Word Representations: a Simple and General Method for Semi-Supervised Learning [Turian et al., 2012]

Massimo Innamorati

University of Edinburgh

February 26, 2016
Overview

Introduction
  Overview of the paper’s aims

Word representations
  Distributional Representations
  Brown Clustering
  Distributed Representations

Evaluation Tasks
  Chunking
  Named Entity Recognition

Experiments & Results
  Induction of Word Representations
  Results
  Conclusion
Motivation

Observation

Semi-supervised NLP systems achieve higher accuracy than their supervised counterparts.
Motivation

Observation
Semi-supervised NLP systems achieve higher accuracy than their supervised counterparts.

Problem
Which features - or combination thereof - to use given a task?
Motivation

Observation
Semi-supervised NLP systems achieve higher accuracy than their supervised counterparts.

Problem
Which features - or combination thereof - to use given a task?

Focus
Clustering-based and distributed representations.
Sequence labelling tasks: NER and chunking.
Distributional Representations

Aim
Generate a cooccurrence matrix $F$ of size $W \times C$. 

Previous Literature
[Salgren, 2006] Improves classification tasks (e.g. IR, WSD). Not known which settings ideal for NER & chunking.
Distributional Representations

Aim
Generate a cooccurrence matrix $F$ of size $W \times C$.

Settings
Choose a context (window direction and size).
Distributional Representations

**Aim**
Generate a cooccurrence matrix $F$ of size $W \times C$.

**Settings**
Choose a context (window direction and size).
Choose a count.
Distributional Representations

Aim
Generate a cooccurrence matrix $F$ of size $W \times C$.

Settings
Choose a context (window direction and size).
Choose a count.
Choose a function $g$ to reduce dimensionality of $F_w$. 

Previous Literature
[Salgren, 2006] Improves classification tasks (e.g. IR, WSD). Not known which settings ideal for NER & chunking.
Distributional Representations

Aim
Generate a cooccurrence matrix $F$ of size $WxC$.

Settings
Choose a context (window direction and size).
Choose a count.
Choose a function $g$ to reduce dimensionality of $F_w$.

Previous Literature
[Salgren, 2006] Improves classification tasks (e.g. IR, WSD). Not known which settings ideal for NER & chunking.
Brown Clustering

Aim
Generate $K$ hierarchical clusters based on bigram mutual information.
Brown Clustering

Aim
Generate $K$ hierarchical clusters based on bigram mutual information.
Sample output:
cat 1101
dog 1100
city 1001
town 1011
Brown Clustering

Aim
Generate $K$ hierarchical clusters based on bigram mutual information.
Sample output:
cat 1101
dog 1100
city 1001
town 1011

Pros & Cons
Hierarchy allows to choose among several levels.
Use of bigrams is restrictive.
Distributed Representations

Aim
Use a neural network to generate word vectors whose features capture latent semantic and syntactic properties.
Distributed Representations
[Collobert & Weston, 2008]

Training

- for each epoch
  - Read an n-gram \( x = (w_1, ..., w_n) \)
  - Calculate \( e(x) = e(w_1) \oplus \cdots \oplus e(w_n) \)
  - Pick a corrupted n-gram \( \tilde{x} = (w_1, ..., w_{n-q}, \tilde{w}_n) \) and calculate \( e(\tilde{x}) \)
  - Get \( s(x) \) by passing \( e(x) \) through SLNN.
  - \( L(x) = \max(0, 1 - s(x) + s(\tilde{x})) \)
Distributed Representations
Hierarchical Log-Bilinear model [Mnih & Hinton, 2009]

Training
Given an n-gram, concatenate embeddings of \( n - 1 \) first words. Learn a linear model to predict the last word.
Aims

Hypothesis
It is a task-independent generalisation that supervised NLP systems can be improved by adding word representations as word vectors (thus turning them into semi-supervised systems).
Aims

Hypothesis
It is a task-independent generalisation that supervised NLP systems can be improved by adding word representations as word vectors (thus turning them into semi-supervised systems).

Method
Compare semi-supervised models obtained with off-the-shelf embeddings to previous ones with task-specific information, in particular [Ando & Zhang, 2005] and [Suzuki & Isozaki, 2008] for chunking and [Lin & Wu, 2009] for NER.
Chunking

Syntactic sequence labelling task, consisting of identifying phrases.
Chunking

Syntactic sequence labelling task, consisting of identifying phrases.

Method
Use publicly available CRFsuite chunker.
Add word embedding features learnt from RCV1 corpus (1.3M sentences).
Train on 8.9K sentences of WSJ newswire in Penn Treebank corpus.
NER

Classification task.
NER

Classification task.

Method
Use publicly available system by [Ratinov & Roth 2009].
Train on 14K sentences of Reuters newswire from CoNLL03 dataset.
Add word embedding features learnt from RCV1 corpus (1.3M sentences).
Test on Reuter + out-of-domain dataset MUC7 (with unseen NE types).
Induction of Word Representations

Brown
Models with 1000, 100, 320, and 3200 clusters.
Used clusters at depth 4, 6, 10, and 20.
Induction of Word Representations

Brown
Models with 1000, 100, 320, and 3200 clusters.
Used clusters at depth 4, 6, 10, and 20.

Collobert & Weston
50 epochs.
Embeddings of dimensionality 25, 50, 100, or 200 learnt over 5-gram windows.
Induction of Word Representations

Brown
Models with 1000, 100, 320, and 3200 clusters.
Used clusters at depth 4, 6, 10, and 20.

Collobert & Weston
50 epochs.
Embeddings of dimensionality 25, 50, 100, or 200 learnt over 5-gram windows.

HLBL
Embeddings of dimensionality 100 learnt over 5-gram windows.
Induction of Word Representations

**Brown**
Models with 1000, 100, 320, and 3200 clusters.
Used clusters at depth 4, 6, 10, and 20.

**Collobert & Weston**
50 epochs.
Embeddings of dimensionality 25, 50, 100, or 200 learnt over 5-gram windows.

**HLBL**
Embeddings of dimensionality 100 learnt over 5-gram windows.
Scaling of Embeddings

In all cases, the features are bounded by a scaling constant $\sigma$ that sets their new standard deviation.

$$E \leftarrow \sigma \cdot /\text{stddev}(E)$$ (1)
Results

Influence of capacity of embeddings on chunking (top) and NER (bottom)
Results

Final results for chunking.

<table>
<thead>
<tr>
<th>System</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>94.16</td>
<td>93.79</td>
</tr>
<tr>
<td>HLBL, 50-dim</td>
<td>94.63</td>
<td>94.00</td>
</tr>
<tr>
<td>C&amp;W, 50-dim</td>
<td>94.66</td>
<td>94.10</td>
</tr>
<tr>
<td>Brown, 3200 clusters</td>
<td>94.67</td>
<td>94.11</td>
</tr>
<tr>
<td>Brown+HLBL, 37M</td>
<td>94.62</td>
<td>94.13</td>
</tr>
<tr>
<td>C&amp;W+HLBL, 37M</td>
<td>94.68</td>
<td>94.25</td>
</tr>
<tr>
<td>Brown+C&amp;W+HLBL, 37M</td>
<td>94.72</td>
<td>94.15</td>
</tr>
<tr>
<td>Brown+C&amp;W, 37M</td>
<td>94.76</td>
<td>94.35</td>
</tr>
<tr>
<td>Ando and Zhang (2005), 15M</td>
<td>-</td>
<td>94.39</td>
</tr>
<tr>
<td>Suzuki and Isozaki (2008), 15M</td>
<td>-</td>
<td>94.67</td>
</tr>
<tr>
<td>Suzuki and Isozaki (2008), 1B</td>
<td>-</td>
<td>95.15</td>
</tr>
</tbody>
</table>
Results

Final results for NER.

<table>
<thead>
<tr>
<th>System</th>
<th>Dev</th>
<th>Test</th>
<th>MUC7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>90.03</td>
<td>84.39</td>
<td>67.48</td>
</tr>
<tr>
<td>Baseline+Nonlocal</td>
<td>91.91</td>
<td>86.52</td>
<td>71.80</td>
</tr>
<tr>
<td>HLBL 100-dim</td>
<td>92.00</td>
<td>88.13</td>
<td>75.25</td>
</tr>
<tr>
<td>Gazetteers</td>
<td>92.09</td>
<td>87.36</td>
<td>77.76</td>
</tr>
<tr>
<td>C&amp;W 50-dim</td>
<td>92.27</td>
<td>87.93</td>
<td>75.74</td>
</tr>
<tr>
<td>Brown, 1000 clusters</td>
<td>92.32</td>
<td><strong>88.52</strong></td>
<td><strong>78.84</strong></td>
</tr>
<tr>
<td>C&amp;W 200-dim</td>
<td><strong>92.46</strong></td>
<td>87.96</td>
<td>75.51</td>
</tr>
<tr>
<td>C&amp;W+HLBL</td>
<td>92.52</td>
<td>88.56</td>
<td>78.64</td>
</tr>
<tr>
<td>Brown+HLBL</td>
<td>92.56</td>
<td>88.93</td>
<td>77.85</td>
</tr>
<tr>
<td>Brown+C&amp;W</td>
<td>92.79</td>
<td>89.31</td>
<td>80.13</td>
</tr>
<tr>
<td>HLBL+Gaz</td>
<td>92.91</td>
<td>89.35</td>
<td>79.29</td>
</tr>
<tr>
<td>C&amp;W+Gaz</td>
<td>92.98</td>
<td>88.88</td>
<td>81.44</td>
</tr>
<tr>
<td>Brown+Gaz</td>
<td><strong>93.25</strong></td>
<td><strong>89.41</strong></td>
<td><strong>82.71</strong></td>
</tr>
<tr>
<td>Lin and Wu (2009), 3.4B</td>
<td>-</td>
<td>88.44</td>
<td>-</td>
</tr>
<tr>
<td>Ando and Zhang (2005), 27M</td>
<td>93.15</td>
<td>89.31</td>
<td>-</td>
</tr>
<tr>
<td>Suzuki and Isozaki (2008), 37M</td>
<td>93.66</td>
<td>89.36</td>
<td>-</td>
</tr>
<tr>
<td>Suzuki and Isozaki (2008), 1B</td>
<td><strong>94.48</strong></td>
<td>89.92</td>
<td>-</td>
</tr>
<tr>
<td>All (Brown+C&amp;W+HLBL+Gaz), 37M</td>
<td>93.17</td>
<td>90.04</td>
<td>82.50</td>
</tr>
<tr>
<td>All+Nonlocal, 37M</td>
<td>93.95</td>
<td>90.36</td>
<td>84.15</td>
</tr>
<tr>
<td>Lin and Wu (2009), 700B</td>
<td>-</td>
<td><strong>90.90</strong></td>
<td>-</td>
</tr>
</tbody>
</table>
Results

Per-token errors given word frequency in chunking (top) and NER (bottom).
Conclusion

These models do not outperform the state-of-the-art semi-supervised by [Ando & Zhang, 2005], [Suzuki & Isozaki, 2008], and [Lin & Wu, 2009].
Conclusion

These models do not outperform the state-of-the-art semi-supervised by [Ando & Zhang, 2005], [Suzuki & Isozaki, 2008], and [Lin & Wu, 2009]. However, they are more general, and prove that task-agnostic embeddings can be used to improve supervised systems.
Conclusion

These models do not outperform the state-of-the-art semi-supervised by [Ando & Zhang, 2005], [Suzuki & Isozaki, 2008], and [Lin & Wu, 2009]. However, they are more general, and prove that task-agnostic embeddings can be used to improve supervised systems.

It is also found that Brown embeddings are better for rare words, and a default method for scaling is presented.
These models do not outperform the state-of-the-art semi-supervised by [Ando & Zhang, 2005], [Suzuki & Isozaki, 2008], and [Lin & Wu, 2009]. However, they are more general, and prove that task-agnostic embeddings can be used to improve supervised systems.

It is also found that Brown embeddings are better for rare words, and a default method for scaling is presented.

Extending the embeddings to phrase representations may be useful.
References

Word representations: A simple and general method for semi-supervised learning
*Proceedings of the 48th annual meeting of the association for computational linguistics*. 12(3), 45 – 678.

The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces
*Institutionen för lingvistik*


A scalable hierarchical distributed language model.
*Advances in neural information processing systems.*
References

*Proceedings of the 43rd annual meeting on association for computational linguistics. Association for Computational Linguistics.*

*Association for Computational Linguistics (pp. 665-673).*

References

The End