Tailoring Continuous Word Representations for Dependency Parsing

Mohit Bansal, Kevin Gimpel, and Karen Livescu
TTI-Chicago
Questions We Want to Answer

- What kind of embeddings will help dependency parsing (in-domain and out-of-domain)?
- How can we convert embeddings to parsing features?
- Are there good intrinsic measures of embedding quality?
Representation Models

- BROWN (Brown et al., 1992)
- SENNA (Collobert et al., 2011, 2008)
- TURIAN (Turian et al., 2010)
- HUANG (Huang et al., 2012)
- SKIP (Mikolov et al., 2013)
SKIP

INPUT  PROJECTION  OUTPUT

w(t)  →  Projection

w(t)  →  w(t-2)

w(t)  →  w(t-1)

w(t)  →  w(t+1)

w(t)  →  w(t+2)

Few mins. vs. days/weeks/months!!

Mikolov et al., 2013
Syntactically Tailored Embeddings

- Context window size (SKIP)
  - Smaller window $\rightarrow$ syntactic/functional similarity
  - Larger window $\rightarrow$ topical similarity

The morning *flight* at the JFK *airport* was *delayed*

- Similar effect in distributional representations (Lin and Wu, 2009)
Syntactically Tailored Embeddings

- Syntactic context (\(\text{SKIP}_{\text{DEP}}\))
  - Condition on dependency context instead of linear
  - First parse a large corpus with baseline parser:

... said that the regulation of safety is ...

(grandparent) (parent) (child)
Syntactically Tailored Embeddings

- Syntactic context (\(\text{SKIP}_{\text{DEP}}\))
  - Condition on dependency context instead of linear
  - Then convert each dependency to a tuple:
    
    \[
    \begin{bmatrix}
    \text{dep label} & \text{grandparent} & \text{parent} & \text{child} & \text{dep label} \\
    \text{PMOD}_{<L>} & \text{regulation}_{<G>} & \text{of} & \text{safety} & \text{PMOD}_{<L>} \\
    \end{bmatrix}
    \]

- Syntactic information in clustering, topic, semantic space models
  (Sagae and Gordon, 2009; Haffari et al., 2011; Grave et al., 2013; Boyd-Graber and Blei, 2008; Pado and Lapata, 2007)
Cluster Examples

- SKIP, \( w = 10 \):

  - [attendant, takeoff, airport, carry-on, airplane, flown, landings, flew, fly, cabins, ...]
  - [maternity, childbirth, clinic, physician, doctor, medical, health-care, day-care, ...]
  - [transactions, equity, investors, capital, financing, stock, fund, purchases, ...]
Cluster Examples

- SKIP, $w = 1$

  - [Mr., Mrs., Ms., Prof., III, Jr., Dr.]
  - [Jeffrey, William, Dan, Robert, Stephen, Peter, John, Richard, ...]
  - [Portugal, Iran, Cuba, Ecuador, Greece, Thailand, Indonesia, ...]

  - [his, your, her, its, their, my, our]
  - [Your, Our, Its, My, His, Their, Her]
  - [truly, wildly, politically, financially, completely, potentially, ...]
## Intrinsic Evaluation

(Finkelstein et al., 2002)

<table>
<thead>
<tr>
<th>Representation</th>
<th>SIM</th>
<th>TAG</th>
</tr>
</thead>
<tbody>
<tr>
<td>BROWN</td>
<td>–</td>
<td>89.3</td>
</tr>
<tr>
<td>SENNA</td>
<td>49.8</td>
<td>85.2</td>
</tr>
<tr>
<td>HUANG</td>
<td>62.6</td>
<td>78.1</td>
</tr>
<tr>
<td>SKIP, $w = 10$</td>
<td>44.6</td>
<td>71.5</td>
</tr>
<tr>
<td>SKIP, $w = 5$</td>
<td>44.4</td>
<td>81.1</td>
</tr>
<tr>
<td>SKIP, $w = 1$</td>
<td>37.8</td>
<td>86.6</td>
</tr>
<tr>
<td>SKIP$_{DEP}$</td>
<td>34.6</td>
<td>88.3</td>
</tr>
</tbody>
</table>

- **Topical**
- **Syntactic/Functional**
Dependency Parsing Features

- Brown Cluster Features (Koo et al., 2008):

```
000 00 0 011 10 11
apple pear Apple IBM bought run of in
```

- `apple` \(\rightarrow\) 00010100010
  - prefix4
  - prefix6

- `prefix6` \(\rightarrow\) 110010
- `prefix4` \(\rightarrow\) 1100
- `tag` \(\rightarrow\) VBD
  - `ate` (parent)
  - `apple` (child)
Dependency Parsing Features

- **Continuous Representation Features:**
  - Per-dimension bucket features:

  \[
  \begin{align*}
  \text{ate} & \rightarrow [0.2, 0.7, -0.6, 0.9] \\
  \text{apple} & \rightarrow [0.6, -0.1, 0.7, 0.2]
  \end{align*}
  \]

- **Hierarchical clustering (bit string) features:**

  \[
  \text{linkage}(E, \text{‘ward’}, \text{‘euclidean’})
  \]

  \[
  \begin{align*}
  \text{apple} & \rightarrow 00010100010 \\
  \text{prefix4} & \\
  \text{prefix6}
  \end{align*}
  \]
 Parsing Experiments

- Setup: MSTParser (2nd order) w/ standard processing
- Per-dim bucket << Hierarchical clustering features:

<table>
<thead>
<tr>
<th>System</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>92.0</td>
</tr>
<tr>
<td>SENNA (Buckets)</td>
<td>92.0</td>
</tr>
<tr>
<td>SENNA (Hier. Clustering)</td>
<td>92.3</td>
</tr>
<tr>
<td>HUANG (Buckets)</td>
<td>91.9</td>
</tr>
<tr>
<td>HUANG (Hier. Clustering)</td>
<td>92.4</td>
</tr>
</tbody>
</table>
Parsing Experiments

Main WSJ results:

<table>
<thead>
<tr>
<th>System</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>91.9</td>
</tr>
<tr>
<td><strong>BROWN</strong></td>
<td>92.7</td>
</tr>
<tr>
<td>Senna</td>
<td>92.3</td>
</tr>
<tr>
<td>Turian</td>
<td>92.3</td>
</tr>
<tr>
<td>Huang</td>
<td>92.4</td>
</tr>
<tr>
<td>Skip</td>
<td>92.3</td>
</tr>
<tr>
<td>Skip_{dep}</td>
<td>92.7</td>
</tr>
</tbody>
</table>

**Ensemble Results**

| ALL – BROWN | 92.9 |
| ALL         | 93.0 |

(faster)  
(complementary)
### Parsing Experiments

- **Main Web results:**

<table>
<thead>
<tr>
<th>System</th>
<th>Test Avg (5 domains)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>83.5</td>
</tr>
<tr>
<td>BROWN</td>
<td>84.2</td>
</tr>
<tr>
<td>SENNA</td>
<td><strong>84.3</strong></td>
</tr>
<tr>
<td>TURIAN</td>
<td>83.9</td>
</tr>
<tr>
<td>HUANG</td>
<td>84.1</td>
</tr>
<tr>
<td>SKIP</td>
<td>83.7</td>
</tr>
<tr>
<td>SKIP(_{DEP})</td>
<td>84.1</td>
</tr>
</tbody>
</table>

- **Ensemble Results**

<table>
<thead>
<tr>
<th>Ensemble Results</th>
<th>Test Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL–BROWN</td>
<td>84.7</td>
</tr>
<tr>
<td>ALL</td>
<td>84.9</td>
</tr>
</tbody>
</table>

(faster)  
(complementary)
Correlation w/ Intrinsic Metrics

- Correlation only for variations of a single model

<table>
<thead>
<tr>
<th>Representation</th>
<th>SIM</th>
<th>TAG</th>
<th>Parsing F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SKIP, $w = 10$</td>
<td>44.6</td>
<td>71.5</td>
<td>92.70</td>
</tr>
<tr>
<td>SKIP, $w = 5$</td>
<td>44.4</td>
<td>81.1</td>
<td>92.86</td>
</tr>
<tr>
<td>SKIP, $w = 1$</td>
<td>37.8</td>
<td>86.6</td>
<td>92.94</td>
</tr>
<tr>
<td>SKIP$\text{DEP}$</td>
<td>34.6</td>
<td>88.3</td>
<td>93.33</td>
</tr>
</tbody>
</table>

- Topical
- Syntactic/Functional
Conclusion

- Improvements ~ Brown but with faster training
- Hierarchical clustering >> bucket (per-dim) features
- Syntactic context helps
- Intrinsic metrics ~correlate with parsing accuracy
Thank you!

Data (dependency embeddings and features) at: ttic.uchicago.edu/~mbansal