From neighborhood to parenthood: 
the advantages of dependency representation 
over bigrams in Brown clustering

Simon Šuster and Gertjan van Noord
University of Groningen
Word representations

• Falls into following categories:
  • Word-space models (DS) + dimensionality reduction
  • Clustering
  • Word embeddings
  • Other probabilistic models

• Improve generalization

• Clustering: grouping similar words (semantic, paradigmatic & orthographic variants)
Brown clusters

- Popular: POS-tagging, NER, parsing, question answering etc.
- Easy-to-understand parameters
- Simplicity and robustness

Word embeddings’ recent momentum:
- Brown clusters hard to beat in real tasks (Turian et al 2010; Bansal et al 2014; Nepal and Yates 2014; Passos et al 2014)

- Brown clusters admittedly less scalable
Our contribution

- Original Brown clustering uses bigram contexts
- Adapt to dependencies: helpful for semantic similarity
- Tool for dependency Brown clustering
How does Brown clustering work

- Maximize data likelihood defined on a class-based bigram LM
- In practice, done through average Mutual Information

1. Assign some word types to unique classes
2. Put each remaining word type to one of these classes by minimizing the MI loss
3. When all word types are merged, further merge the resulting classes to create a hierarchy
From Brown to dependency Brown

Original formulation

• Factorization includes class transitions with conditioning on previous word’s class
• Such representation is local

Our modification

• Adopt dependency representation: less local, more precise (assuming we can trust the parser)
• (Class-based) dependency language model: conditioning on parent’s class
We took some shelter from the rain
We took some shelter from the rain

dependency contexts →
bigram contexts →
(We, took), (took, shelter), (shelter, some), ...
(Optional) 2nd order dependency: collapse on preposition (shelter, rain)
Contexts

Extract parent–child pairs:
(\textit{took}, \textit{We}),
(\textit{took}, \textit{shelter}),
(\textit{shelter}, \textit{some}),
...
(Optional) 2nd order dependency: collapse on preposition
(\textit{shelter}, \textit{rain})
Evaluation

- Parse a 46M-word reference corpus sample
- Obtain counts of dependency instances as input for the clustering tool
- Semantic similarity task on Dutch wordnet
  - Average similarity over all clusters, as measured by Lin score
<table>
<thead>
<tr>
<th>Group</th>
<th>Cluster id</th>
<th>Most frequent words</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>001010001011100</td>
<td>contractor, family doctor, baker, lawyer, pharmacist, real estate agent, property developer, postman, . . .</td>
</tr>
<tr>
<td>A2</td>
<td>0010100010110111</td>
<td>analyst, reviewer, observer, expert, commentator, people’s rights organisation, insider, . . .</td>
</tr>
<tr>
<td>A3</td>
<td>0010100010111110</td>
<td>entrepreneur, businessman, manager, self-employed, merchant, starter, craftsman, . . .</td>
</tr>
<tr>
<td>B1</td>
<td>011101111011110</td>
<td>me</td>
</tr>
<tr>
<td>B2</td>
<td>011101111011100</td>
<td>him/herself, myself, yourself</td>
</tr>
<tr>
<td>B3</td>
<td>011101111011000</td>
<td>them</td>
</tr>
<tr>
<td>C1</td>
<td>001100100100</td>
<td>Bush, Obama, Clinton, Putin, . . .</td>
</tr>
<tr>
<td>C2</td>
<td>0011000111010</td>
<td>Sarah, Kim, Nathalie, Justine, . . .</td>
</tr>
<tr>
<td>C3</td>
<td>0011000111011</td>
<td>David, Jimmy, Benjamin, . . .</td>
</tr>
<tr>
<td>D1</td>
<td>001011100010101</td>
<td>email, mail, sms, sms_DIM, e-mail, mail_DIM, . . .</td>
</tr>
<tr>
<td>D2</td>
<td>001011100010100</td>
<td>telephone, satellite, telephony, telephone line, Explorer, music player, iTunes, . . .</td>
</tr>
<tr>
<td>E</td>
<td>001000010110101</td>
<td>income, energy consumption, minimum wage, cholesterol, IQ, alcohol content, . . .</td>
</tr>
</tbody>
</table>
Different contexts, different clusters?

- No strong evidence in our case (manual inspection)
- Word embedding and distributional-semantic literature:
  - **Bow**: words associated with target word (topical similarity)
  - **Dep**: words behaving like target word
- Bigram contexts in original Brown clustering too narrow for topical similarity
Varying $k$ number of clusters

Graph showing the average quality (Lin) for varying $k$ with different minfreq values.

- **Standard Brown**
- **DepBrown**

(minfreq=10) and (minfreq=50)
Varying minfreq

- DepBrown
- Standard Brown

Minimum frequency vs. Average quality (Lin)
Varying minfreq + Nouns only

- ▲ Nouns–DepBrown
- ▲- ▲ Nouns–Standard Brown
- ▲ ▲ ▲ ▲ ▲ DepBrown
- ▲ ▲ ▲ ▲ ▲ Standard Brown

Average quality (Lin)

Minimum frequency

0.18 0.20 0.22 0.24 0.26 0.28 0.30 0.32 0.34 0.36 0.38
3 5 10 20 30 50 100
Amount of data

- Standard Brown
- DepBrown

Number of words vs. Average quality (Lin)
Leveraging syntactic functions

- Select parent–child pairs based on dependency label
- Further improvements in semantic similarity by using:
  - subjects
  - direct objects
  - directional complements
  - 2-nd order relations (intervening preposition)
    - directional and prepositional complements
Thank you!

Data & clustering tool at:

github.com/rug-compling/dep-brown-cluster