Word representations: a simple and general method for semi-supervised learning

GUANNAN LU

MARCH 17, 2015
Outline

- Motivation
- Word representations
  - Distributional representations
  - Clustering-based representations
  - Distributed representations
- Supervised evaluation tasks
  - Chunking
  - Named entity recognition (NER)
- Experiments & Results
- Summary
Motivation

- Semi-supervised approaches can improve accuracy
- It can be tricky and time-consuming
Motivation

- Semi-supervised approaches can improve accuracy
- It can be tricky and time-consuming
- A popular approach:
  - use unsupervised methods to induce word features
Motivation

- Semi-supervised approaches can improve accuracy
- It can be tricky and time-consuming
- A popular approach:
  - use unsupervised methods to induce word features
    - clustering
    - word embeddings
Motivation

- Semi-supervised approaches can improve accuracy
- It can be tricky and time-consuming
- A popular approach:
  - use unsupervised methods to induce word features
    - clustering
    - word embeddings
- Questions:
  - Which features are good for what tasks?
  - Should we prefer certain word features?
  - Can we combine them?
Word Representations

- **Word representation:**
  - A mathematical object associated with each word, often a vector

- **Word feature:** each dimension’s value

- **Conventional representation**
  - E.g. One-hot representation
  - Problems:
    - Data sparsity
Distributional representations

- Co-occurrence matrix $F: W \times C$
  - Each row $F_w$ is initial representation of word $w$
  - Each column $F_c$ is some context
Distributional representations

- Co-occurrence matrix $F: W \times C$
  - Each row $F_w$ is initial representation of word $w$
  - Each column $F_c$ is some context
- Function $g: f = g(F)$
  - Map $F$ to $f: W \times d$ where $d << C$
Distributional representations

- Co-occurrence matrix $F: W \times C$
  - Each row $F_w$ is initial representation of word $w$
  - Each column $F_c$ is some context
- Function $g: f = g(F)$
  - Map $F$ to $f: W \times d$ where $d << C$
- LSA: term-document matrix (Landauer et al., 1998)
Clustering-based representations

- **Brown clustering** (Brown et al., 1992)
  - A hierarchical clustering algorithm
  - A class-based bigram language model
  - Time complexity: $O(V^*K^2)$
    - $V$ is the size of the vocabulary, $K$ is the number of clusters.
  - Limitations:
    - Only based on bigram statistics
    - Not consider word usage
Distributed representations

- Not to be confused with distributional representations!
Distributed representations

- Not to be confused with distributional representations!
- Also known as word embeddings
- Dense, real-valued, low-dimensional
- Neural language models
Distributed representations

- Collobert and Weston embeddings (2008)
  - Neural language model
  - Discriminative and non-probabilistic
  - General architecture (e.g. SRL, NER, POS tagging)

- Differences on implementation
  - Not achieve the low log-rank
  - Corrupt the last word for each n-gram
  - Learning rates are separated
Distributed representation

- HLBL embeddings (2009)
  - Log-bilinear model
    - Predict the feature vector of the next word
  - Hierarchical structure (binary tree)
    - Represent each word as a leaf with a particular path
    - Calculate the product of the probability of each binary choice
Evaluation tasks

• Chunking: syntactic sequence labeling
  ○ CoNLL-2000 shared task
  ○ CRFsuite
  ○ Data
    ▷ The Penn Treebank
    ▷ 7936 sentences (training)
    ▷ 1000 sentences (development)
Evaluation tasks

- **NER**: sequence prediction problem
  - The regularized averaged perceptron model (Ratinov and Roth, 2009)
  - CoNLL03 shared task
    - 204k words for training, 51k words for development, 46K words for testing
  - Out-of-domain dataset: MUC7 formal run (59K words)
Evaluation---Features

- Word features: $w_i$ for $i$ in $\{-2, -1, 0, +1, +2\}$, $w_i \land w_{i+1}$ for $i$ in $\{-1, 0\}$.
- Tag features: $w_i$ for $i$ in $\{-2, -1, 0, +1, +2\}$, $t_i \land t_{i+1}$ for $i$ in $\{-2, -1, 0, +1\}$. $t_i \land t_{i+1} \land t_{i+2}$ for $i$ in $\{-2, -1, 0\}$.
- Embedding features [if applicable]: $e_i[d]$ for $i$ in $\{-2, -1, 0, +1, +2\}$, where $d$ ranges over the dimensions of the embedding $e_i$.
- Brown features [if applicable]: $\text{substr}(b_i, 0, p)$ for $i$ in $\{-2, -1, 0, +1, +2\}$, where $\text{substr}$ takes the $p$-length prefix of the Brown cluster $b_i$.

- Previous two predictions $y_{i-1}$ and $y_{i-2}$
- Current word $x_i$
- $x_i$ word type information: all-capitalized, is-capitalized, all-digits, alphanumeric, etc.
- Prefixes and suffixes of $x_i$, if the word contains hyphens, then the tokens between the hyphens
- Tokens in the window $c = (x_{i-2}, x_{i-1}, x_i, x_{i+1}, x_{i+2})$
- Capitalization pattern in the window $c$
- Conjunction of $c$ and $y_{i-1}$.
Experiment

- Unlabeled data
- RCV1 corpus (63 millions words in 3.3 million sentences)
- Preprocessing technique (Liang, 2005)
  - Remove all sentences that are less than 90% lowercase a-z.
**Results**

- **Scaling of word embeddings**

Figure 1: Effect as we vary the scaling factor $\sigma$ (Equation 1) on the validation set F1. We experiment with Collobert and Weston (2008) and HLBL embeddings of various dimensionality. (a) Chunking results. (b) NER results.
Results

- Capacity of word representations

Figure 2: Effect as we vary the capacity of the word representations on the validation set F1. (a) Chunking results. (b) NER results.
## Results

<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>

<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>88.52</td>
<td>78.84</td>
</tr>
<tr>
<td>C&amp;W 200-dim</td>
<td>92.46</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>93.25</td>
<td>89.41</td>
<td>82.71</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>94.48</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>90.90</td>
<td>-</td>
</tr>
</tbody>
</table>
Results

![Graphs showing results for Chunking and NER](image)

- **Chunking**: Graphs displaying the number of per-token errors in the test set as a function of the frequency of words in the unlabeled data, with different clustering methods comparing C&W, 50-dim, and Brown, 3200 clusters.

- **NER**: Graphs showing similar trends but with Brown, 1000 clusters for comparison.
Summary

- Word features
  - in an unsupervised, task-inspecific, and model-agnostic manner
- The disadvantage
  - Accuracy might be lower than a task-specific semi-supervised method
- The contributions
  - The first work to compare different word representations
  - Combining different word representations can improve accuracy further
- Future work
  - Induce phrase representations
  - Apply to other supervised NLP systems
References

Any questions?

Thank you!