Qual Presentation

Daniel Khashabi
Outline

- My own line of research

- Papers:
  - Fast Dropout training, ICML, 2013
Outline

- My own line of research
- Papers:
  - Fast Dropout training, ICML, 2013
Motivation

- Developing tools for *word-similarity*
- Useful in many applications
Motivation

- Developing tools for word-similarity
- Useful in many applications
  - For example paraphrase detection:
Motivation

- Developing tools for **word-similarity**
- Useful in many applications
  - For example **paraphrase detection**:

  The Iraqi foreign minister warned of disastrous consequences if Turkey launched an invasion.
Motivation

- Developing tools for word-similarity
- Useful in many applications
  - For example paraphrase detection:

  The Iraqi foreign minister warned of disastrous consequences if Turkey launched an invasion.

  Iraq has warned that a Turkish incursion would have disastrous results.
Motivation

- Developing tools for **word-similarity**
- Useful in many applications
  - For example **paraphrase detection**:

| The Iraqi foreign minister warned of disastrous consequences if Turkey launched an invasion. |
| Iraq has warned that a Turkish incursion would have disastrous results. |
Motivation

- Developing tools for **word-similarity**
- Useful in many applications
  - For example **paraphrase detection**:

The Iraqi foreign minister warned of disastrous consequences if Turkey launched an invasion.

Iraq has warned that a Turkish incursion would have disastrous results.

I can be there on time.
Motivation

- Developing tools for **word-similarity**
- Useful in many applications
  - For example **paraphrase detection**:

  The Iraqi foreign minister warned of disastrous consequences if Turkey launched an invasion.

  Iraq has warned that a Turkish incursion would have disastrous results.

  I can be there on time.

  I can’t be there on time.
Motivation

- Developing tools for **word-similarity**
- Useful in many applications
  - For example **paraphrase detection**:

```
The Iraqi foreign minister warned of disastrous consequences if Turkey launched an invasion.
```

```
Iraq has warned that a Turkish incursion would have disastrous results.
```

```
I can be there on time.
```

```
I can’t be there on time.
```
Motivation (2)

- Developing tools for word-similarity
Motivation (2)

- Developing tools for word-similarity
- We need to solve easier problem
Motivation (2)

- Developing tools for word-similarity
- We need to solve easier problem
  - For example, SAT test:

<table>
<thead>
<tr>
<th>Stem:</th>
<th>word:language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choices:</td>
<td></td>
</tr>
<tr>
<td>(1)</td>
<td>paint:portrait</td>
</tr>
<tr>
<td>(2)</td>
<td>poetry:rhythm</td>
</tr>
<tr>
<td>(3)</td>
<td>note:music</td>
</tr>
<tr>
<td>(4)</td>
<td>tale:story</td>
</tr>
<tr>
<td>(5)</td>
<td>week:year</td>
</tr>
<tr>
<td>Solution:</td>
<td>(3) note:music</td>
</tr>
</tbody>
</table>

- Very important for understanding hierarchies of word semantics
Motivation (3)

- Developing tools for word-similarity
- We need to solve easier problem
Motivation (3)

- Developing tools for word-similarity
- We need to solve easier problem
  - Compositional behavior of the word semantics
Motivation (3)

- Developing tools for word-similarity
- We need to solve easier problem
  - Compositional behavior of the word semantics
Motivation (3)

- Developing tools for word-similarity
- We need to solve easier problem
  - Compositional behavior of the word semantics

Search
Motivation (3)

- Developing tools for word-similarity
- We need to solve easier problem
  - Compositional behavior of the word semantics
Motivation (3)

- Developing tools for word-similarity
- We need to solve easier problem
  - Compositional behavior of the word semantics
Motivation (3)

- Developing tools for word-similarity
- We need to solve easier problem
  - Compositional behavior of the word semantics
Motivation (3)

- Developing tools for word-similarity
- We need to solve easier problem
  - Compositional behavior of the word semantics
Motivation (4)

- Developing tools for word-similarity
- We need to solve easier problem
  - Compositional behavior of the word semantics
Motivation (4)

- Developing tools for word-similarity
- We need to solve easier problem
  - Compositional behavior of the word semantics
  - For example: understanding noun-modifier questions

<table>
<thead>
<tr>
<th>Stem:</th>
<th>fantasy world</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choices:</td>
<td>(1) fairyland</td>
</tr>
<tr>
<td>(2) fantasy</td>
<td></td>
</tr>
<tr>
<td>(3) world</td>
<td></td>
</tr>
<tr>
<td>(4) phantasy</td>
<td></td>
</tr>
<tr>
<td>(5) universe</td>
<td></td>
</tr>
<tr>
<td>(6) ranter</td>
<td></td>
</tr>
<tr>
<td>(7) souring</td>
<td></td>
</tr>
</tbody>
</table>

Solution: (1) fairyland
Designing semantic features

- Feature engineering
  - An important step in semantic modeling of words
Designing semantic features

- Feature engineering
  - An important step in semantic modeling of words
  - The rest is just learning the task in a fully supervised fashion
Designing semantic features

- Feature engineering
  - An important step in semantic modeling of words
- The rest is just learning the task in a **fully supervised** fashion
- Type of the features:
  - Log-Frequency: \( LF(x_i) = \log(freq(x_i) + 1) \)
Designing semantic features

- Feature engineering
  - An important step in semantic modeling of words
- The rest is just learning the task in a fully supervised fashion
- Type of the features:
  - Log-Frequency: $LF(x_i) = \log(freq(x_i) + 1)$
  - PPMI (Positive Pointwise Mutual Information)
Designing semantic features

- Feature engineering
  - An important step in semantic modeling of words
- The rest is just learning the task in a fully supervised fashion
- Type of the features:
  - Log-Frequency: $LF(x_i) = \log(\text{freq}(x_i) + 1)$
  - PPMI (Positive Pointwise Mutual Information)
  - Semantic Similarity
  - Functional Similarity
Feature generation
Feature generation

- Features
  - Log-Frequency
  - PPMI (Positive Pointwise Mutual Information)
  - Semantic Similarity
  - Functional Similarity
Feature generation

- Features
  - Log-Frequency
  - PPMI (Positive Pointwise Mutual Information)
  - Semantic Similarity
  - Functional Similarity

- All features are generated on a collection of documents
  - Of size $5 \times 10^{10}$ words
Feature generation

- Features
  - Log-Frequency
  - PPMI (Positive Pointwise Mutual Information)
  - Semantic Similarity
  - Functional Similarity

- All features are generated on a collection of documents
  - Of size $5 \times 10^{10}$ words

- Definition
  - Word-Context:
    - (Left) Context
    - Word
    - (Right) Context
Feature generation

- **Features**
  - Log-Frequency
  - PPMI (Positive Pointwise Mutual Information)
  - Semantic Similarity
  - Functional Similarity

- **All features are generated on a collection of documents**
  - Of size $5 \times 10^{10}$ words

- **Definition**
  - **Word-Context:**

- **Three word-context matrices**
  - Rows correspond to words/phrases in Wordnet
Pointwise Mutual Information
Pointwise Mutual Information

- **PMI (Pointwise Mutual Information):**

\[
PMI(a, b) = \log \frac{p(a, b)}{p(a)p(b)}
\]
Pointwise Mutual Information

- PMI (Pointwise Mutual Information):
  \[ PMI(a, b) = \log \frac{p(a, b)}{p(a)p(b)} \]

- PPMI (Positive PPMI)
  \[ PPMI(a, b) = \max(0, PMI(a, b)) \]
Pointwise Mutual Information

- **PMI (Pointwise Mutual Information):**
  \[
  PMI(a,b) = \log \frac{p(a,b)}{p(a)p(b)}
  \]

- **PPMI(Positive PPMI)**
  \[
  PPMI(a,b) = \max(0, PMI(a,b))
  \]

- One useful definition for probabilities
  - The ratio of the times a context appears with a words
Pointwise Mutual Information (2)
Pointwise Mutual Information (2)

- Only the words or phrases that exist in the Wordnet
Pointwise Mutual Information (2)

- Only the words or phrases that exist in the Wordnet
- And appear with frequency more than 100 in the corpus
Pointwise Mutual Information (2)

- Only the words or phrases that exist in the Wordnet
- And appear with frequency more than 100 in the corpus
- Find words to the left and right of the word (context) in phrases:

  Table shows forty paradigm words
Pointwise Mutual Information (2)

- Only the words or phrases that exist in the Wordnet
- And appear with frequency more than 100 in the corpus
- Find words to the left and right of the word (context) in phrases:
  
  **Table shows forty paradigm words**

- Create word-context frequency matrix $F$:
Pointwise Mutual Information (2)

- Only the words or phrases that exist in the Wordnet
- And appear with frequency more than 100 in the corpus
- Find words to the left and right of the word (context) in phrases:
  
  Table shows **forty paradigm words**

- Create word-context frequency matrix $F$:

  $f_{ij}$ is the number of times $w_i$ appear in context $c_j$. 

\[ f_{ij} \]
Pointwise Mutual Information (3)
Pointwise Mutual Information (3)

- Create word-context frequency matrix $F$:

$$F$$
Pointwise Mutual Information (3)

- Create word-context frequency matrix $F$:
- Create PPMI matrix:
Pointwise Mutual Information (3)

- Create word-context frequency matrix $F$:
- Create PPMI matrix:
Pointwise Mutual Information (3)

- Create word-context frequency matrix $F$:

- Create PPMI matrix:

$$ p_{ij} = \frac{f_{ij}}{\sum_{j=1}^{n_r} \sum_{i=1}^{n_c} f_{ij}} $$
Pointwise Mutual Information (3)

- Create word-context frequency matrix $F$:

- Create PPMI matrix:

\[
p_{ij} = \frac{f_{ij}}{\sum_{j=1}^{n_r} \sum_{i=1}^{n_c} f_{ij}}
\]

\[
p_{*j} = \frac{\sum_{i=1}^{n_c} f_{ij}}{\sum_{j=1}^{n_r} \sum_{i=1}^{n_c} f_{ij}}
\]
Pointwise Mutual Information (3)

- Create word-context frequency matrix $F$:

- Create PPMI matrix:

$$p_{ij} = \frac{f_{ij}}{\sum_{j=1}^{n_r} \sum_{i=1}^{n_c} f_{ij}}$$

$$p_{*j} = \frac{\sum_{i=1}^{n_c} f_{ij}}{\sum_{j=1}^{n_r} \sum_{i=1}^{n_c} f_{ij}}$$

$$p_{i*} = \frac{\sum_{j=1}^{n_c} f_{ij}}{\sum_{j=1}^{n_r} \sum_{i=1}^{n_c} f_{ij}}$$
Pointwise Mutual Information (3)

- Create word-context frequency matrix $F$:

- Create PPMI matrix:

$$p_{ij} = \frac{f_{ij}}{\sum_{j=1}^{n_r} \sum_{i=1}^{n_c} f_{ij}}$$

$$p_{*j} = \sum_{i=1}^{n_c} f_{ij} / \sum_{j=1}^{n_r} \sum_{i=1}^{n_c} f_{ij}$$

$$p_{i*} = \sum_{j=1}^{n_c} f_{ij} / \sum_{j=1}^{n_r} \sum_{i=1}^{n_c} f_{ij}$$

$$x_{ij} = \max\left(\log\left(\frac{p_{ij}}{p_{i*}p_{*j}}\right), 0\right)$$
Pointwise Mutual Information (4)
Pointwise Mutual Information (4)

- Given PPMI matrix:

\[
\begin{array}{c}
\downarrow \\
\downarrow \\
X \\
\downarrow \\
\end{array}
\]

context

words
Pointwise Mutual Information (4)

- Given PPMI matrix:
  - Word $W_i$ in the $i$-th row
Pointwise Mutual Information (4)

- Given PPMI matrix:
  - Word $w_i$ in the $i$-th row
  - Word $w_j$ in the $j$-th column
Pointwise Mutual Information (4)

- Given PPMI matrix:
  - Word $w_i$ in the $i$-th row
  - Word $w_j$ in the $j$-th column

$$\text{PPMI}(w_i, w_j, \text{left}) = x_{ij}^{\text{left}}$$
Pointwise Mutual Information (4)

- Given PPMI matrix:
- Word $w_i$ in the $i$-th row
- Word $w_j$ in the $j$-th column

\[
\begin{align*}
\text{PPMI}(w_i, w_j, \text{left}) &= x_{ij}^{\text{left}} \\
\text{PPMI}(w_i, w_j, \text{right}) &= x_{ij}^{\text{right}}
\end{align*}
\]
Pointwise Mutual Information (4)

- Given PPMI matrix:
  - Word $w_i$ in the $i$-th row
  - Word $w_j$ in the $j$-th column

\[
\begin{align*}
\text{PPMI}(w_i, w_j, \text{left}) &= x_{ij}^{\text{left}} \\
\text{PPMI}(w_i, w_j, \text{right}) &= x_{ij}^{\text{right}} \\
\text{PPMI}(w_j, w_i, \text{left}) &= x_{ji}^{\text{left}} \\
\end{align*}
\]
Pointwise Mutual Information (4)

- Given PPMI matrix:
  - Word \( W_i \) in the \( i \)-th row
  - Word \( W_j \) in the \( j \)-th column

\[
\begin{align*}
\text{PPMI}(w_i, w_j, \text{left}) &= x_{ij}^{\text{left}} \\
\text{PPMI}(w_i, w_j, \text{right}) &= x_{ij}^{\text{right}} \\
\text{PPMI}(w_j, w_i, \text{left}) &= x_{ji}^{\text{left}} \\
\text{PPMI}(w_j, w_i, \text{right}) &= x_{ji}^{\text{right}}
\end{align*}
\]
Pointwise Mutual Information (4)

- Given PPMI matrix:
- Word $W_i$ in the $i$-th **row**
- Word $W_j$ in the $j$-th **column**

\[
\begin{align*}
\text{PPMI}(w_i, w_j, \text{left}) &= x_{ij}^{\text{left}} \\
\text{PPMI}(w_i, w_j, \text{right}) &= x_{ij}^{\text{right}} \\
\text{PPMI}(w_j, w_i, \text{left}) &= x_{ji}^{\text{left}} \\
\text{PPMI}(w_j, w_i, \text{right}) &= x_{ji}^{\text{right}}
\end{align*}
\]

- For an $n$-tuple one can generate $2n(n-1)$ PPMI features
A vector space for domain similarity
A vector space for domain similarity

- Designed to capture the topic of a word.
A vector space for domain similarity

- Designed to capture the topic of a word.
- Construct a frequency matrix:
  - Rows: correspond to words in Wordnet
  - Columns: Nearby nouns

\[ D \]
A vector space for domain similarity

- Designed to capture the topic of a word.
- Construct a frequency matrix:
  - Rows: correspond to words in Wordnet
  - Columns: Nearby nouns
- Given a term $x_i$, search the corpus for it
A vector space for domain similarity

- Designed to capture the topic of a word.
- Construct a frequency matrix:
  - Rows: correspond to words in Wordnet
  - Columns: Nearby nouns
- Given a term $x_i$ search the corpus for it
- Choose $x_j$ a “noun” closest to the right/left of

\[ D \]
A vector space for domain similarity

- Designed to capture the topic of a word.
- Construct a frequency matrix:
  - Rows: correspond to words in Wordnet
  - Columns: Nearby nouns

- Given a term $x_i$ search the corpus for it
- Choose $x_j$ a “noun” closest to the right/left of
- And increment $d_{ij}$ by one.
A vector space for functional similarity

- Exactly the same as the domain similarity measures
  - Except that it is made using the "verbal" context
A vector space for functional similarity

- Exactly the same as the domain similarity measures
  - Except that it is made using the “verbal” context
- Construct a frequency matrix:

```
words
```

```markdown
S
```
A vector space for functional similarity

- Exactly the same as the domain similarity measures
  - Except that it is made using the “verbal” context
- Construct a frequency matrix:
  - Rows: correspond to terms in Wordnet
  - Columns: Nearby verbs
A vector space for functional similarity

- Exactly the same as the domain similarity measures
  - Except that it is made using the “verbal” context

- Construct a frequency matrix:
  - Rows: correspond to terms in Wordnet
  - Columns: Nearby verbs

- Given a term $x_i$, search the corpus for it
A vector space for functional similarity

- Exactly the same as the domain similarity measures
  - Except that it is made using the “verbal” context

- Construct a frequency matrix:
  - Rows: correspond to terms in Wordnet
  - Columns: Nearby verbs

- Given a term $x_i$ search the corpus for it
- Choose $x_j$ a “verbs” closest to the right/left of $x_i$
A vector space for functional similarity

- Exactly the same as the domain similarity measures
  - Except that it is made using the "verbal" context

- Construct a frequency matrix:
  - Rows: correspond to terms in Wordnet
  - Columns: Nearby verbs

- Given a term $x_i$, search the corpus for it
- Choose $x_j$ a "verbs" closest to the right/left of
- And increment $d_{ij}$ by one.
Using semantic and functional similarity
Using semantic and functional similarity

- Given the frequency matrix:
Using semantic and functional similarity

- Given the frequency matrix:
- Keep the lower-dimensional representation

\[ F = U \Sigma V \]
Using semantic and functional similarity

- Given the frequency matrix:
- Keep the lower-dimensional representation
  \[ F = U \Sigma V \]
- Keep the values corresponding to the \( k \) biggest eigenvalues
  \[ F \approx U_k \Sigma_k V_k \]
Using semantic and functional similarity

- Given the frequency matrix:
- Keep the lower-dimensional representation
  \[ F = U \Sigma V \]
- Keep the values corresponding to the \( k \) biggest eigenvalues
  \[ F \approx U_k \Sigma_k V_k \]
- Given word \( w_i \), \( U_k \Sigma_k^p \) is the corresponding vector
Using semantic and functional similarity

- Given the frequency matrix:

- Keep the lower-dimensional representation

  \[ F = U\Sigma V \]

- Keep the values corresponding to the \( k \) biggest eigenvalues

  \[ F \approx U_k \Sigma_k V_k \]

- Given word \( w_i \), \( U_k \Sigma_k^p \) is the corresponding vector

  \( p \in [0,1] \) is used to tune sensitivity with respect to eigenvalues
Using semantic and functional similarity

- Given the frequency matrix:
- Keep the lower-dimensional representation
  \[ F = U \Sigma V \]
- Keep the values corresponding to the \( k \) biggest eigenvalues
  \[ F \approx U_k \Sigma_k V_k \]
- Given word \( w_i \), \( U_k \Sigma_k^p \) is the corresponding vector
- \( p \in [0,1] \) is used to tune sensitivity with respect to eigenvalues
  - Given \( w_i \) and \( w_j \) to find: \( \text{Dom}(w_i, w_j, k, p) \)

Nearby verbs

\[ \text{words} \]
Using semantic and functional similarity

- Given the frequency matrix:
- Keep the lower-dimensional representation
  \[ F = U \Sigma V \]
- Keep the values corresponding to the \( k \) biggest eigenvalues
  \[ F \approx U_k \Sigma_k V_k \]
- Given word \( w_i \), \( U_k \Sigma_k^p \) is the corresponding vector
- \( p \in [0,1] \) is used to tune sensitivity with respect to eigenvalues
  - Given \( w_i \) and \( w_j \) to find: \( \text{Dom}(w_i, w_j, k, p) \)
  - find corresponding vectors
Using semantic and functional similarity

- Given the frequency matrix:
- Keep the lower-dimensional representation
  \[ F = U \Sigma V \]

- Keep the values corresponding to the \( k \) biggest eigenvalues
  \[ F \approx U_k \Sigma_k V_k \]
- Given word \( w_i \), \( U_k \Sigma_k^p \) is the corresponding vector
- \( p \in [0,1] \) is used to tune sensitivity with respect to eigenvalues
  - Given \( w_i \) and \( w_j \) to find: \( Dom(w_i, w_j, k, p) \)
  - find corresponding vectors
  - find cosine distance between the vectors
5-choice SAT tests

- 374 five-choice SAT questions

<table>
<thead>
<tr>
<th>Stem:</th>
<th>word:language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choices:</td>
<td></td>
</tr>
<tr>
<td>(1)</td>
<td>paint:portrait</td>
</tr>
<tr>
<td>(2)</td>
<td>poetry:rhythm</td>
</tr>
<tr>
<td>(3)</td>
<td>note:music</td>
</tr>
<tr>
<td>(4)</td>
<td>tale:story</td>
</tr>
<tr>
<td>(5)</td>
<td>week:year</td>
</tr>
</tbody>
</table>

Solution: (3) note:music
5-choice SAT tests

- 374 five-choice SAT questions

<table>
<thead>
<tr>
<th>Stem:</th>
<th>word:language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choices:</td>
<td>(1) paint:portrait</td>
</tr>
<tr>
<td></td>
<td>(2) poetry:rhythm</td>
</tr>
<tr>
<td></td>
<td>(3) note:music</td>
</tr>
<tr>
<td></td>
<td>(4) tale:story</td>
</tr>
<tr>
<td></td>
<td>(5) week:year</td>
</tr>
</tbody>
</table>

Solution: (3) note:music

- Could be converted into 5 4-tuples:

\[ \langle \text{word}, \text{language}, \text{note}, \text{music} \rangle \]
5-choice SAT tests

- 374 five-choice SAT questions

<table>
<thead>
<tr>
<th>Stem:</th>
<th>word:language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choices:</td>
<td>(1) paint:portrait</td>
</tr>
<tr>
<td></td>
<td>(2) poetry:rhythm</td>
</tr>
<tr>
<td></td>
<td>(3) note:music</td>
</tr>
<tr>
<td></td>
<td>(4) tale:story</td>
</tr>
<tr>
<td></td>
<td>(5) week:year</td>
</tr>
<tr>
<td>Solution:</td>
<td>(3) note:music</td>
</tr>
</tbody>
</table>

- Could be converted into 5 4-tuples:

  \[ \langle \text{word}, \text{language}, \text{note}, \text{music} \rangle \]

- Each positive 4-tuple \( \langle a, b, c, d \rangle \) could be converted to:
5-choice SAT tests

- 374 five-choice SAT questions

<table>
<thead>
<tr>
<th>Stem:</th>
<th>word:language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choices:</td>
<td></td>
</tr>
<tr>
<td>(1)</td>
<td>paint:portrait</td>
</tr>
<tr>
<td>(2)</td>
<td>poetry:rhythm</td>
</tr>
<tr>
<td>(3)</td>
<td>note:music</td>
</tr>
<tr>
<td>(4)</td>
<td>tale:story</td>
</tr>
<tr>
<td>(5)</td>
<td>week:year</td>
</tr>
<tr>
<td>Solution:</td>
<td>(3) note:music</td>
</tr>
</tbody>
</table>

- Could be converted into 5 4-tuples:

  \(<word, language, note, music>\)

- Each positive 4-tuple \(<a, b, c, d>\) could be converted to:

  \(<b, a, d, c>, <c, d, a, b>, <d, c, b, a>\)
Results on 5-choice SAT

The top ten results with the SAT analogy questions.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Reference</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Know-Best</td>
<td>Veale (2004)</td>
<td>43.0</td>
</tr>
<tr>
<td>k-means</td>
<td>Biçici &amp; Yuret (2006)</td>
<td>44.0</td>
</tr>
<tr>
<td>BagPack</td>
<td>Herdağdelen &amp; Baroni (2009)</td>
<td>44.1</td>
</tr>
<tr>
<td>VSM</td>
<td>Turney &amp; Littman (2005)</td>
<td>47.1</td>
</tr>
<tr>
<td>Dual-Space</td>
<td>Turney (2012)</td>
<td>51.1</td>
</tr>
<tr>
<td>BMI</td>
<td>Bollegala et al. (2009)</td>
<td>51.1</td>
</tr>
<tr>
<td>PairClass</td>
<td>Turney (2008b)</td>
<td>52.1</td>
</tr>
<tr>
<td>PERT</td>
<td>Turney (2006a)</td>
<td>53.5</td>
</tr>
<tr>
<td>SuperSim</td>
<td>—</td>
<td>54.8</td>
</tr>
<tr>
<td>LRA</td>
<td>Turney (2006b)</td>
<td>56.1</td>
</tr>
<tr>
<td>Human</td>
<td>Average college applicant</td>
<td>57.0</td>
</tr>
</tbody>
</table>
## Results on 5-choice SAT

The top ten results with the SAT analogy questions.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Reference</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Know-Best</td>
<td>Veale (2004)</td>
<td>43.0</td>
</tr>
<tr>
<td>k-means</td>
<td>Biçici &amp; Yuret (2006)</td>
<td>44.0</td>
</tr>
<tr>
<td>BagPack</td>
<td>Herdağdelen &amp; Baroni (2009)</td>
<td>44.1</td>
</tr>
<tr>
<td>VSM</td>
<td>Turney &amp; Littman (2005)</td>
<td>47.1</td>
</tr>
<tr>
<td>Dual-Space</td>
<td>Turney (2012)</td>
<td>51.1</td>
</tr>
<tr>
<td>BMI</td>
<td>Bollegala et al. (2009)</td>
<td>51.1</td>
</tr>
<tr>
<td>PairClass</td>
<td>Turney (2008b)</td>
<td>52.1</td>
</tr>
<tr>
<td>PERT</td>
<td>Turney (2006a)</td>
<td>53.5</td>
</tr>
<tr>
<td>SuperSim</td>
<td>—</td>
<td>54.8</td>
</tr>
<tr>
<td>LRA</td>
<td>Turney (2006b)</td>
<td>56.1</td>
</tr>
<tr>
<td>Human</td>
<td>Average college applicant</td>
<td>57.0</td>
</tr>
</tbody>
</table>

The results are not significantly different according to Fisher’s exact test at the 95% confidence level.
Results on 5-choice SAT

- The top ten results with the SAT analogy questions.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Reference</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Know-Best</td>
<td>Veale (2004)</td>
<td>43.0</td>
</tr>
<tr>
<td>k-means</td>
<td>Biçici &amp; Yuret (2006)</td>
<td>44.0</td>
</tr>
<tr>
<td>BagPack</td>
<td>Herdağdelen &amp; Baroni (2009)</td>
<td>44.1</td>
</tr>
<tr>
<td>VSM</td>
<td>Turney &amp; Littman (2005)</td>
<td>47.1</td>
</tr>
<tr>
<td>Dual-Space</td>
<td>Turney (2012)</td>
<td>51.1</td>
</tr>
<tr>
<td>BMI</td>
<td>Bollegala et al. (2009)</td>
<td>51.1</td>
</tr>
<tr>
<td>PairClass</td>
<td>Turney (2008b)</td>
<td>52.1</td>
</tr>
<tr>
<td>PERT</td>
<td>Turney (2006a)</td>
<td>53.5</td>
</tr>
<tr>
<td>SuperSim</td>
<td>—</td>
<td>54.8</td>
</tr>
<tr>
<td>LRA</td>
<td>Turney (2006b)</td>
<td>56.1</td>
</tr>
<tr>
<td>Human</td>
<td>Average college applicant</td>
<td>57.0</td>
</tr>
</tbody>
</table>

- SuperSim answers the SAT questions in a few minutes
- LRA requires nine days

not significantly different according to Fisher’s exact test at the 95% confidence level
SAT with 10 choices
SAT with 10 choices

- Adding more negative instances
SAT with 10 choices

- Adding more negative instances
- In general if $\langle a, b, c, d \rangle$ is positive $\langle a, d, c, b \rangle$ is negative
SAT with 10 choices

- Adding more negative instances
- In general if \( \langle a, b, c, d \rangle \) is positive \( \langle a, d, c, b \rangle \) is negative
- For example: Positive: \( \langle word, language, note, music \rangle \)
SAT with 10 choices

- Adding more negative instances
- In general if $\langle a, b, c, d \rangle$ is positive $\langle a, d, c, b \rangle$ is negative
- For example: Positive: $\langle word, language, note, music \rangle$
  Negative: $\langle word, music, note, language \rangle$
SAT with 10 choices

- Adding more negative instances
- In general if $\langle a, b, c, d \rangle$ is positive $\langle a, d, c, b \rangle$ is negative
- For example: Positive: $\langle$word, language, note, music$\rangle$
  
  Negative: $\langle$word, music, note, language$\rangle$
- This generates 5 more negative instances
SAT with 10 choices

- Adding more negative instances
- In general if \( \langle a, b, c, d \rangle \) is positive \( \langle a, d, c, b \rangle \) is negative
- For example: **Positive:** \( \langle \text{word}, \text{language}, \text{note}, \text{music} \rangle \)
  **Negative:** \( \langle \text{word}, \text{music}, \text{note}, \text{language} \rangle \)
- This generates 5 more negative instances

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Features</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LF</td>
<td>PPMI</td>
</tr>
<tr>
<td>Dual-Space</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SuperSim</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SuperSim</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>SuperSim</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>SuperSim</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SuperSim</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SuperSim</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>SuperSim</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>SuperSim</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SuperSim</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
SemEval-2012 Task 2

- Class and subclasses labels + examples

```plaintext
CLASS-INCLUSION,Taxonomic
50.0 "weapon:spear"
...
34.7 "vegetable:carrot"
...
-1.9 "mammal:porpoise"
...
-29.8 "pen:ballpoint"
...
-55.1 "wheat:bread"
```

- Gather using Mechanical Turk:
- 75 subcategories
- Average of 41 word-pairs per subcategories
SemEval-2012 Task 2

- SuperSim Trained on 5-choice SAT and tested on SemEval data
- It gives the best correlation coefficient

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Reference</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>BUAP</td>
<td>Tovar et al. (2012)</td>
<td>0.014</td>
</tr>
<tr>
<td>Duluth-V2</td>
<td>Pedersen (2012)</td>
<td>0.038</td>
</tr>
<tr>
<td>Duluth-V1</td>
<td>Pedersen (2012)</td>
<td>0.039</td>
</tr>
<tr>
<td>Duluth-V0</td>
<td>Pedersen (2012)</td>
<td>0.050</td>
</tr>
<tr>
<td>UTD-SVM</td>
<td>Rink &amp; Harabagiu (2012)</td>
<td>0.116</td>
</tr>
<tr>
<td>UTD-NB</td>
<td>Rink &amp; Harabagiu (2012)</td>
<td>0.229</td>
</tr>
<tr>
<td>RNN-1600</td>
<td>Mikolov et al. (2013)</td>
<td>0.275</td>
</tr>
<tr>
<td>UTD-LDA</td>
<td>Rink &amp; Harabagiu (2013)</td>
<td>0.334</td>
</tr>
<tr>
<td>Com</td>
<td>Zhila et al. (2013)</td>
<td>0.353</td>
</tr>
<tr>
<td>SuperSim</td>
<td>—</td>
<td>0.408</td>
</tr>
</tbody>
</table>
Compositional similarity: the data

- Noun-modifier question based on WordNet

<table>
<thead>
<tr>
<th>Stem:</th>
<th>fantasy world</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choices:</td>
<td>(1) fairyland</td>
</tr>
<tr>
<td>(2) fantasy</td>
<td></td>
</tr>
<tr>
<td>(3) world</td>
<td></td>
</tr>
<tr>
<td>(4) phantasy</td>
<td></td>
</tr>
<tr>
<td>(5) universe</td>
<td></td>
</tr>
<tr>
<td>(6) ranter</td>
<td></td>
</tr>
<tr>
<td>(7) souring</td>
<td></td>
</tr>
<tr>
<td>Solution:</td>
<td>(1) fairyland</td>
</tr>
</tbody>
</table>

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

21
## Compositional similarity: the data

- **Noun-modifier question based on WordNet**

<table>
<thead>
<tr>
<th>Stem:</th>
<th>fantasy world</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choices:</td>
<td>(1) fairyland</td>
</tr>
<tr>
<td></td>
<td>(2) fantasy</td>
</tr>
<tr>
<td></td>
<td>(3) world</td>
</tr>
<tr>
<td></td>
<td>(4) phantasy</td>
</tr>
<tr>
<td></td>
<td>(5) universe</td>
</tr>
<tr>
<td></td>
<td>(6) ranter</td>
</tr>
<tr>
<td></td>
<td>(7) souring</td>
</tr>
<tr>
<td>Solution:</td>
<td>(1) fairyland</td>
</tr>
</tbody>
</table>

- Create tuples of the form: $\langle a, b, c \rangle$
Compositional similarity: the data

- Noun-modifier question based on WordNet

<table>
<thead>
<tr>
<th>Stem:</th>
<th>fantasy world</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choices:</td>
<td>(1) fairyland</td>
</tr>
<tr>
<td></td>
<td>(2) fantasy</td>
</tr>
<tr>
<td></td>
<td>(3) world</td>
</tr>
<tr>
<td></td>
<td>(4) phantasy</td>
</tr>
<tr>
<td></td>
<td>(5) universe</td>
</tr>
<tr>
<td></td>
<td>(6) ranter</td>
</tr>
<tr>
<td></td>
<td>(7) souring</td>
</tr>
</tbody>
</table>

Solution: (1) fairyland

- Create tuples of the form: \( \langle a, b, c \rangle \)
  - Example: \( \langle \text{fantasy}, \text{world}, \text{fairyland} \rangle \)
Compositional similarity: the data

- Noun-modifier question based on WordNet

<table>
<thead>
<tr>
<th>Stem:</th>
<th>fantasy world</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choices:</td>
<td>(1) fairyland</td>
</tr>
<tr>
<td></td>
<td>(2) fantasy</td>
</tr>
<tr>
<td></td>
<td>(3) world</td>
</tr>
<tr>
<td></td>
<td>(4) phantasy</td>
</tr>
<tr>
<td></td>
<td>(5) universe</td>
</tr>
<tr>
<td></td>
<td>(6) ranter</td>
</tr>
<tr>
<td></td>
<td>(7) souring</td>
</tr>
<tr>
<td>Solution:</td>
<td>(1) fairyland</td>
</tr>
</tbody>
</table>

- Create tuples of the form: $\langle a, b, c \rangle$
  - Example: $\langle$fantasy, world, fairyland$\rangle$
  - Any question gives one positive instance and six negative instance
Compositional similarity: 7-choices questions

- 680 questions for training
- 1,500 questions for testing

Total of 2,180 questions
Compositional similarity: 7-choices questions

- 680 questions for training
- 1,500 questions for testing
- Any question gives one positive instance and six negative instance

Total of 2,180 questions
Compositional similarity: 7-choices questions

- 680 questions for training
- 1,500 questions for testing
- Any question gives one positive instance and six negative instance
- And train a classifier given the tuple for probabilities

Total of 2,180 questions
Compositional similarity: 7-choices questions

- 680 questions for training
- 1,500 questions for testing
- Any question gives one positive instance and six negative instance
- And train a classifier given the tuple for probabilities

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>7-choices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vector addition</td>
<td>50.1</td>
</tr>
<tr>
<td>Element-wise multiplication</td>
<td>57.5</td>
</tr>
<tr>
<td>Dual-Space model</td>
<td>58.3</td>
</tr>
<tr>
<td>SuperSim</td>
<td>75.9</td>
</tr>
<tr>
<td>Holistic model</td>
<td>81.6</td>
</tr>
</tbody>
</table>

Total of 2,180 questions
Compositional similarity: 7-choices questions

- 680 questions for training
- 1,500 questions for testing

Any question gives one positive instance and six negative instance

And train a classifier given the tuple for probabilities

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>7-choices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vector addition</td>
<td>50.1</td>
</tr>
<tr>
<td>Element-wise multiplication</td>
<td>57.5</td>
</tr>
<tr>
<td>Dual-Space model</td>
<td>58.3</td>
</tr>
<tr>
<td>SuperSim</td>
<td>75.9</td>
</tr>
<tr>
<td>Holistic model</td>
<td>81.6</td>
</tr>
</tbody>
</table>

The holistic approach is noncompositional.

- The stem bigram is represented by a single context vector
- As if it were a unigram.
Compositional similarity: 14-choices questions
Compositional similarity: 14-choices questions

- Any question gives one positive instance and six negative instance
Compositional similarity: 14-choices questions

- Any question gives one positive instance and six negative instance
- Positive instance: \( \langle a, b, c \rangle \)
Compositional similarity: 14-choices questions

- Any question gives one positive instance and six negative instance.
- Positive instance: \( \langle a, b, c \rangle \)
- Negative instance: \( \langle b, a, c \rangle \) e.g. word fantasy \( \neq \) wonderland
Compositional similarity: 14-choices questions

- Any question gives one positive instance and six negative instances.
- Positive instance: $\langle a, b, c \rangle$
- Negative instance: $\langle b, a, c \rangle$ e.g. word fantasy $\neq$ wonderland
- This gives 7 more negative instances (14-choices)
Compositional similarity: 14-choices questions

- Any question gives one **positive** instance and six **negative** instance
- **Positive** instance: \( \langle a, b, c \rangle \)
- **Negative** instance: \( \langle b, a, c \rangle \) e.g. word fantasy \( \neq \) wonderland
- This gives 7 more negative instances (14-choices)

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>7-choices</td>
</tr>
<tr>
<td>Vector addition</td>
<td>50.1</td>
</tr>
<tr>
<td>Element-wise multiplication</td>
<td>57.5</td>
</tr>
<tr>
<td>Dual-Space model</td>
<td>58.3</td>
</tr>
<tr>
<td>SuperSim</td>
<td>75.9</td>
</tr>
<tr>
<td>Holistic model</td>
<td>81.6</td>
</tr>
</tbody>
</table>
Compositional similarity: ablation experiment

- Analyzing effect of each feature type on the 14-choice test

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>LF</th>
<th>PPMI</th>
<th>Dom</th>
<th>Fun</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dual-Space</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>41.5</td>
</tr>
<tr>
<td>SuperSim</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>68.0</td>
</tr>
<tr>
<td>SuperSim</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>66.6</td>
</tr>
<tr>
<td>SuperSim</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>52.3</td>
</tr>
<tr>
<td>SuperSim</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>69.3</td>
</tr>
<tr>
<td>SuperSim</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>65.9</td>
</tr>
<tr>
<td>SuperSim</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14.1</td>
</tr>
<tr>
<td>SuperSim</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>59.7</td>
</tr>
<tr>
<td>SuperSim</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>34.6</td>
</tr>
<tr>
<td>SuperSim</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>32.9</td>
</tr>
</tbody>
</table>

- PPMI features are the most important
Compositional similarity: closer look at PPMI
Compositional similarity: closer look at PPMI

- PPMI features for \( \langle a, b, c \rangle \) into three subsets:
  \[ \langle a, b \rangle, \langle a, c \rangle, \langle b, c \rangle \]
Compositional similarity: closer look at PPMI

- PPMI features for $\langle a, b, c \rangle$ into three subsets:
  $\langle a, b \rangle$, $\langle a, c \rangle$, $\langle b, c \rangle$

- For example for $\langle a, b \rangle$: $\text{PPMI}(a, b, \text{left})$, $\text{PPMI}(a, b, \text{right})$
  $\text{PPMI}(b, a, \text{left})$, $\text{PPMI}(b, a, \text{right})$
Compositional similarity: closer look at PPMI

- PPMI features for $\langle a, b, c \rangle$ into three subsets:
  $\langle a, b \rangle$, $\langle a, c \rangle$, $\langle b, c \rangle$

- For example for $\langle a, b \rangle$:

<table>
<thead>
<tr>
<th>PPMI feature subsets</th>
<th>$\langle a, b \rangle$</th>
<th>$\langle a, c \rangle$</th>
<th>$\langle b, c \rangle$</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\langle a, b, left \rangle$, $\langle a, b, right \rangle$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>68.0</td>
</tr>
<tr>
<td>$\langle b, a, left \rangle$, $\langle b, a, right \rangle$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>59.9</td>
</tr>
<tr>
<td>$\langle b, a, left \rangle$, $\langle b, a, right \rangle$</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>65.4</td>
</tr>
<tr>
<td>$\langle b, a, left \rangle$, $\langle b, a, right \rangle$</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>66.5</td>
</tr>
<tr>
<td>$\langle a, b, left \rangle$, $\langle a, b, right \rangle$</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>62.6</td>
</tr>
<tr>
<td>$\langle a, b, left \rangle$, $\langle a, b, right \rangle$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>58.1</td>
</tr>
<tr>
<td>$\langle a, b, left \rangle$, $\langle a, b, right \rangle$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>55.6</td>
</tr>
<tr>
<td>$\langle a, b, left \rangle$, $\langle a, b, right \rangle$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>52.3</td>
</tr>
</tbody>
</table>
Compositional similarity: closer look at PPMI

- PPMI features for $\langle a, b, c \rangle$ into three subsets:
  $\langle a, b \rangle$, $\langle a, c \rangle$, $\langle b, c \rangle$

- For example for $\langle a, b \rangle$: $\text{PPMI}(a, b, \text{left}), \text{PPMI}(a, b, \text{right})$
  $\text{PPMI}(b, a, \text{left}), \text{PPMI}(b, a, \text{right})$

$\downarrow \langle a, b \rangle$ subset are more important.

<table>
<thead>
<tr>
<th>PPMI feature subsets</th>
<th>$\langle a, b \rangle$</th>
<th>$\langle a, c \rangle$</th>
<th>$\langle b, c \rangle$</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>68.0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>59.9</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>65.4</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>67.5</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>62.6</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>58.1</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>55.6</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>52.3</td>
</tr>
</tbody>
</table>
Compositional similarity: holistic training
Compositional similarity: holistic training

- A holistic training data
Compositional similarity: holistic training

- A holistic training data

<table>
<thead>
<tr>
<th>Stem:</th>
<th>search engine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choices:</td>
<td></td>
</tr>
<tr>
<td>(1) search</td>
<td>engine</td>
</tr>
<tr>
<td>(2) search</td>
<td>language</td>
</tr>
<tr>
<td>(3) engine</td>
<td>warrant</td>
</tr>
<tr>
<td>(4) search</td>
<td>engine</td>
</tr>
<tr>
<td>(5) search</td>
<td>engine</td>
</tr>
<tr>
<td>(6) diesel</td>
<td>engine</td>
</tr>
<tr>
<td>(7) steam</td>
<td>engine</td>
</tr>
</tbody>
</table>

Solution: (1) search_engine
Compositional similarity: holistic training

- A holistic training data
- Extract noun-modifier pairs from WordNet

<table>
<thead>
<tr>
<th>Stem:</th>
<th>search engine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choices:</td>
<td></td>
</tr>
<tr>
<td>(1) search_engine</td>
<td></td>
</tr>
<tr>
<td>(2) search</td>
<td></td>
</tr>
<tr>
<td>(3) engine</td>
<td></td>
</tr>
<tr>
<td>(4) search_language</td>
<td></td>
</tr>
<tr>
<td>(5) search_warrant</td>
<td></td>
</tr>
<tr>
<td>(6) diesel_engine</td>
<td></td>
</tr>
<tr>
<td>(7) steam_engine</td>
<td></td>
</tr>
</tbody>
</table>

Solution: (1) search_engine
Compositional similarity: holistic training

- A holistic training data
- Extract noun-modifier pairs from WordNet
- Call $a_b$ a pseudo-unigram and treat it as unigram

<table>
<thead>
<tr>
<th>Stem</th>
<th>Choices</th>
</tr>
</thead>
<tbody>
<tr>
<td>search_engine</td>
<td>(1) search_engine</td>
</tr>
<tr>
<td></td>
<td>(2) search</td>
</tr>
<tr>
<td></td>
<td>(3) engine</td>
</tr>
<tr>
<td></td>
<td>(4) search_language</td>
</tr>
<tr>
<td></td>
<td>(5) search_warrant</td>
</tr>
<tr>
<td></td>
<td>(6) diesel_engine</td>
</tr>
<tr>
<td></td>
<td>(7) steam_engine</td>
</tr>
</tbody>
</table>

Solution: (1) search_engine
Compositional similarity: holistic training

- A holistic training data
- Extract noun-modifier pairs from WordNet
- Call $a\_b$ a pseudo-unigram and treat it as unigram
- Use the components as distracters

<table>
<thead>
<tr>
<th>Stem:</th>
<th>search engine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choices:</td>
<td>(1) search_engine</td>
</tr>
<tr>
<td></td>
<td>(2) search</td>
</tr>
<tr>
<td></td>
<td>(3) engine</td>
</tr>
<tr>
<td></td>
<td>(4) search_language</td>
</tr>
<tr>
<td></td>
<td>(5) search_warrant</td>
</tr>
<tr>
<td></td>
<td>(6) diesel_engine</td>
</tr>
<tr>
<td></td>
<td>(7) steam_engine</td>
</tr>
<tr>
<td>Solution:</td>
<td>(1) search_engine</td>
</tr>
</tbody>
</table>
Compositional similarity: holistic training (2)

- Training on the holistic questions: “Holistic”
- Compared with the standard training
- Test is the standard testing

<table>
<thead>
<tr>
<th>Training</th>
<th>Correct 7-choices</th>
<th>Correct 14-choices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holistic</td>
<td>61.8</td>
<td>54.4</td>
</tr>
<tr>
<td>Standard</td>
<td>75.9</td>
<td>68.0</td>
</tr>
</tbody>
</table>

- There is a drop when training with the holistic samples
- Not very clear, but seems to be because of the nature of the
References

SVM

- **Primal form:**
  \[
  \begin{aligned}
  &\min_{\beta} \frac{1}{2}\|\beta\|^2 \\
  &y_i (\beta \cdot x_i) - 1 \geq 0
  \end{aligned}
  \]

- **Relaxed form:**
  \[
  \begin{aligned}
  &\min_{\beta} \frac{1}{2}\|\beta\|^2 + C \sum_i \varepsilon_i \\
  &y_i (\beta \cdot x_i) - 1 \geq -\varepsilon_i
  \end{aligned}
  \]

- **Dual form:**
  \[
  \begin{aligned}
  &\max_\alpha \sum \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \\
  &\sum_i \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C
  \end{aligned}
  \]
Proof: Hinge regression closed form

- Suppose we want to minimize:

\[
\|Ax - b\|^2 + \|\Gamma x\|^2
\]

\[
x = \left( A^T A + \Gamma^T \Gamma \right)^{-1} A^T b
\]
Proof: Hinge regression closed form

Proof:

\[ L = \frac{1}{2} (Ax - b)^T (Ax - b) \]

\[ \frac{dL}{dx} = A^T (Ax - b) = 0 \]

\[ x = (A^T A)^{-1} A^T b \]
Proof: Hinge regression closed form

Proof:

\[ L = \frac{1}{2} (A x - b)^T (A x - b) + \frac{1}{2} (\Gamma x)^T (\Gamma x) \]

\[ \frac{dL}{dx} = A^T (A x - b) + \Gamma^T (\Gamma x) = 0 \]

\[ x = (A^T A + \Gamma^T \Gamma)^{-1} A^T b \]