3 \right]^\top \right)
\end{align*}
\]

- **Disco2**: Distributional compositional semantics for discourse.
- Same architecture as Socher, Huang, Pennington, Ng & Manning (Socher et al.).

Jacob Eisenstein: Distributed Semantics for Automatically Classifying Discourse Relations
A bilinear model

\[ \hat{y} = \arg \max_{y \in Y} (u^{(\ell)})^\top A_y u^{(r)} + b_y \]

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

\[ A_y = a_{y,1} a_{y,2}^\top + \text{diag}(a_{y,3}). \]
Learning

- Word representations are fixed to \textsc{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

<table>
<thead>
<tr>
<th>Method</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most common class</td>
<td>26.0</td>
</tr>
<tr>
<td>Additive word representations</td>
<td>28.7</td>
</tr>
</tbody>
</table>

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

Jacob Eisenstein: Distributed Semantics for Automatically Classifying Discourse Relations
## PDTB Results

<table>
<thead>
<tr>
<th>Most common class</th>
<th>26.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Additive word representations</td>
<td>28.7</td>
</tr>
<tr>
<td>Lin et al. (2009)</td>
<td>40.2</td>
</tr>
<tr>
<td>$\text{SFM}$: Our reimplementation of Lin et al. (2009)</td>
<td>39.7</td>
</tr>
<tr>
<td>$\text{SFM}_B$: Lin et al. (2009) + Brown clusters</td>
<td>40.7</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most common class</td>
<td>26.0</td>
</tr>
<tr>
<td>Additive word representations</td>
<td>28.7</td>
</tr>
<tr>
<td>Lin et al. (2009)</td>
<td>40.2</td>
</tr>
<tr>
<td><strong>SFM</strong>: Our reimplementation of Lin et al. (2009)</td>
<td>39.7</td>
</tr>
<tr>
<td><strong>SFM-B</strong>: Lin et al. (2009) + Brown clusters</td>
<td>40.7</td>
</tr>
<tr>
<td>Disco2</td>
<td>37.0</td>
</tr>
<tr>
<td>Disco2 + <strong>SFM-B</strong></td>
<td>43.8</td>
</tr>
</tbody>
</table>

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

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

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

Jacob Eisenstein: Distributed Semantics for Automatically Classifying Discourse Relations
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...

You can't cram the meaning of a whole %&!$# sentence into a single $&!#* vector!
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.

\[ d_i = \tanh \left( \mathbf{V} \left[ \begin{array}{c} d_{\rho(i)} \\ u_{s(i)} \end{array} \right] \right) \]

This computation preserves the feedforward architecture.
A new bilinear model

\[ \hat{y} = \arg \max_{y \in \mathcal{Y}} (u^{(\ell)})^\top A_y u^{(r)} + \sum_{\langle i, j \rangle \in \mathcal{A}} (d_i^{(\ell)})^\top B_y d_j^{(r)} + b_y \]

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

Jacob Eisenstein: Distributed Semantics for Automatically Classifying Discourse Relations
### 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 |

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most common class</td>
<td>26.0</td>
</tr>
<tr>
<td>Additive word representations</td>
<td>28.7</td>
</tr>
<tr>
<td>Lin et al. (2009)</td>
<td>40.2</td>
</tr>
<tr>
<td><strong>SFM</strong>: Our reimplementation of Lin et al. (2009)</td>
<td>39.7</td>
</tr>
<tr>
<td><strong>SFM-B</strong>: Lin et al. (2009) + Brown clusters</td>
<td>40.7</td>
</tr>
<tr>
<td>Disco2</td>
<td>37.0</td>
</tr>
<tr>
<td>Disco2 + SFMB</td>
<td>43.8</td>
</tr>
<tr>
<td>Disco2 + SFMB + entity semantics</td>
<td><strong>44.6</strong></td>
</tr>
</tbody>
</table>
## PDTB Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most common class</td>
<td>26.0</td>
</tr>
<tr>
<td>Additive word representations</td>
<td>28.7</td>
</tr>
<tr>
<td>Lin et al. (2009)</td>
<td>40.2</td>
</tr>
<tr>
<td><strong>SFM:</strong> Our reimplementation of Lin et al. (2009)</td>
<td>39.7</td>
</tr>
<tr>
<td><strong>SFM-B:</strong> Lin et al. (2009) + Brown clusters</td>
<td>40.7</td>
</tr>
<tr>
<td>Disco2</td>
<td>37.0</td>
</tr>
<tr>
<td>Disco2 + SFMB</td>
<td>43.8</td>
</tr>
<tr>
<td>Disco2 + SFMB + entity semantics</td>
<td><strong>44.6</strong></td>
</tr>
</tbody>
</table>

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

Jacob Eisenstein: Distributed Semantics for Automatically Classifying Discourse Relations
Project 3: A Generative Model

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

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

RNNLM (Mikolov et al 2010)

\[ h_{i,n} = f(E_{y_{i,n}} + U h_{i,n-1}) \]  
\[ y_{i,n+1} \sim \text{SoftMax}(V h_{i,n}) \]
The Discourse Relation Language Model

RNNLM (Mikolov et al 2010)

\[ h_{i,n} = f(E_{y_{i,n}} + Uh_{i,n-1}) \]  \hspace{1cm} (1)

\[ y_{i,n+1} \sim \text{SoftMax}(Vh_{i,n}) \]  \hspace{1cm} (2)

\[ z_{i} \sim \text{SoftMax}(\beta \cdot h_{i,n-1} + b) \]  \hspace{1cm} (3)
DCLM (Ji et al 2015)

\begin{align*}
\mathbf{h}_{i,n} &= f(E_{y_{i,n}} + U\mathbf{h}_{i,n-1}) \quad (1) \\
y_{i,n+1} &\sim \text{SoftMax}(V\mathbf{h}_{i,n} + Uh_{i-1,i,N_{i-1}}) \quad (2) \\
\end{align*}
The Discourse Relation Language Model

\[ h_{i,n} = f(E_{y_{i,n}} + U_{h_{i,n-1}}) \]  \hspace{1cm} (1)

\[ y_{i,n+1} \sim \text{SoftMax}(V^{(z_i)}_{h_{i,n}} + U^{(z_i)}_{h_{i-1,N_{i-1}}}) \]  \hspace{1cm} (2)

\[ z_i \sim \text{SoftMax}(\beta \cdot h_{i-1,N_{i-1}} + b) \]  \hspace{1cm} (3)

Jacob Eisenstein: Distributed Semantics for Automatically Classifying Discourse Relations
One Model, Two Tasks

Discourse-driven language modeling

\[ p(y_{i,n+1} \mid y_{i,1:n}, y_{i-1}) = \sum_{z_i} p(y_{i,n+1}, z_i \mid y_{i,1:n}, y_{i-1}) \]

Discourse relation prediction

\[ p(z_i \mid y_i, y_{i-1}) = \frac{p(z_i, y_i \mid y_{i-1})}{\sum_{z'} p(z', y_i \mid y_{i-1})} \]
One Model, Two Tasks

Discourse-driven language modeling

\[ p(y_{i,n+1} \mid y_{i,1:n}, y_{i-1}) = \sum_{z_i} p(y_{i,n+1}, z_i \mid y_{i,1:n}, y_{i-1}) \]

Discourse relation prediction

\[ p(z_i \mid y_i, y_{i-1}) = \frac{p(z_i, y_i \mid y_{i-1})}{\sum_{z'} p(z', y_i \mid y_{i-1})} \]
Figure: Evaluation on the first-level implicit discourse relation identification in the PDTB.
One Model, Two Tasks

Discourse-driven language modeling

\[ p(y_{i,n+1} \mid y_{i,1:n}, y_{i-1}) = \sum_{z_i} p(y_{i,n+1}, z_i \mid y_{i,1:n}, y_{i-1}) \]

Discourse relation prediction

\[ p(z_i \mid y_i, y_{i-1}) = \frac{p(z_i, y_i \mid y_{i-1})}{\sum_{z'} p(z', y_i \mid y_{i-1})} \]
Discourse-driven language modeling

\[ p(y_i, n+1 \mid y_{i,1:n}, y_{i-1}) = \sum_{z_i} p(y_i, n+1, z_i \mid y_{i,1:n}, y_{i-1}) \]

Discourse relation prediction

\[ p(z_i \mid y_i, y_{i-1}) = \frac{p(z_i, y_i \mid y_{i-1})}{\sum_{z'} p(z', y_i \mid y_{i-1})} \]
Discourse-driven Language Modeling

Figure: Language modeling on the PDTB
Dialogue act labeling

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

Utterance$_i$ — Dialog_act — Utterance$_{i+1}$
Dialogue act labeling

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

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Dialog Act</th>
<th>Utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td><strong>Yes-No-Question</strong></td>
<td>So do you go to college right now?</td>
</tr>
<tr>
<td>B</td>
<td><strong>Yes-Answer</strong></td>
<td>Yeah,</td>
</tr>
<tr>
<td>B</td>
<td><strong>Statement</strong></td>
<td>it’s my last year</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Jacob Eisenstein: Distributed Semantics for Automatically Classifying Discourse Relations
Dialog act tagging

Figure: Dialog act tagging on the Switchboard Dialog Act Corpus (Stolcke et al., 2000).
Discourse-driven Language Modeling

Figure: Language modeling on the Switchboard Dialog Act Corpus

Jacob Eisenstein: Distributed Semantics for Automatically Classifying Discourse Relations
Linking discourse and semantics

Annotating semantics is hard! Maybe we should give up (Clarke et al., 2010; Artzi & Zettlemoyer, 2011; Berant et al., 2013).

In comparison, annotating and predicting discourse relations is relatively easy.

Can discourse structure be learned from downstream tasks?
Next steps

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

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

- Discourse-coherent machine translation (long live JSALT 2015!)


Jacob Eisenstein: Distributed Semantics for Automatically Classifying Discourse Relations
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
(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