Adaptive Joint Learning of Compositional and Non-Compositional Phrase Embeddings

Kazuma Hashimoto & Yoshimasa Tsuruoka
University of Tokyo
Learning Phrase Embeddings

- Embedding meanings of phrases into a vector space
  - e.g. Recurrent neural networks

\[ \nu(\text{catch eye}) \quad \nu(\text{grab attention}) \]

Composition function

\[ \nu(\text{catch}) \quad \nu(\text{eye}) \quad \nu(\text{grab}) \quad \nu(\text{attention}) \]

word embeddings

phrase embeddings

catch eye

grab attention

Compositional semantics
Compositional or Non-Compositional?

- Not all phrases are compositional

Selecting one of them is not always the best option!

Compositional embedding
$$c(\text{catch eye})$$

$$n(\text{catch}_\text{eye})$$

Non-compositional embedding

Composition function

$$\nu(\text{catch}) \quad \nu(\text{eye})$$

Grab attention

catch eye

Catch ear

catch e-mail
This Work: Adaptive Joint Learning

- Compositionality detection
- Non-compositional embedding
- Compositional embedding

\[
\alpha(p) \in [0, 1]
\]

\[
\begin{align*}
\phi(p) &= 1 - \alpha(p) \\
\psi(p) &= \alpha(p)
\end{align*}
\]

\[p = VO\]

State-of-the-art results on:
- Compositionality detection for **Verb-Object (VO)** pairs
- **Transitive verb** disambiguation
Adaptive Weighted Addition

- Very simple weighted addition of $c(p)$ and $n(p)$
  - $p$: phrase
  - $\alpha(p)$: compositionality score

$$v(p) = \alpha(p)c(p) + (1 - \alpha(p))n(p)$$

$$\alpha(p) = \sigma(W \cdot \phi(p)) \quad \text{← logistic function}$$
Overview of the General Training Process

- All model parameters are \textit{jointly} learned according to the objective function for a specific task
  - The compositionality levels are automatically adjusted to the task

\[ v(p) = \alpha(p) c(p) + (1 - \alpha(p)) n(p) \]

\[ \alpha(p) = \sigma(W \cdot \phi(p)) \]
Focusing on Verb-Object Compositionality

- Co-occurrence-based **Subject-Verb-Object (SVO)** phrase embeddings *(Hashimoto and Tsuruoka, 2015)*
  - Applying our joint learning method to VO phrases
    - $p = VO$ in our method

\[
\alpha(VO) = \sigma(W \cdot \phi(VO)) \\
v(VO) = \alpha(VO)c(VO) \\
+ (1 - \alpha(VO))n(VO)
\]
Simple Statistical Features for $\alpha(VO)$

- Indices (binary features) of
  - Verb
  - Object
  - Verb-object pair

- Count-based features ([Venkatapathy and Joshi, 2005](#))
  - Frequency
  - PMI

  - Effective in detecting the compositionality of verb-object pairs ([McCarthy et al., 2007](#))
Experimental Settings

• Training data
  – BNC
    • SVO: 1.38M pairs, SVOPN 0.93M pairs
  – Wikipedia
    • SVO 23.6M pairs, SVOPN 17.3M pairs
  – BNC-Wikipedia
    • The concatenation of the BNC and Wikipedia data
• 25-dimensional embeddings (and matrices)
  – Random initialization
  – AdaGrad  (Duchi et al., 2011)
Evaluation 1: Compositionality Detection

- Measuring compositionality levels of verb-object pairs
  - Human rating: [1, 6]
  - Higher score → more compositional
    - VJ’05: 765 pairs (Venkatapathy and Joshi, 2005)
    - MC’07: 638 pairs (McCarthy et al., 2007)

<table>
<thead>
<tr>
<th>Phrase</th>
<th>Human rating</th>
<th>System output</th>
</tr>
</thead>
<tbody>
<tr>
<td>buy car</td>
<td>6</td>
<td>0.78</td>
</tr>
<tr>
<td>own land</td>
<td>6</td>
<td>0.79</td>
</tr>
<tr>
<td>take toll</td>
<td>1.5</td>
<td>0.14</td>
</tr>
<tr>
<td>shed light</td>
<td>1</td>
<td>0.21</td>
</tr>
<tr>
<td>bear fruit</td>
<td>1</td>
<td>0.15</td>
</tr>
<tr>
<td>make noise</td>
<td>6</td>
<td>0.37</td>
</tr>
<tr>
<td>have reason</td>
<td>5</td>
<td>0.26</td>
</tr>
</tbody>
</table>

\[ \alpha(VO) \]

Spearman’s rank correlation
Our method outperforms the state-of-the-art supervised and WordNet-based methods

- Without explicit training data for the task

<table>
<thead>
<tr>
<th>Method</th>
<th>MC’07</th>
<th>VJ’05</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed method (Wikipedia)</td>
<td>0.508</td>
<td>0.514</td>
</tr>
<tr>
<td>Proposed method (BNC)</td>
<td>0.507</td>
<td>0.507</td>
</tr>
<tr>
<td>Proposed method (BNC-Wikipedia)</td>
<td>0.518</td>
<td>0.527</td>
</tr>
<tr>
<td>Proposed method (Ensemble)</td>
<td>0.550</td>
<td>0.552</td>
</tr>
<tr>
<td>Kiela and Clark (2013) w/ WordNet</td>
<td>n/a</td>
<td>0.461</td>
</tr>
<tr>
<td>Kiela and Clark (2013)</td>
<td>n/a</td>
<td>0.420</td>
</tr>
<tr>
<td>DSPORTO+ (McCarthy et al., 2007)</td>
<td>0.454</td>
<td>n/a</td>
</tr>
<tr>
<td>DSPORTO (McCarthy et al., 2007)</td>
<td>0.398</td>
<td>n/a</td>
</tr>
<tr>
<td>PMI (McCarthy et al., 2007)</td>
<td>0.274</td>
<td>n/a</td>
</tr>
<tr>
<td>Frequency (McCarthy et al., 2007)</td>
<td>0.141</td>
<td>n/a</td>
</tr>
<tr>
<td>Human agreement</td>
<td>0.702</td>
<td>0.716</td>
</tr>
</tbody>
</table>
Trends of $\alpha(VO)$ during the Training

- $\alpha(VO)$ converges according to its corresponding phrase
  - Room for improvement when treating light verbs

![Graph showing the trends of $\alpha(VO)$ during training](image-url)
The results match our intuition:

- $\alpha(VO)$ is low for phrases which form idiomatic phrases.

<table>
<thead>
<tr>
<th>Highest average scores</th>
<th>Lowest average scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>approve 0.83</td>
<td>bear 0.37</td>
</tr>
<tr>
<td>reject 0.72</td>
<td>play 0.38</td>
</tr>
<tr>
<td>discuss 0.71</td>
<td>have 0.38</td>
</tr>
<tr>
<td>visit 0.70</td>
<td>make 0.39</td>
</tr>
<tr>
<td>want 0.70</td>
<td>break 0.40</td>
</tr>
<tr>
<td>describe 0.70</td>
<td>take 0.40</td>
</tr>
<tr>
<td>involve 0.69</td>
<td>raise 0.41</td>
</tr>
<tr>
<td>own 0.68</td>
<td>reach 0.41</td>
</tr>
<tr>
<td>attend 0.68</td>
<td>gain 0.42</td>
</tr>
<tr>
<td>reflect 0.67</td>
<td>draw 0.42</td>
</tr>
</tbody>
</table>
Evaluation 2: Verb Disambiguation

- Measuring similarities of transitive verbs paired with the same subjects and objects
  - Human rating: [1, 7] (Grefenstette and Sadrzadeh, 2011)
  - Higher score → more similar

Transitive verb disambiguation task (GS’11)

<table>
<thead>
<tr>
<th>phrase pair</th>
<th>human rating</th>
<th>system output</th>
</tr>
</thead>
<tbody>
<tr>
<td>student write name</td>
<td>7</td>
<td>0.91</td>
</tr>
<tr>
<td>student spell name</td>
<td>6</td>
<td>0.85</td>
</tr>
<tr>
<td>child show sign</td>
<td>6</td>
<td>0.85</td>
</tr>
<tr>
<td>child express sign</td>
<td>1</td>
<td>-0.1</td>
</tr>
<tr>
<td>system meet criterion</td>
<td>1</td>
<td>-0.1</td>
</tr>
</tbody>
</table>

Spearman’s rank correlation

- Cosine distance between SVO embeddings

08/08/2016 ACL 2016 in Berlin, Germany
Outperforming the baseline and other methods

- Our adaptive joint learning method is more effective than just using compositional embeddings

<table>
<thead>
<tr>
<th>Method</th>
<th>GS’11a</th>
<th>GS’11b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed method (Wikipedia)</td>
<td>0.598</td>
<td>0.461</td>
</tr>
<tr>
<td>Proposed method (BNC)</td>
<td>0.595</td>
<td>0.463</td>
</tr>
<tr>
<td>Proposed method (BNC-Wikipedia)</td>
<td>0.623</td>
<td>0.483</td>
</tr>
<tr>
<td>Proposed method (Ensemble A)</td>
<td>0.661</td>
<td>0.511</td>
</tr>
<tr>
<td>Proposed method (Ensemble B)</td>
<td>0.680</td>
<td>0.524</td>
</tr>
<tr>
<td>$\alpha(VO) = 1$ (Wikipedia)</td>
<td>0.576</td>
<td>n/a</td>
</tr>
<tr>
<td>$\alpha(VO) = 1$ (BNC)</td>
<td>0.574</td>
<td>n/a</td>
</tr>
<tr>
<td>Milajevs et al. (2014)</td>
<td>0.456</td>
<td>n/a</td>
</tr>
<tr>
<td>Polajnar et al. (2014)</td>
<td>n/a</td>
<td>0.370</td>
</tr>
<tr>
<td>Hashimoto et al. (2014)</td>
<td>0.420</td>
<td>0.340</td>
</tr>
<tr>
<td>Polajnar et al. (2015)</td>
<td>n/a</td>
<td>0.330</td>
</tr>
<tr>
<td>Grefenstette and Sadrzadeh (2011)</td>
<td>n/a</td>
<td>0.210</td>
</tr>
<tr>
<td>Human agreement</td>
<td>0.750</td>
<td>0.620</td>
</tr>
</tbody>
</table>

Baseline of $v(VO) = c(VO)$
Examples of Closest Neighbor Embeddings

<table>
<thead>
<tr>
<th></th>
<th>Proposed method</th>
<th>Baseline of $v(VO) = c(VO)$</th>
</tr>
</thead>
</table>
| take toll      | $\alpha$(take toll) = 0.11

*capturing idiomatic aspect*

- put strain
- place strain
- cause strain
- have effect
- exacerbate injury

- deplete division
- necessitate monitoring
- deplete pool
- create pollution
- deplete field

| catch eye      | $\alpha$(catch eye) = 0.14

*should be higher*

- catch attention
- grab attention
- make impression
- lift spirit
- become favorite

- catch ear
- catch heart
- catch e-mail
- catch imagination
- catch attention

| bear fruit     | $\alpha$(bear fruit) = 0.19

- accentuate effect
- enhance beauty
- enhance atmosphere
- rejuvenate earth
- enhance habitat

- bear herb
- bear grain
- bear spore
- bear variety
- bear seed

| make noise     | $\alpha$(make noise) = 0.33

*should be higher*

- attack intruder
- attack trespasser
- avoid predator
- attack diver
- attack pedestrian

- make sound
- do beating
- get bounce
- get pulse
- lose bit
Summary

- Adaptive joint learning of
  - Compositional and non-compositional embeddings
  - Compositionality detection function
    - State-of-the-art results on compositionality detection and transitive verb disambiguation tasks

- Future work:
  - Context-sensitive compositionality detection
  - General types of phrases