Dependency-based Convolutional Neural Networks for Sentence Embedding

What is Hawaii’s state flower?

Mingbo Ma  Liang Huang  Bing Xiang  Bowen Zhou
CUNY  IBM T. J. Watson

ACL 2015
Beijing
Kalchbrenner et al. (2014) and Kim (2014) apply CNNs to sentence modeling

- alleviates data sparsity by word embedding
- sequential order (sentence) instead of spatial order (image)

Should use more linguistic and structural information!
Sequential Convolution

Sequential convolution

What is Hawaii's state flower?
Sequential Convolution

What is Hawaii’s state flower?
What is Hawaii’s state flower?
Sequential Convolution

Sequential convolution

What
is
Hawaii’s
state
flower

1
2
3
4
5
6

word rep.

convolution direction
Sequential Convolution

Sequential convolution

What is Hawaii’s state flower?

Word rep. convolution direction
Try different convolution filters and repeat the same process
What is Hawaii's state flower?

Sequential convolution

word rep.

1 2 3 4

state flower

convolution direction
Sequential Convolution

Sequential convolution

What       is       Hawaii       's       state       flower

Max pooling

word rep.

convolution direction
Convolution direction

Sequential convolution

What is Hawaii's state flower?

Max pooling
Classification

Feed into NN
What is Hawaii's state flower?

Sequential Convolution: Location

Gold standard: Entity
What is Hawaii’s state flower?

Sequential convolution
Sequential convolution

What is Hawaii’s state flower.

Sequential Convolution
What is Hawaii’s state flower?
What is Hawaii's state flower?

Sequential convolution
What is Hawaii’s state flower?
Sequential convolution:

- Traditional convolution operates in surface order
- Cons: No structural information is captured
  No long distance relationships
Dependency-based Convolution

Sequential convolution:

- Traditional convolution operates in surface order
- Cons: No structural information is captured
  No long distance relationships

Structural Convolution:

- operates the convolution filters on dependency tree
- more “important” words are convolved more often
- long distance relationships is naturally obtained
What is Hawaii's state flower?
What is Hawaii’s state flower?
What is Hawaii's state flower?

Convolution on Tree

Word rep.

ROOT

Convolution direction

Dependency convolution

Child

Parent

1

2

3

4

5

6
What is Hawaii's state flower?

Convolution on Tree:

- **ROOT**: What
- **Child**: is
- **Parent**: Hawaii
- **Child**: 's
- **Parent**: state
- **Child**: flower

**Convolution Direction**

**Dependency Convolution**

**Word Rep.**
What is Hawaii's state flower?
What is Hawaii’s state flower?
What is Hawaii’s state flower?
Try different Bigram convolution filters and repeat the same process
What is Hawaii's state flower?
What is Hawaii's state flower?

Convolution on Tree

- **ROOT**
- **word rep.**
  - What 1
  - is 2
  - Hawaii's 3
  - state 4
  - flower 6

**Dependency convolution**

**Max pooling**
What is Hawaii’s state flower?
What is Hawaii's state flower?

Convolution on Tree

dependency convolution

Max pooling

convolution direction

ROOT

child

parent

word rep.

1

2

3

4

5

6
What is Hawaii's state flower?
Trigram Convolution on Trees
What is Hawaii's state flower?

Convolution on Tree

Convolution direction

Trigram convolution

word rep.
What is Hawaii’s state flower?

Convolution on Tree

ROOT* ➔ ROOT**

Convolution direction

Trigram convolution

Word rep.
What is Hawaii's state flower?
follow the same steps as before...
What is Hawaii's state flower?

Convolution on Tree

Convolution direction

more important words are convolved more often!
What is Hawaii’s state flower?
What is Hawaii's state flower?

Convolution on Tree

ROOT

What 1
is 2
Hawaii 3
's 4
state 5
flower 6

bigram

trigram

Fully connected NN with softmax output
Besides convolution on ancestor path, we also can capture conjunction information from siblings.
## Experiments

**Tasks:**
- Sentimental analysis
- Question classification

**Datasets:**

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Dataset</th>
<th># Classes</th>
<th>Size</th>
<th>Testset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentimental Analysis</td>
<td>MR</td>
<td>2</td>
<td>10662</td>
<td>10-CV</td>
</tr>
<tr>
<td></td>
<td>SST1</td>
<td>5</td>
<td>11855</td>
<td>2210</td>
</tr>
<tr>
<td>Question Classification</td>
<td>TREC</td>
<td>6</td>
<td>5952</td>
<td>500</td>
</tr>
<tr>
<td></td>
<td>TREC-2</td>
<td>50</td>
<td>5952</td>
<td>500</td>
</tr>
</tbody>
</table>
Sentimental Analysis Data Examples

Sentimental analysis from Rotten Tomatoes (MR & SST-1)

**straightforward statements:**
- simplistic, silly and tedious  
  Negative

**subtle statements:**
- the film tunes into a grief that could lead a  
  man across centuries  
  Positive

**sentences with adversative:**
- not for everyone, but for those with whom it  
  will connect, it's a nice departure from  
  standard moviegoing fare  
  Positive
### Sentimental Analysis Experiments Results

<table>
<thead>
<tr>
<th>Category</th>
<th>Model</th>
<th>MR</th>
<th>SST-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>This work</td>
<td>ancestor</td>
<td>80.4</td>
<td>47.7</td>
</tr>
<tr>
<td></td>
<td>ancestor+sibling</td>
<td>81.7</td>
<td>48.3</td>
</tr>
<tr>
<td></td>
<td>ancestor+sibling+sequential</td>
<td>81.9</td>
<td>49.5</td>
</tr>
<tr>
<td>CNNs</td>
<td>CNNs-non-static (Kim ’14) — baseline</td>
<td>81.5</td>
<td>48.0</td>
</tr>
<tr>
<td></td>
<td>CNNs-multichannel (Kim ’14)</td>
<td>81.1</td>
<td>47.4</td>
</tr>
<tr>
<td></td>
<td>Deep CNNs (Kalchbrenner+ ’14)</td>
<td>-</td>
<td>48.5</td>
</tr>
<tr>
<td>Recursive NNs</td>
<td>Recursive Autoencoder (Socher+ ’11)</td>
<td>77.7</td>
<td>43.2</td>
</tr>
<tr>
<td></td>
<td>Recursive Neural Tensor (Socher+ ’13)</td>
<td>-</td>
<td>45.7</td>
</tr>
<tr>
<td></td>
<td>Deep Recursive NNs (Irsoy+ ’14)</td>
<td>-</td>
<td>49.8</td>
</tr>
<tr>
<td>Recurrent NNs</td>
<td>LSTM on tree (Zhu+ ’15)</td>
<td>81.9</td>
<td>48.0</td>
</tr>
<tr>
<td>Other</td>
<td>Paragraph-Vec (Le+ ’14)</td>
<td>-</td>
<td>48.7</td>
</tr>
<tr>
<td>Sentence</td>
<td>Top-level (TREC)</td>
<td>Fine-grained (TREC-2)</td>
<td></td>
</tr>
<tr>
<td>-------------------------------------------------------------------------</td>
<td>------------------</td>
<td>-----------------------</td>
<td></td>
</tr>
<tr>
<td>How did serfdom develop in and then leave Russia?</td>
<td>DESC</td>
<td>manner</td>
<td></td>
</tr>
<tr>
<td>What is Hawaii 's state flower ?</td>
<td>ENTY</td>
<td>plant</td>
<td></td>
</tr>
<tr>
<td>What sprawling U.S. state boasts the most airports ?</td>
<td>LOC</td>
<td>state</td>
<td></td>
</tr>
<tr>
<td>When was Algeria colonized ?</td>
<td>NUM</td>
<td>date</td>
<td></td>
</tr>
<tr>
<td>What person 's head is on a dime ?</td>
<td>HUM</td>
<td>ind</td>
<td></td>
</tr>
<tr>
<td>What does the technical term ISDN mean ?</td>
<td>ABBR</td>
<td>exp</td>
<td></td>
</tr>
</tbody>
</table>
## Question Classification Experiments Results

<table>
<thead>
<tr>
<th>Category</th>
<th>Model</th>
<th>TREC</th>
<th>TREC2</th>
</tr>
</thead>
<tbody>
<tr>
<td>This work</td>
<td>ancestor</td>
<td>95.4</td>
<td>88.4</td>
</tr>
<tr>
<td></td>
<td>ancestor+sibling</td>
<td>95.6</td>
<td>89.0</td>
</tr>
<tr>
<td></td>
<td>ancestor+sibling+sequential</td>
<td>95.4</td>
<td>88.8</td>
</tr>
<tr>
<td>CNNs</td>
<td>CNNs-non-static (Kim ’14) — baseline</td>
<td>93.6</td>
<td>86.4</td>
</tr>
<tr>
<td></td>
<td>CNNs-multichannel (Kim ’14)</td>
<td>92.2</td>
<td>86.0</td>
</tr>
<tr>
<td></td>
<td>Deep CNNs (Kalchbrenner+ ’14)</td>
<td>93.0</td>
<td>-</td>
</tr>
<tr>
<td>Hand-coded</td>
<td>SVMs (Silva+ ’11)*</td>
<td>95.0</td>
<td>90.8</td>
</tr>
</tbody>
</table>

We achieved the highest published accuracy on TREC.
Cases which we do better than Baseline:

- Gold/Ours: Enty  Baseline: Loc
- Gold/Ours: Enty  Baseline: Desc
- Gold/Ours: Desc  Baseline: Enty
- Gold/Ours: Mild Neg  Baseline: Mild Pos

http://cogcomp.cs.illinois.edu/Data/QA/QC/definition.html
Error Analysis :-(

Cases which we make mistakes:

黄金: Num   我们: Enty   基线: Num

Gold: Num   Ours: Enty   Baseline: Num

Cases which we and baseline make mistakes:

Gold: Num   Ours: Enty   Baseline: Desc

http://cogcomp.cs.illinois.edu/Data/QA/QC/definition.html
Conclusions

Pros:

- Dependency-based convolution captures long-distance information.
- It outperforms sequential CNN in all four datasets.
  - highest published accuracy on TREC.

Cons:

- Our model’s accuracy depends on parser quality.
Deep Learning can and should be combined with linguistic intuitions.

Thank you!