Modelling the Mental Lexicon

Computational Cognitive Science, Lecture 17
Stella Frank, stella.frank@ed.ac.uk
November 13, 2018
Last Lecture: Representations in Mental Lexicon

Items in mental lexicon are related in form, meaning & use

Form:

- Phonological form: sound similarity
- Morphological form: shared morphemes (also imply shared meaning)

Meaning:

- Semantic similarity, relatedness

Use:

- Collocations
Today: Modelling the Mental Lexicon

Goal: build a model of mental lexicon that captures human behaviour

• Can predict which words humans consider similar
• across multiple domains of similarity
• robustly, and at scale
Today: Modelling the Mental Lexicon

Roadmap

• Recap network models

• Introduce *vector embedding* models from NLP
  • Two types: count versus predict

• Evaluate them as cognitive models:
  • What do they capture? What is missing?
Hierarchical Network Model Of Semantic Memory

- Organise concepts in a hierarchy
- Associate properties at highest possible node
- Retrieval (Reaction time, RT) correlated with moving through moving through graph: Predicts Can a penguin fly? faster than Can a canary fly?

Adapted from the Hierarchical Model of Collins and Quillian (1969)

By Nathanael Crawford - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=13268578
Hierarchical Network Model Of Semantic Memory

Issues:

• Rule-based semantics can't capture typicality effects: *Is a canary a bird?* is faster than *Is a penguin a bird?*

• Hard to extend to all concepts (but see WordNet)

• Can only capture semantics

Adapted from the Hierarchical Model of Collins and Quillian (1969)
By Nathanael Crawford - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=13268578
Network models of semantics & phonology

Link words using

- phonological distance
- semantic similarity

Use spreading activation to predict relatedness.

Hard to distinguish different relations or to integrate multiple domains; hard to scale up.

Collins & Loftus (1975)

Goldstein and Vitevich, 2017
The mental lexicon as high-dimensional space

Each point is a word, represented as a 200D vector

(here projected onto 2D)

http://projector.tensorflow.org/
The mental lexicon as high-dimensional space

- If you are not used to thinking about hyper-planes in high-dimensional spaces, now is the time to learn.
- To deal with hyper-planes in a 14-dimensional space, visualize a 3-D space and say “fourteen” to yourself very loudly. Everyone does it.
- But remember that going from 13-D to 14-D creates as much extra complexity as going from 2-D to 3-D.
Word Embeddings

Words are mapped \((embedded)\) from

- a discrete high-dimensional (Vocabulary size-\(D\)) space

\[
dog = [0, 0, 1, 0, \ldots, 0, 0, 0]_V
\]

- a continuous lower-dimensional (still large!) space

\[
dog = [0.14, 12.546, 34.564, -0.235, \ldots, 63.566, -3.435]_D
\]

Important: these dimensions are \(not\) inherently meaningful, they're just useful.
Word Embeddings

Word embeddings are based on an old idea in computational linguistics - vector representations of semantics.

Modern NLP models rely heavily on embeddings.

Google trend search for "word embeddings"
Old Idea: Distributional Semantics

We can represent a word's meaning using patterns of co-occurrence in text: its *distribution*.

"You shall know a word by the company it keeps."

J. R. Firth (1957)

Words with the same patterns of co-occurrence, appearing in similar contexts, will tend to have similar meanings.

Zelig Harris (1954)

*I bought a _ at the pet shop*

*If you don't _ you'll be late!*

*Have a good _*
Count-based Embeddings

Start with a large, sparse matrix of words x contexts gathered from a large corpus.

- Hyperspace Analogy to Language (1996) using local context words

Make this matrix denser using dimensionality reduction, e.g., using SVD, singular value decomposition.

Smooth over rare counts by using PPMI, positive pointwise mutual information: replace each \((w,c)\) cell with

\[
PPMI(w, c) = \max(0, \log \frac{P(w, c)}{P(w)P(c)}) = \max(0, \log \frac{N(w, c)N}{N(w)N(c)})
\]
Embeddings using Prediction (Mikolov et al., 2013)

Train a neural language-model (-ish) over all the data, then extract and use the vector representations (embeddings).

**word2Vec Models:**

- **Continuous bag of words (cbow):**
  - Input layer
  - Hidden layer
  - Output layer
  - Parameters: $W_{V \times N}$, $W'_{N \times V}$

- **skip-gram:**
  - Input layer
  - Hidden layer
  - Output layer
  - Parameters: $W_{V \times N}$, $W'_{N \times V}$
Embeddings from Corpora

Which linguistic domains will these embeddings capture?

• Phonology
• Morphological
• Syntactic
• Semantic
• Collocational?
Embedding space encodes semantic relations

Analogical reasoning as algebra:

king - man + woman = queen
walked - walking + swimming = swam
Embedding Similarity

Similarity is measured using cosine similarity:

\[
\cos(x, y) = \frac{\vec{x} \cdot \vec{y}}{\| \vec{x} \| \| \vec{y} \|} = \frac{\sum_{i}^{D} x_i y_i}{\sqrt{\sum_{i}^{D} x_i^2} \sqrt{\sum_{i}^{D} y_i^2}}
\]

France and Italy are quite similar
\[\theta \text{ is close to } 0^\circ\]
\[\cos(\theta) \approx 1\]

Ball and crocodile are not similar
\[\theta \text{ is close to } 90^\circ\]
\[\cos(\theta) \approx 0\]

The two vectors are similar but opposite
\[\theta \text{ is close to } 180^\circ\]
\[\cos(\theta) \approx -1\]
Evaluating Embedding Space

- Correlate embedding similarities with similarity ratings (explicit ratings, e.g. 1-5 Likert scale)

- Analogies: Performance on SAT-style analogy questions

Sample question

Stem: mason:stone
Choices: (a) teacher:chalk
(b) carpenter:wood
(c) soldier:gun
(d) photograph:camera
(e) book:word
Solution: (b) carpenter:wood

https://aclweb.org/aclwiki/SAT_Analogy_Questions_(State_of_the_art)
Evaluating Embedding Space

- Correlate embedding similarities with similarity ratings (explicit ratings, e.g. 1-5 Likert scale)
- Analogies: Performance on SAT-style analogy questions

**Sample question**

<table>
<thead>
<tr>
<th>Stem:</th>
<th>mason:stone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choices:</td>
<td></td>
</tr>
<tr>
<td>(a) teacher:</td>
<td>chalk</td>
</tr>
<tr>
<td>(b) carpenter:</td>
<td>wood</td>
</tr>
<tr>
<td>(c) soldier:</td>
<td>gun</td>
</tr>
<tr>
<td>(d) photograph</td>
<td>camera</td>
</tr>
<tr>
<td>(e) book:</td>
<td>word</td>
</tr>
<tr>
<td>Solution:</td>
<td>(b) carpenter: wood</td>
</tr>
</tbody>
</table>

What will these measures test?
What will these measures miss?

https://aclweb.org/aclwiki/SAT_Analogy_Questions_(State_of_the_art)
Evaluating embeddings with implicit measures

Can embedding similarity predict primed reaction times?

Journal of Memory and Language 92 (2017) 57–78

Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation

Paweł Mandera *, Emmanuel Keuleers, Marc Brysbaert

Department of Experimental Psychology, Ghent University, Belgium

(See also Ettinger & Linzen, 2016; Hollis & Westbury, 2016, Auguste et al., 2017)
Mandera et al. (2017)

Question: Can a linear regression model that includes embedding similarity predict primed reaction time?

Compare "count" and "predict" models, trained on a 2B word web corpus and/or a 4M word subtitle corpus.

Test on **Semantic Priming Project** data (Hutchinson et al, 2013):

- 6644 prime-target pairs, either semantically related (with different strengths) or unrelated (with matched frequency)
- Tasks: Lexical Decision (see *table*; is *chair* a word Y/N?)
  Speeded Naming (see *bird*, then see & name *egg*)
Baseline: Regression model with word frequency, word length, and orthographic neighbourhood density
Mandera et al. (2017): Findings

- Embedding similarity greatly outperforms baseline: semantics encoded in embeddings is a strong signal for lexical decision RT, naming RT, word association

- Training corpus has an effect: small but naturalistic (subtitle/speech) corpus can match massive corpus

- "Predict" models tend to outperform "count" models: Is "predict" model architecture more cognitively plausible?

- Limitation: evaluating semantic similarity only (due to Semantic Priming Project dataset)
Embeddings as cognitive representations

- Corpus data (digested & represented in embeddings) contains many of the links/relations that humans have -

- despite the fact that humans use language (and learn associated concepts) in an interactive and grounded way, in the physical world,

- while representations from corpus data are ungrounded, based on textual co-occurrence only.

"Yellow banana" problem: we don't mention the obvious - even if the association exists in our mental lexicon.
Word association embeddings

Predicting human similarity judgments with distributional models: The value of word associations

Simon De Deyne and Amy Perfors
Computational Cognitive Science Lab
School of Psychology
University of Adelaide
simon.dedeyne@adelaide.edu.au
amy.perfors@adelaide.edu.au

Daniel J Navarro
School of Psychology
University of New South Wales
dan.navarro@unsw.edu.au

COLING 2016

Create embeddings from (lots of) word association data:

[SML WRLD OF WRDZ]
Discover what words mean for people worldwide
https://smallworldofwords.org/

Embeddings: count-based and "random-walk" (spreading activation through similarity network)
Q: Can these internal-language embeddings outperform corpus (external-language) embeddings on standard explicit similarity ratings datasets?

A: Yes:

Table 1: Spearman rank order correlations between human relatedness and similarity judgments, and the predictions from all four models described earlier. Word association results presented here are based on $G_{123}$. Further details for $G_1$ are available in the text.

<table>
<thead>
<tr>
<th>Data set</th>
<th>$n$</th>
<th>$n(overlap)$</th>
<th>Text Corpus</th>
<th>Word Associations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Count</td>
<td>word2vec</td>
</tr>
<tr>
<td>WordSim-353 Related</td>
<td>252</td>
<td>207</td>
<td>.67</td>
<td>.70</td>
</tr>
<tr>
<td>WordSim-353 Similarity</td>
<td>203</td>
<td>175</td>
<td>.74</td>
<td>.79</td>
</tr>
<tr>
<td>MTURK-771</td>
<td>771</td>
<td>6788</td>
<td>.67</td>
<td>.71</td>
</tr>
<tr>
<td>SimLex-999</td>
<td>998</td>
<td>927</td>
<td>.37</td>
<td>.43</td>
</tr>
<tr>
<td>Radinsky2011</td>
<td>287</td>
<td>137</td>
<td>.75</td>
<td>.78</td>
</tr>
<tr>
<td>RG1965</td>
<td>65</td>
<td>52</td>
<td>.78</td>
<td>.83</td>
</tr>
<tr>
<td>MEN</td>
<td>3000</td>
<td>2611</td>
<td>.75</td>
<td>.79</td>
</tr>
<tr>
<td>Remote Triads</td>
<td>300</td>
<td>300</td>
<td>.65</td>
<td>.52</td>
</tr>
<tr>
<td><strong>mean</strong></td>
<td></td>
<td></td>
<td>.67</td>
<td>.69</td>
</tr>
</tbody>
</table>
Embeddings in the Mental Lexicon?

- Natural language processing deal with External Language - so using representations trained on E-language works well
- Cognitive representations (internal language) are product of
  - E-language (though less exposure than NLP models get)
  - grounded experiences
  - physical language production

Argument about type of training data, not representation itself (dense high-dimensional vectors)
What's missing?

• An account of *learning* from realistic amounts of data: one-shot learning, small-sample learning

• Learning from realistic kinds of data: grounding problem

• Cognitively plausible, well-understood inductive biases (e.g. priors)

• ... including higher-level priors/biases learned from the data: hierarchical model structure

Keeping high-dimensional representations that can capture the patterns of regularities and relations in mental lexicon.