Representation Learning:
What is it and how do you teach it?

Clayton Greenberg

Spoken Language Systems Group (LSV)
Graduate School of Computer Science (GSCS)
Saarland Informatics Campus, Saarland University (UdS)

May 8, 2017
“I . . . I hardly know, sir, just at present . . . at least I know who I WAS when I got up this morning, but I think I must have been changed several times since then.”
$x_t$ is the set of features about you that is given to the system.

$A$ is where other sources of information are incorporated.

$h_t$ is the task-specific output.

Representation learning is how the computer decides on the $x_t$ to $A$ arrow.
Some natural language processing (NLP) applications

Speech Recognition  Machine Translation  Information Retrieval
The disciplines of NLP
Educational objectives for representation learning

- **CS** Students can collect high-quality data in an organized fashion.
- **M** Students can extrapolate from exploring representative samples.
- **L** Students can recognize and describe useful patterns in data.
- Students can build appropriate models of data.
- Students can apply their models to relevant tasks.
Outline

1. Characters and words: SWORDSS
2. Words and events: thematic fit modeling
3. General teaching philosophy
Outline

1. **Characters and words: SWORDSS**
   - The Collect objective
   - The Extrapolate objective
   - The Recognize objective
   - The Build objective
   - The Apply objective

2. **Words and events: thematic fit modeling**
   - The Collect objective
   - The Extrapolate objective
   - The Recognize objective
   - The Build objective
   - The Apply objective

3. **General teaching philosophy**
Limited data: unbirthday

This is one of very few occurrences of the word “unbirthday”.

- It is especially difficult to learn a good representation for a rare word.
- Distributional hypothesis: the context of the word captures its meaning.
- Fewer occurrences $\Rightarrow$ fewer contexts $\Rightarrow$ less meaning
Sub-word units

- unbirthday $\approx$ un + birthday

- Linguists would separate the word into its morphemes

- Morphological analysis is high precision, but also expensive

- Data that can be collected automatically: subword units of fixed length

- unbirthday $=$ \{unb, nbi, bir, irt, rth, thd, hda, day\}
Rare words are common

<table>
<thead>
<tr>
<th>Language</th>
<th>Vocabulary</th>
<th>Rare Words</th>
<th>Representation Not Found</th>
</tr>
</thead>
<tbody>
<tr>
<td>German</td>
<td>37k</td>
<td>16k</td>
<td>13k</td>
</tr>
<tr>
<td>Tagalog</td>
<td>22k</td>
<td>11k</td>
<td>8k</td>
</tr>
<tr>
<td>Turkish</td>
<td>25k</td>
<td>14k</td>
<td>10k</td>
</tr>
</tbody>
</table>

- Roughly half of the words in the vocabulary are rare.

- Most rare words fail to receive a representation.

- We extrapolate that humans must have backoff strategies for rare words.
Form and function

Which is the Jabberwocky and which is the Bandersnatch?

We recognize universal form-function correspondences.
SWORDSS model: better rare word embeddings

Workflow for applying enhanced word embeddings (Singh et al., 2016).

- **Builds** better embeddings in 4 steps: map, index, search, and combine.
- At end of search step: obtain a list $W$ of words $w$ overlapping target $t$ by at least 3 characters.
- Combine step: $v_t = \sum_{w \in W} StringSimilarity(t, w) \times v_w$
SWORDSS Results

Applied the enhanced representations to two specialized tasks:

1. Rare word similarity task
   - Ask humans to rate how similar two words are on scale from 1 to 10.
   - Compute the cosine of the angle between the two representations.
   - Test the correlation between the ratings and cosines.
   - Achieved near state-of-the-art (Spearman’s $\rho = 0.51$) despite no hard-coded morphological information.

2. Rare word perplexity task
   - Incorporate new representations into a language model.
   - Compute surprisals for each rare word in a corpus.
   - Perplexity is the exponential of the average surprisal value.
   - Outperformed many high performing language model architectures.
Outline

1. Characters and words: SWORDSS
   - The Collect objective
   - The Extrapolate objective
   - The Recognize objective
   - The Build objective
   - The Apply objective

2. Words and events: thematic fit modeling
   - The Collect objective
   - The Extrapolate objective
   - The Recognize objective
   - The Build objective
   - The Apply objective

3. General teaching philosophy
McRae et al. (1997) procedure for patients

On a scale from 1 (not common) to 7 (very common), how common is it for

- soccer
- croquet
- the harpsichord
- cheese

to be played?
Using log polysemy and frequency to predict thematic fit on McRaeNN dataset.
Greenberg et al. (2015a) procedure for patients

On a scale from 1 (strongly disagree) to 7 (strongly agree), how much do you agree that

- soccer
- croquet
- the harpsichord
- cheese

is something that is played?
Confounds become insignificant due to less biased data collection.
ANOVA results: *Polysemy-Fit* interaction

Polysemy and Fit effects on thematic fit scores from Greenberg et al. (2015a)

We *extrapolate* that unequal sense frequencies harm word word representations.
Extrapolate: having one sense that fits is much better than having none.  
Extrapolate: fitting the most preferred sense is a little better than that.  
Extrapolate: polysemy makes judgements less extreme.
Extrapolate: sense frequency distinctions look much less important.

Extrapolate: the softening effect of polysemy is weaker for bad fillers.
An “instrument” example

Alice hit the hedgehog with a gavel, a flamingo, pink quills, and the queen.
Instrument thematic fit judgements

Ferretti et al. (2001): “[On a scale from 1 to 7, how]ow common is it to use each of the following to perform the action of eating?”

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>cup</td>
<td>3.3</td>
</tr>
<tr>
<td>fork</td>
<td>6.7</td>
</tr>
<tr>
<td>knife</td>
<td>6.3</td>
</tr>
<tr>
<td>napkin</td>
<td>3.8</td>
</tr>
<tr>
<td>pliers</td>
<td>1.0</td>
</tr>
<tr>
<td>spoon</td>
<td>6.3</td>
</tr>
<tr>
<td>toothpick</td>
<td>2.1</td>
</tr>
</tbody>
</table>

We recognize that a verb may need a separate representation for each role.
The existing method, step 1: count dependencies

Count verb-role-filler triples & adjust counts by local mutual information (LMI).

\[
LMI(V, R, F) = O_{VRF} \log \frac{O_{VRF}}{E_{VRF}}
\]
The existing method, step 2: compute centroid from top 20

Query the top 20 highest scoring fillers and compute the centroid.

The most typical with-PP arguments of the verb “eat” according to TypeDM.
The existing method, step 3: calculate thematic fit score

Return cosine similarity of test role-filler and centroid.

Sample thematic fit scores using the Baroni and Lenci (2010) method.
Key idea

Build a model in which each verb-role has multiple prototypes (vectors). Use only the closest prototype to determine the thematic fit score.
The *Centroid* method

Illustration of the *Centroid* method for prototype generation, using the most typical with-PP arguments of the verb “eat” according to *TypeDM*. 
The OneBest method

Illustration of the OneBest method for prototype generation, using the most typical with-PP arguments of the verb “eat” according to TypeDM.
The *kClusters* method

Illustration of the *kClusters* method for prototype generation, using the most typical with-PP arguments of the verb “eat” according to *TypeDM* (Greenberg et al., 2015b).
Greenberg et al. (2015a) dataset: results by verb type

<table>
<thead>
<tr>
<th>Method</th>
<th>Polysemeous</th>
<th>Monosemeous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Centroid</td>
<td>0.43</td>
<td>0.66</td>
</tr>
<tr>
<td>OneBest</td>
<td>0.47</td>
<td>0.66</td>
</tr>
<tr>
<td>kClusters</td>
<td>0.47</td>
<td>0.68</td>
</tr>
</tbody>
</table>

Correlation between human judgements from the Greenberg et al. (2015a) dataset (patients) and automatic scores using LMIs from *TypeDM*, by prototype generation method and verb type.

Applying the *kClusters* method correlates best with the human ratings.
Outline

1. **Characters and words: SWORDDSS**
   - The Collect objective
   - The Extrapolate objective
   - The Recognize objective
   - The Build objective
   - The Apply objective

2. **Words and events: thematic fit modeling**
   - The Collect objective
   - The Extrapolate objective
   - The Recognize objective
   - The Build objective
   - The Apply objective

3. **General teaching philosophy**
Interactive Lectures

My lectures are interactive, energetic, and dynamic, with exercises / activities built-in.
Green statements

Students can **collect** high-quality data in an organized fashion.

Students can **extrapolate** from exploring representative samples.

Students can **recognize** and describe useful patterns in data.

Students can **build** appropriate models of data.

Students can **apply** their models to relevant tasks.

We recite short definitions of the most important terms in class.
### Class discussions

<table>
<thead>
<tr>
<th>Method</th>
<th>Polysemous</th>
<th>Monosemous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Centroid</td>
<td>0.43</td>
<td>0.66</td>
</tr>
<tr>
<td>OneBest</td>
<td>0.47</td>
<td>0.66</td>
</tr>
<tr>
<td>kClusters</td>
<td>0.47</td>
<td>0.68</td>
</tr>
</tbody>
</table>

Correlation between human judgements from the Greenberg et al. (2015a) dataset (patients) and automatic scores using LMI from *TypeDM*, by prototype generation method and verb type.

After laying the foundation, we discuss advanced concepts in class.
Quick feedback

Which is the Jabberwocky and which is the Bandersnatch?

Receiving feedback as soon as possible after an activity reinforces learning.
Math has a clear punchline

- **Builds** better embeddings in 4 steps: map, index, search, and combine.
- At end of search step: obtain a list $W$ of words $w$ overlapping target $t$ by at least 3 characters.
- Combine step: $v_t = \sum_{w \in W} \text{StringSimilarity}(t, w) \times v_w$

Count verb-role-filler triples & adjust counts by local mutual information (LMI).

$$LMI(V, R, F) = O_{VRF} \log \frac{O_{VRF}}{E_{VRF}}$$

Sometimes there is no replacement for board work.
Data Science at UCSD

Computer Science  Mathematics  Cognitive Science
Data Science: a map of the field

A map of the data science field from the European Data Science Summer School, Saarland University
