**Vector-based Models of Semantic Composition**

Mirella Lapata

School of Informatics
University of Edinburgh
mlap@inf.ed.ac.uk

---

**Outline**

1. **Introduction**
   - Semantic Space Models
   - Logic-based View
   - Connectionism

2. **Composition Models**

3. **Evaluation**
   - Phrase Similarity Task
   - Paraphrase Detection

4. **Conclusions**

---

**Introduction**

**Distributional Semantics**

- You shall know a word by the company it keeps (Firth, 1957)
- A word’s context provides information about its meaning
- Words are similar if they share similar linguistic contexts

---

**A Simple Semantic Space**

Stuart B. Opotowsky was named vice president for this company with interests in insurance, tobacco, hotels and broadcasting.

|        | vice | president | interests | insurance | ...
|--------|------|-----------|-----------|-----------|
| company| 1    | 1         | 1         | 1         | ...
| company| 25   | 103       | 19        | 55        | ...
| company| 0.06 | 0.26      | 0.05      | 0.14      | ...
| company| 1.52 | 2.32      | 1.14      | 1.06      | ...
Distributional Semantics

Words are represented through their relations to other words.

Key Idea: documents are mixtures of topics, topics are probability distributions over words (Blei et al., 2003; Griffiths and Steyvers, 2002; 2003; 2004).

Topic models are generative and structured. For a new document:

1. Choose a distribution over topics
2. Choose a topic at random according to distribution
3. draw a word from that topic

Statistical techniques used to invert the process: infer set of topics that were responsible for generating a collection of documents.
Meaning Representation

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic n</th>
</tr>
</thead>
<tbody>
<tr>
<td>practical</td>
<td>0.39</td>
<td>0.02</td>
</tr>
<tr>
<td>difficulty</td>
<td>0.03</td>
<td>0.44</td>
</tr>
<tr>
<td>produce</td>
<td>0.06</td>
<td>0.17</td>
</tr>
</tbody>
</table>

- Topics are the dimensions of the space (500, 1000)
- Vector components: probability of word given topic
- Topics correspond to coarse-grained sense distinctions
- Cosine similarity can be used (probabilistic alternatives)

Semantic Space Models

Semantic space models are extremely popular across disciplines!
- Semantic Priming (Lund and Burgess, 1996)
- Text comprehension (Landauer and Dumais, 1997)
- Word association (McDonald, 2000)
- Information Retrieval (Salton et al., 1975)
- Thesaurus extraction (Grefenstette, 1994)
- Word Sense disambiguation (Schütze, 1998)
- Text Segmentation (Hirst, 1997)
- **Automatic, language independent**

**Catch:** representation of the meaning of single words. What about phrases or sentences?

Quick Fix

It was not the sales manager who hit the bottle that day, but the office worker with the serious drinking problem.

That day the office manager, who was drinking, hit the problem sales worker with the bottle, but it was not serious.

- Vector averaging: \( \mathbf{p} = \frac{1}{2}(\mathbf{u} + \mathbf{v}) \) (Foltz et al., 1998; Landauer et al., 1997); **syntax insensitive**
- Add a neighbor to the sum: \( \mathbf{p} = \mathbf{u} + \mathbf{v} + \mathbf{n} \) (Kintsch, 2001);
  - **meaning of predicate depends on its argument**

Logic-based View

Meaning of whole is function of meaning of its parts (Frege, 1957).

\[
\lambda u. \lambda v. \exists x (u @ x \land v @ x) \quad \lambda y. \text{HORSE}(y) \quad \lambda z. \text{RUN}(z)
\]

\[
\exists x (\text{HORSE}(x) \land \text{RUN}(x))
\]

- Logic can account for sentential meaning (Montague, 1974).
- Differences in meaning are **qualitative** rather than **quantitative**.
- Cannot express **degrees of similarity**.
Compositionality

Partee (1995): the meaning of the whole is a function of the meaning of the parts and of the way they are syntactically combined.

Lakoff (1977): the meaning of the whole is a greater than the meaning of the parts.

Frege (1884): never ask the meaning of a word in isolation but only in the context of a statement.

Pinker (1994): composition of simple elements must allow the construction of novel meanings which go beyond those of the individual elements.

Composition Models

A Framework for Semantic Composition

\[ p = f(u, v, f(u, v, OBJ, K)) \]

composition of \( u, v \)

syntactic relationship

Assumptions:

1. eliminate background knowledge \( K \)
2. vary syntactic relationship \( R \)
3. \( p \) is in same space as \( u \) and \( v \)
4. \( f() \) is a linear function of Cartesian product (additive model)
5. \( f() \) is a linear function of tensor product (multiplicative model)

Additive Models

\[ p = Au + Bv \]

Instances

\[ p = u + v \]
\[ p = u + v + \sum_{i} n_i \]
\[ p = \alpha u + \beta v \]
\[ p = v \]

Connectionism

- Tensor products: \( p = u \otimes v \) (Smolensky, 1990); dimensionality
- Circular convolution: \( p = u \ast v \) (Plate, 1991); components are randomly distributed
- Spatter codes: take the XOR of two vectors (Kanerva, 1998); components are random bits

Models (Mitchell and Lapata, 2010)

<table>
<thead>
<tr>
<th></th>
<th>music</th>
<th>solution</th>
<th>economy</th>
<th>craft</th>
<th>create</th>
</tr>
</thead>
<tbody>
<tr>
<td>practical</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>difficulty</td>
<td>1</td>
<td>8</td>
<td>4</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>problem</td>
<td>2</td>
<td>15</td>
<td>7</td>
<td>9</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ \text{practical} + \text{difficulty} = [1 \ 14 \ 6 \ 14 \ 4] \]
\[ \text{practical} + \text{difficulty} + \text{problem} = [3 \ 29 \ 13 \ 23 \ 5] \]
\[ 0.4 \cdot \text{practical} + 0.6 \cdot \text{difficulty} = [0.6 \ 5.6 \ 3.2 \ 6.4 \ 1.6] \]
\[ \text{difficulty} = [1 \ 8 \ 4 \ 4 \ 0] \]
**Composition Models**

Models (Mitchell and Lapata, 2010)

### Multiplicative Models

\[ p = Cuv \]

**Instances**

\[ p = u \odot v \]
\[ p = u \otimes v \]
\[ p = u \oslash v \]

\[ p = \sum_j u_i \cdot v_{i-j} \]

---

### Evaluation

Phrase Similarity Task (Kintsch, 2002)

- Elicit similarity judgments for adjective-noun, noun-noun, verb-object combinations.
- Phrase pairs from three bands: High, Medium, Low.
- Compute vectors for phrases, measure their similarity.
- Correlate model similarities with human ratings.

<table>
<thead>
<tr>
<th>High</th>
<th>Medium</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>old person</td>
<td>elderly lady</td>
<td>small house</td>
</tr>
<tr>
<td>kitchen door</td>
<td>bed room window</td>
<td>housing department</td>
</tr>
<tr>
<td>produce effect</td>
<td>achieve result</td>
<td>consider matter</td>
</tr>
<tr>
<td>start work</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

### Results for verb-obj (Mitchell and Lapata, 2010)

**Dilation Models**

\[ p = Cuv = Uv \]
\[ U_{ij} = 0, U_{ii} = u_i \]
\[ x = \frac{uv}{u^2}u \]
\[ y = v - x = v - \frac{uv}{u^2}u \]
\[ v' = \lambda x + y = (\lambda - 1)\frac{uv}{u^2}u + v \]
\[ p = (\lambda - 1)(u \cdot v)u + (u \cdot u)v \]
Evaluation Phrase Similarity Task

Results for verb-obj (Mitchell and Lapata, 2010)

Results for adj-noun (Mitchell and Lapata, 2010)

Results for noun-noun (Mitchell and Lapata, 2010)
Summary

- Multiplicative and dilation models best for simple space
- Dilation and additive models best for LDA model
- Circular convolution is worst performing model
- Different composition functions appropriate for different representations (additive vs. multiplicative)

What are composition models good for?
- modeling brain activity (Chang et al., 2009)
- language modeling (Mitchell and Lapata, 2009)
- modeling eye tracking data (Mitchell et al., 2010)
- paraphrase detection (Blacoe and Lapata, 2011)

Paraphrase Detection

Given: A pair of sentences $S_1 = (w_1 \ldots w_m)$ and $S_2 = (w_1 \ldots w_n)$

Task: Classify whether $S_1$ and $S_2$ are paraphrases or not

The lies and deceptions from Saddam have been well documented over 12 years.

It has been well documented over 12 years of lies and deception from Saddam.

Microsoft Research Paraphrase Corpus (Dolan et al., 2004).

Features: sentence vectors concatenated (con), subtracted (sub), encoding of words in sentence (enc), sentence vector similarity, unigram overlap, sentence lengths

<table>
<thead>
<tr>
<th>Model</th>
<th>Acc</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>66.5</td>
<td>79.9</td>
</tr>
<tr>
<td>Mihalcea et al. (2006)</td>
<td>70.3</td>
<td>81.3</td>
</tr>
<tr>
<td>Rus et al. (2008)</td>
<td>70.6</td>
<td>80.5</td>
</tr>
<tr>
<td>Qiu et al. (2006)</td>
<td>72.0</td>
<td>81.6</td>
</tr>
<tr>
<td>Islam et al. (2007)</td>
<td>72.6</td>
<td>81.3</td>
</tr>
<tr>
<td><strong>Mitchell and Lapata (2010)</strong></td>
<td><strong>73.0</strong></td>
<td><strong>83.3</strong></td>
</tr>
<tr>
<td>Fernando et al. (2008)</td>
<td>74.1</td>
<td>82.4</td>
</tr>
<tr>
<td>Wan et al. (2006)</td>
<td>75.6</td>
<td>83.0</td>
</tr>
<tr>
<td>Socher et al. (76.4)</td>
<td>76.4</td>
<td>83.6</td>
</tr>
</tbody>
</table>

Conclusions

- University of Edinburgh
  (http://homepages.inf.ed.ac.uk/s1066731/)
- Stanford University, Richard Socher
  (http://www.socher.org/)
- University of Trento, COMPOSES
  (http://clic.cimec.unitn.it/composes/)
- Universities of Oxford and Cambridge, DiscoCat
- Carnegie Mellon University (Alona Fyshe, Brian Murphy)
- Workshop at ACL-2013: Continuous Vector Space Models and their Compositionality
Conclusions

LDA Topics

Mirella Lapata