The Role of Context Types and Dimensionality in Learning Word Embeddings

Oren Melamud, David McClosky, Siddharth Patwardhan, Mohit Bansal

NAACL, 2016
What’s a good word embedding for my task?

so many choices...
What’s a good word embedding for my task?

Useful in supervised tasks:

- As pre-training initialization
- With limited supervised data

so many choices...
What's a good word embedding for my task?

Useful in supervised tasks:

- As pre-training initialization
- With limited supervised data

Applied to various tasks:

- Dependency Parsing
- Named Entity Recognition
- Co-reference Resolution
- Sentiment Analysis
- More...

so many choices...
Plethora of Word Embeddings

so many choices...

Easy to obtain

- Off-the-shelf
- Do-it-yourself toolkits
Plethora of Word Embeddings

Lots of choices to make
Plethora of Word Embeddings

Lots of choices to make

1. Input
   - Context type
     (BOW-N, syntactic, ...)
   - Learning corpus
Plethora of Word Embeddings

Lots of choices to make

1. Input
   - Context type (BOW-N, syntactic, ...)
   - Learning corpus

2. Computational model
   - Model type (word2vec, GloVe, ...)
   - Hyperparameters
Plethora of Word Embeddings

Lots of choices to make

1. Input
   - Context type
     (BOW-N, syntactic, ...)
   - Learning corpus

2. Computational model
   - Model type
     (word2vec, GloVe, ...)
   - Hyperparameters

3. Output
   - Dimensionality
     (is higher always better?)
Plethora of Word Embeddings

Lots of choices to make

1. Input
   • Context type
     (BOW-N, syntactic, ...)
   • Learning corpus

2. Computational model
   • Model type
     (word2vec, GloVe, ...)
   • Hyperparameters

3. Output
   • Dimensionality
     (is higher always better?)

4. Post-processing
   • Ensembles, retrofitting, ...
Our Focus

Choices we explore:

1. Input
   - Context type
     (BOW-N, syntactic, substitute)
   - Wikipedia + Gigaword + UMBC (web)

2. Computational model
   - word2vec

3. Output
   - Dimensionality
     (is higher always better?)

4. Post-processing
   - Embeddings combinations
     (concat, SVD, CCA)

Evaluated extensively on intrinsic and extrinsic tasks
Research questions:

- Do intrinsic benchmarks predict extrinsic performance?
Our Focus

Research questions:

• Do intrinsic benchmarks predict extrinsic performance?
• Tune context type and dimensionality per extrinsic task?
Research questions:

- Do intrinsic benchmarks predict extrinsic performance?
- Tune context type and dimensionality per extrinsic task?
- Can we benefit from combining different context types?
A new word2vec context type (substitute-based)

- Based on $n$-gram language modeling
Outline

• Context types and dimensionality

• Combining context types

• Conclusions
Context Types and Dimensionality
The Italian chef baked the cake in the oven.

**Common Context Types**

**BOW-2 Contexts**

<table>
<thead>
<tr>
<th>t</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>baked</td>
<td>Italian</td>
</tr>
<tr>
<td>baked</td>
<td>chef</td>
</tr>
<tr>
<td>baked</td>
<td>the</td>
</tr>
<tr>
<td>baked</td>
<td>cake</td>
</tr>
</tbody>
</table>

**Dependency Contexts**

<table>
<thead>
<tr>
<th>t</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>baked</td>
<td>nsubj:chef</td>
</tr>
<tr>
<td>baked</td>
<td>dobj:cake</td>
</tr>
<tr>
<td>baked</td>
<td>prep_in:oven</td>
</tr>
</tbody>
</table>
Learning word2vec Skip-gram Embeddings

BOW-2 Contexts

<table>
<thead>
<tr>
<th>t</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>baked</td>
<td>Italian</td>
</tr>
<tr>
<td>baked</td>
<td>chef</td>
</tr>
<tr>
<td>baked</td>
<td>the</td>
</tr>
<tr>
<td>baked</td>
<td>cake</td>
</tr>
</tbody>
</table>

Dependency Contexts

<table>
<thead>
<tr>
<th>t</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>baked</td>
<td>nsubj:chef</td>
</tr>
<tr>
<td>baked</td>
<td>dobj:cake</td>
</tr>
<tr>
<td>baked</td>
<td>prep_in:oven</td>
</tr>
</tbody>
</table>

\[
\sum_{(t,c) \in PAIRS} \left( \log \sigma(v'_c \cdot v_t) + \sum_{neg \in NEG_{(t,c)}} \log \sigma(-v'_{neg} \cdot v_t) \right)
\]
Substitute-based Contexts

Potential substitutes encode the context (Yuret, 2012)

The Italian chef *baked* the cake in the oven
Substitute-based Contexts

Potential substitutes encode the context (Yuret, 2012)

The Italian chef ____ the cake in the oven

0.50  put
0.25  baked
0.15  cooked
0.10  forgot
Substitute-based Contexts

Potential substitutes encode the context (Yuret, 2012)

The Italian chef **baked** the cake in the oven

- 0.50 put
- 0.25 baked
- 0.15 cooked
- 0.10 forgot

<table>
<thead>
<tr>
<th>Substitute Contexts</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>t</strong></td>
</tr>
<tr>
<td>baked</td>
</tr>
<tr>
<td>baked</td>
</tr>
<tr>
<td>baked</td>
</tr>
<tr>
<td>baked</td>
</tr>
</tbody>
</table>
word2vec with Substitute-based Contexts

<table>
<thead>
<tr>
<th>Substitute Contexts</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
</tr>
<tr>
<td>baked</td>
</tr>
<tr>
<td>baked</td>
</tr>
<tr>
<td>baked</td>
</tr>
<tr>
<td>baked</td>
</tr>
</tbody>
</table>

\[
\sum_{(t,s) \in PAIRS} \ W_{t,s} \cdot \left( \log \sigma(v'_s \cdot v_t) + \sum_{\text{neg} \in \text{NEGS}_{(t,s)}} \log \sigma(-v'_{\text{neg}} \cdot v_t) \right)
\]
Top-5 closest words to ‘**playing**’

<table>
<thead>
<tr>
<th>W-10</th>
<th>DEP</th>
<th>SUB</th>
</tr>
</thead>
<tbody>
<tr>
<td>played</td>
<td>play</td>
<td>singing</td>
</tr>
<tr>
<td>play</td>
<td>played</td>
<td>rehearsing</td>
</tr>
<tr>
<td>plays</td>
<td>understudying</td>
<td>performing</td>
</tr>
<tr>
<td>professionally</td>
<td>caddying</td>
<td>composing</td>
</tr>
<tr>
<td>player</td>
<td>plays</td>
<td>running</td>
</tr>
</tbody>
</table>

**Topical**  
**Functional**

Small context windows also yield ‘functional’ similarity
Intrinsic Evaluations - Word Similarity

Topical
( lion:zoo )

Functional
( lion:cat )

* Similar results for SimLex-999

- Context type matters
- Higher dimensionality is generally better
Can we find similar patterns in extrinsic tasks?
## Extrinsic Evaluations

<table>
<thead>
<tr>
<th>System</th>
<th>Benchmark</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Stanford NN Dependency Parser</strong></td>
<td></td>
</tr>
<tr>
<td>Chen &amp; Manning (2014)</td>
<td>PTB</td>
</tr>
<tr>
<td><strong>Named Entity Recognition</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>shared task</td>
</tr>
<tr>
<td><strong>Co-reference Resolution</strong></td>
<td></td>
</tr>
<tr>
<td>Full features + embeddings</td>
<td>shared task</td>
</tr>
<tr>
<td><strong>Sentiment Analysis</strong></td>
<td></td>
</tr>
<tr>
<td>Average of embeddings</td>
<td>Sentence-level</td>
</tr>
<tr>
<td>with logistic regression</td>
<td>Sentiment Treebank</td>
</tr>
<tr>
<td></td>
<td>Socher et al. (2013)</td>
</tr>
</tbody>
</table>

*Only dev-set experiments*
Extrinsic Evaluations - Parsing

- Preference for ‘functional’ embeddings
- Best performance at $d = 50$ (due to limited supervision?)
Extrinsic Evaluations - Parsing

- Similar context type preferences
- But different dimensionality preferences
Extrinsic Evaluations - NER

- Best performance at $d = 50$
- No clear context type preference
Extrinsic Evaluations - Sentiment Analysis

- No context type preference
- Higher dimensionality is better
Extrinsic Evaluations - Coreference Resolution

- Small performance diffs (competitive non-embedding features)
Extrinsic Evaluations - Summary

- Correlation with intrinsic results
Extrinsic Evaluations - Summary

- Correlation with intrinsic results
- Dimensionality preferences
Extrinsic Evaluations - Summary

- Correlation with intrinsic results
- Dimensionality preferences
- Context type preferences
Context Combinations
Embeddings Concatenation

Let the classifier choose the valuable information:

<table>
<thead>
<tr>
<th></th>
<th>boy</th>
<th>girl</th>
<th>dog</th>
</tr>
</thead>
<tbody>
<tr>
<td>dim1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dim2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>boy</th>
<th>girl</th>
<th>dog</th>
</tr>
</thead>
<tbody>
<tr>
<td>dim1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dim2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dim3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dim4</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

25
Concatenation
Concat helps when ‘regular’ increase in dimensionality is ‘exhausted’
Concat helps when ‘regular’ increase in dimensionality is ‘exhausted’
‘Topical’ + ‘Functional’ concats worked best

- W10 + SUB
- W10 + W1
- W10 + DEP
• Compression via SVD or CCA degrades performance

• Better let the task-specific classifier ‘choose’ the relevant information
Conclusions
Summary

• Do intrinsic benchmarks predict extrinsic performance?

• Tune context type and dimensionality per extrinsic task?

• Can we benefit from combining different context types?

Thank you and happy cooking!
Summary

• Do intrinsic benchmarks predict extrinsic performance? NO

• Tune context type and dimensionality per extrinsic task? YES

• Can we benefit from combining different context types? MAYBE

Thank you and happy cooking!
Summary

• Do intrinsic benchmarks predict extrinsic performance? **NO**
• Tune context type and dimensionality per extrinsic task?

Thank you and happy cooking!
Summary

- Do intrinsic benchmarks predict extrinsic performance? **NO**
- Tune context type and dimensionality per extrinsic task? **YES**

Thank you and happy cooking!
Summary

• Do intrinsic benchmarks predict extrinsic performance? NO
• Tune context type and dimensionality per extrinsic task? YES
• Can we benefit from combining different context types?
Summary

- Do intrinsic benchmarks predict extrinsic performance? **NO**
- Tune context type and dimensionality per extrinsic task? **YES**
- Can we benefit from combining different context types? **MAYBE**
Summary

• Do intrinsic benchmarks predict extrinsic performance? NO
• Tune context type and dimensionality per extrinsic task? YES
• Can we benefit from combining different context types? MAYBE

Thank you and happy cooking!