Pattern-based Solutions to Limitations of Leading Word Embeddings

Roy Schwartz
University of Washington NLP Seminar, February 8th, 2016

Joint work with Roi Reichart and Ari Rappoport
• **Background**
  – *Word embeddings are great!*

• **Problem**
  – *They also suffer from major limitations*

• **Solution**
  – *Pattern-based methods overcome many of these limitations*
Publications

- **Symmetric Patterns: Fast and Enhanced Representation of Verbs and Adjectives** (Schwartz, Reichart & Rappoport, *in review*)
- **Symmetric Pattern Based Word Embeddings for Improved Word Similarity Prediction** (Schwartz, Reichart & Rappoport, *CoNLL 2015*)
- Minimally Supervised Classification to Semantic Categories using Automatically Acquired Symmetric Patterns (Schwartz, Reichart & Rappoport, *COLING 2014*)
- **Authorship Attribution of Micro-Messages** (Schwartz, Tsur, Rappoport & Koppel, *EMNLP 2013*)
- **Learnability-based Syntactic Annotation Design** (Schwartz, Abend & Rappoport, *COLING 2012*)
- **Neutralizing Linguistically Problematic Annotations in Unsupervised Dependency Parsing Evaluation** (Schwartz, Abend, Reichart & Rappoport, *ACL 2011*)
Word Embedding Models
A.K.A. Vector Space Models

• Design vector representations of linguistic units (words, phrases, …)

• Distributional Semantics hypothesis (Harris, 1954)
  – Words that occur in similar contexts are likely to have similar meanings
Word Embedding Models
A.K.A Vector Space Models

• Design vector representations of linguistic units (words, phrases, ...)

• Distributional Semantics hypothesis (Harris, 1954)
  – Words that occur in similar contexts are likely to have similar meanings
Word Embedding Models
A.K.A Vector Space Models

- Design vector representations of linguistic units (words, phrases, ...)

- Distributional Semantics hypothesis (Harris, 1954)
  - Words that occur in similar contexts are likely to have similar meanings

- Most embedding models use *bag-of-words* contexts
  - Without taking into account *order* or *directionality*
Word Embedding Models
A.K.A Vector Space Models

• Design vector representations of linguistic units (words, phrases, ...)

• Distributional Semantics hypothesis (Harris, 1954)
  – Words that occur in similar contexts are likely to have similar meanings

• Most embedding models use bag-of-words contexts
  – Without taking into account order or directionality

  John is a good friend of Mary
Word Embeddings are Great, But...

- **Great results** on word relatedness, word analogy, synonym detection, etc. (Baroni et al., 2014)

- Also useful for downstream applications
  - Sentiment Analysis (Maas et al., ACL 2011, Socher et al., EMNLP 2013)
  - Parsing (Socher et al, EMNLP 2012; Lazaridou et al., EMNLP 2013)
Word Embeddings are Great, But...

• **Great results** on word relatedness, word analogy, synonym detection, etc. (Baroni et al., 2014)

• Also useful for downstream applications
  – Sentiment Analysis (Maas et al., ACL 2011, Socher et al., EMNLP 2013)
  – Parsing (Socher et al, EMNLP 2012; Lazaridou et al., EMNLP 2013)

• **But …**

• They also suffer from major limitations
Limitations of Word Embeddings

50 shades of “Relatedness”

• Failure to distinguish between correlation and similarity (Schwartz et al., CoNLL 2015)
  – cup/coffee vs. cup/glass
  – dog/leash vs. dog/cat
  – car/wheel vs. car/train
Limitations of Word Embeddings

50 shades of “Relatedness”

- Failure to distinguish between correlation and similarity (Schwartz et al., CoNLL 2015)
  - cup/coffee vs. cup/glass
  - dog/leash vs. dog/cat
  - car/wheel vs. car/train

- Failure to distinguish between similarity and (dis)similarity (Schwartz et al., CoNLL 2015)
  - good/great vs. good/bad
  - big/large vs. big/small
Limitations of Word Embeddings
50 shades of “Relatedness”

• Failure to distinguish between correlation and similarity (Schwartz et al., CoNLL 2015)
  – cup/coffee vs. cup/glass
  – dog/leash vs. dog/cat
  – car/wheel vs. car/train

• Failure to distinguish between similarity and (dis)similarity (Schwartz et al., CoNLL 2015)
  – good/great vs. good/bad
  – big/large vs. big/small

• Failure to capture hyponyms and entailment (Levy et al., NAACL 2015)
  – dog/animal, flu/fever
Limitations of Word Embeddings

No **Attributive** Knowledge

- Word embeddings are very good at capturing taxonomic properties
  - *cat, dog* and *elephant* belong to the same class (*animals*)
Limitations of Word Embeddings

No *Attributive* Knowledge

- Word embeddings are very good at capturing taxonomic properties
  - *cat*, *dog* and *elephant* belong to the same class (*animals*)

- They are much worse at capturing *attributive* properties (Rubinstein, Levi, Schwartz and Rappoport, ACL 2015)
  - *bananas*, *the sun* and *school buses* share the same color (*yellow*)
Limitations of Word Embeddings

No **Attributive** Knowledge

- Word embeddings are very good at capturing taxonomic properties
  - *cat, dog* and *elephant* belong to the same class (*animals*)

- They are much worse at capturing **attributive** properties (Rubinstein, Levi, Schwartz and Rappoport, ACL 2015)
  - *bananas, the sun* and *school buses* share the same color (*yellow*)
Limitations of Word Embeddings

Failure to Model Verb Similarity

• Verbs received relatively little attention in the word embedding literature
  – Significantly less than nouns
  – Very few verb datasets
Limitations of Word Embeddings

Failure to Model **Verb** Similarity

- Verbs received relatively little attention in the word embedding literature
  - Significantly less than **nouns**
  - Very few verb datasets

- Word embeddings perform substantially worse on **verb** similarity, as compared to **noun** similarity
  (**Schwartz** et al., CoNLL 2015; **Schwartz** et al., *in review*)
Limitations of Word Embeddings

Failure to Model *Verb* Similarity

• Verbs received relatively little attention in the word embedding literature
  – Significantly less than *nouns*
  – Very few verb datasets

• Word embeddings perform substantially worse on *verb* similarity, as compared to *noun* similarity (*Schwartz* et al., CoNLL 2015; *Schwartz* et al., *in review*)

• Spearman’s $\rho$ scores on SimLex999 (*Hill* et al., 2014):

<table>
<thead>
<tr>
<th>Model</th>
<th>Nouns</th>
<th>Verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>GloVe (Pennington et al., 2014)</td>
<td>0.377</td>
<td>0.163</td>
</tr>
<tr>
<td>word2vec skip-gram (Mikolov et al., 2013)</td>
<td>0.501</td>
<td>0.307</td>
</tr>
</tbody>
</table>
Recap:
Shortcomings of Word Embeddings

• They do not support distinctions finer than “relatedness”
  Similarity, dissimilarity, hyponymy, entailment …

• They fail to capture attributive similarity
  Bananas and school buses are yellow, elephants and mountains are large

• Their suffer from low performance on verb similarity
Solution:
Lexico-syntactic Patterns

• Patterns are sequences of *words* and *wildcards*
  – “*X and Y***”
  – “*X is a Y***”
  – “*wow, what a great X!***”
Solution: Lexico-syntactic Patterns

• Patterns are sequences of words and wildcards
  – “X and Y”
  – “X is a Y”
  – “wow, what a great X!”

• Hearst (1992) introduced the concept of patterns
  – Used “X such as Y” to detect hyponyms (“animals such as dogs”)
  – This method is still considered one of the most efficient ways of extracting hyponyms
Relation Extraction Using Patterns

• Patterns were found useful for recognizing other coarse-grained relations:
  – Antonyms (opposite meaning, Lin et al., 2003)
  – General verb relations (happens-before, stronger-than, Chklovski and Pantel, 2004)

• Patterns can also represent a wide range of semantic relations from different domains
  – Geography: capital-of, river-in (Davidov, Rappoport & Koppel, ACL 2007)
  – Technology: accessory-of (Davidov & Rappoport, ACL 2008)
Relation Extraction Using Patterns

• Patterns were found useful for recognizing other coarse-grained relations:
  – Antonyms (opposite meaning, Lin et al., 2003)
  – General verb relations (happens-before, stronger-than, Chklovskii and Pantel, 2004)

• Patterns can also represent a wide range of semantic relations from different domains
  – Geography: capital-of, river-in (Davidov, Rappoport & Koppel, ACL 2007)
  – Technology: accessory-of (Davidov & Rappoport, ACL 2008)

• **Symmetric Patterns**
Symmetric Patterns

- \( X \) and \( Y \)
- from \( X \) to \( Y \)
- \( X \) or \( Y \)
- neither \( X \) nor \( Y \)
- \( X \) as well as \( Y \)
Symmetric Patterns

X and Y

beds and sofas

sofas and beds
Symmetric Patterns

X and Y

beds and sofas

sofas and beds

X is a Y

Rihanna is a singer

*singer is a Rihanna
Symmetric Patterns

• Words that co-occur in *symmetric patterns* often take the same semantic role
  – *John* and *Mary* went to school
  – Is it better to *walk* or *run*?
  – Jane is *smart* as well as *funny*
Symmetric Patterns for Word \textit{Similarity}

- Symmetric patterns have shown useful for capturing different aspects of word \textit{similarity} in semantic tasks
  - Lexical acquisition (Widdows & Dorow, COLING 2002),
  - Semantic clustering (Davidov & Rappoport, ACL 2006)
  - Construction of connotative lexicon (Feng et al., ACL 2013)
  - Minimally supervised word classification (Schwartz et al., COLING 2014)
Symmetric Patterns for Word Similarity

- Symmetric patterns have shown useful for capturing different aspects of word similarity in semantic tasks
  - Lexical acquisition (Widdows & Dorow, COLING 2002),
  - Semantic clustering (Davidov & Rappoport, ACL 2006)
  - Construction of connotative lexicon (Feng et al., ACL 2013)
  - Minimally supervised word classification (Schwartz et al., COLING 2014)

Symmetric-Pattern-based methods can overcome many of the limitations of general word embeddings!
Similarity vs. Relatedness

• Recall:
  – Related words are not necessarily similar (*cow/milk*)
  – Word embeddings (based on bag-of-words context) fail to make this distinction
Similarity vs. Relatedness

• Recall:
  – Related words are not necessarily similar (cow/milk)
  – Word embeddings (based on bag-of-words context) fail to make this distinction

<table>
<thead>
<tr>
<th>Type</th>
<th>Example</th>
<th>#instances</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Bag-of-words</td>
</tr>
<tr>
<td>similar</td>
<td>(car, train)</td>
<td>2418</td>
</tr>
<tr>
<td></td>
<td>(coffee, tea)</td>
<td>6324</td>
</tr>
<tr>
<td></td>
<td>(dog, cat)</td>
<td>3645</td>
</tr>
</tbody>
</table>
Similarity vs. Relatedness

- Recall:
  - Related words are not necessarily similar (cow/milk)
  - Word embeddings (based on bag-of-words context) fail to make this distinction

<table>
<thead>
<tr>
<th>Type</th>
<th>Example</th>
<th>#instances</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Bag-of-words</td>
</tr>
<tr>
<td>similar</td>
<td>(car, train)</td>
<td>2418</td>
</tr>
<tr>
<td></td>
<td>(coffee, tea)</td>
<td>6324</td>
</tr>
<tr>
<td></td>
<td>(dog, cat)</td>
<td>3645</td>
</tr>
<tr>
<td>related</td>
<td>(car, wheel)</td>
<td>333</td>
</tr>
<tr>
<td></td>
<td>(coffee, cup)</td>
<td>7247</td>
</tr>
<tr>
<td></td>
<td>(dog, walking)</td>
<td>2837</td>
</tr>
</tbody>
</table>
Similarity vs. Relatedness

- Recall:
  - Related words are not necessarily similar (cow/milk)
  - Word embeddings (based on bag-of-words context) fail to make this distinction

<table>
<thead>
<tr>
<th>Type</th>
<th>Example</th>
<th>#instances</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bag-of-words</td>
<td>Symmetric Patterns</td>
</tr>
<tr>
<td>similar</td>
<td>(car,train)</td>
<td>2418</td>
</tr>
<tr>
<td></td>
<td>(coffee,tea)</td>
<td>6324</td>
</tr>
<tr>
<td></td>
<td>(dog,cat)</td>
<td>3645</td>
</tr>
<tr>
<td>related</td>
<td>(car,wheel)</td>
<td>333</td>
</tr>
<tr>
<td></td>
<td>(coffee,cup)</td>
<td>7247</td>
</tr>
<tr>
<td></td>
<td>(dog,walking)</td>
<td>2837</td>
</tr>
</tbody>
</table>
Symmetric Patterns as Word Embeddings Contexts

**Schwartz**, Reichart and Rappoport, CoNLL 2015

\[ V_{\text{dog}} = \begin{pmatrix} \vdots & \vdots & \vdots \\ \text{count}(\text{dog}, w_i) & \vdots & \vdots \end{pmatrix} \]
Symmetric Patterns as Word Embeddings Contexts

Schwartz, Reichart and Rappoport, CoNLL 2015

\[ V_{\text{dog}} = \begin{cases} \text{count}(\text{dog}, w_i) \end{cases} \quad \text{X} \quad V_{\text{dog}}^{SP} = \begin{cases} \text{symmetric-pattern\_count}(\text{dog}, w_i) \end{cases} \]
Symmetric Patterns as Word Embeddings Contexts

Schwartz, Reichart and Rappoport, CoNLL 2015

The goal:
Distinguish between similarity and relatedness
**Similar Contexts**

dog, cat

\[ V_{dog} = \begin{pmatrix} \text{count}(dog, \text{cat}) \\ \vdots \\ \vdots \end{pmatrix} \quad V_{SP}^{dog} = \begin{pmatrix} \text{symmetric-pattern\_count}(dog, \text{cat}) \\ \vdots \\ \vdots \end{pmatrix} \]

Pattern-based Solutions to Limitations of Leading Word Embeddings @ Roy Schwartz
Similar Contexts

dog, cat

\[ V_{dog} = \begin{pmatrix} \text{count}(\text{dog}, \text{cat}) \\ \vdots \\ \vdots \end{pmatrix}, \quad V^{SP}_{dog} = \begin{pmatrix} \text{symmetric-pattern\_count}(\text{dog}, \text{cat}) \\ \vdots \\ \vdots \end{pmatrix} \]

positive
small/zero
Pattern-based Solutions to Limitations of Leading Word Embeddings @ Roy Schwartz
**Related Contexts**

dog, leash

\[
V_{\text{dog}} = \begin{pmatrix}
\vdots \\
\vcenter{\text{count(dog, leash)}} \\
\vdots
\end{pmatrix}
\]

\[
V^{SP} = \begin{pmatrix}
\vdots \\
\vcenter{\text{symmetric-pattern_count(dog, leash)}} \\
\vdots
\end{pmatrix}
\]

positive
small/zero

Pattern-based Solutions to Limitations of Leading Word Embeddings @ Roy Schwartz
Symmetric-pattern embeddings distinguish between *similarity* and *relatedness*
Similarity vs. Dissimilarity

• Recall:
  – Word embeddings fail to distinguish between similar and opposite pairs of words (good/great vs. good/bad)
Similarity vs. Dissimilarity

• Recall:
  – Word embeddings fail to distinguish between similar and opposite pairs of words (good/great vs. good/bad)

• Some patterns are indicative of antonymy (Lin et al. 2003)
  – Antonym patterns = { “either X or Y”, “from X to Y” }
  – either big or small, from poverty to richness
Similarity vs. Dissimilarity

• Recall:
  – Word embeddings fail to distinguish between similar and opposite pairs of words (good/great vs. good/bad)

• Some patterns are indicative of antonymy (Lin et al. 2003)
  – Antonym patterns = { “either X or Y”, “from X to Y” }
  – *either big or small, from poverty to richness*

<table>
<thead>
<tr>
<th>Type</th>
<th>Example</th>
<th>#instances</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Bag-of-words</td>
</tr>
<tr>
<td>related</td>
<td>(bad, dream)</td>
<td>1208</td>
</tr>
<tr>
<td>similar</td>
<td>(bad, evil)</td>
<td>561</td>
</tr>
<tr>
<td>opposite</td>
<td>(bad, good)</td>
<td>23532</td>
</tr>
</tbody>
</table>
Negative Weighting

• A feature of our model that assigns dissimilar vectors to antonym pairs
Negative Weighting

• A feature of our model that assigns dissimilar vectors to antonym pairs

• For each word $w$, compute $V_w^{AP}$ similarly to $V_w^{SP}$, but using the set of antonym patterns (AP)

\[ V_w^{SP+} = V_w^{SP} - \beta \cdot V_w^{AP} \]

❖ $\beta$ is tuned using a development set
Values for *Related* Contexts are **small**
bad, *dream*

\[
V^{SP^+}_{bad} = \left( \begin{array}{c}
\text{symmetric-pattern_count}(bad, dream) \\
\vdots \\
\vdots \\
\vdots \\
\text{antonym-pattern_count}(bad, dream) \\
\end{array} \right) - \beta \left( \begin{array}{c}
\text{symmetric-pattern_count}(bad, dream) \\
\vdots \\
\vdots \\
\vdots \\
\text{antonym-pattern_count}(bad, dream) \\
\end{array} \right)
\]

*positive*

*small/zero*
Values for *Related* Contexts are **small**
bad, **dream**

\[
V_{bad}^{SP^+} = \begin{pmatrix}
\text{symmetric-pattern\_count(bad,dream)} \\
\vdots \\
\text{symmetric-pattern\_count(bad,dream)} \\
\end{pmatrix} - \beta \begin{pmatrix}
\text{antonym-pattern\_count(bad,dream)} \\
\vdots \\
\text{antonym-pattern\_count(bad,dream)} \\
\end{pmatrix}
\]
Values for *Similar* Contexts are **large**

bad, **evil**

$$V_{bad}^{SP^+} = \left( \text{symmetric-pattern\_count}(bad,evil) \right) - \beta \left( \text{antonym-pattern\_count}(bad,evil) \right)$$
Values for Similar Contexts are large bad, evil

\[ V_{bad}^{SP^+} = \left( \text{symmetric-pattern\_count(bad,evil)} \right) - \beta \left( \text{antonym-pattern\_count(bad,evil)} \right) \]

Positive

small/zero
Values for **Opposite** Contexts are **small**
bad, **good**

\[
V_{bad}^{SP^+} = \left( \begin{array}{c} \text{symmetric-pattern\_count}(\text{bad,good}) \\ \vdots \\ \text{symmetric-pattern\_count}(\text{bad,good}) \end{array} \right) - \beta \left( \begin{array}{c} \text{antonym-pattern\_count}(\text{bad,good}) \\ \vdots \\ \text{antonym-pattern\_count}(\text{bad,good}) \end{array} \right)
\]

**positive**

**small/zero**

Pattern-based Solutions to Limitations of Leading Word Embeddings @ Roy Schwartz
Values for **Opposite** Contexts are **small**

bad, good

\[
V_{bad}^{SP^+} = \left( \begin{array}{c}
\text{symmatric-pattern\_count(bad,good)} \\
\vdots \\
\end{array} \right) - \beta \left( \begin{array}{c}
\text{antonym-pattern\_count(bad,good)} \\
\vdots \\
\end{array} \right)
\]

positive

small/zero
Values for **Opposite** Contexts are **small**

bad, good

\[
V_{bad}^{SP^+} = \left( \begin{array}{c}
\text{symmetric-pattern\_count(bad,good)} \\
\vdots \\
\end{array} \right) - \beta \cdot \left( \begin{array}{c}
\text{antonym-pattern\_count(bad,good)} \\
\vdots \\
\end{array} \right)
\]

Negative Weighting is able to distinguish between **similar** and **opposite** pairs

Positive | Small/zero
Experiments

• More about the SP$^+$ model
  – Set of symmetric pattern types is extracted from plain text using the (Davidov & Rappoport, 2006) algorithm
  – Positive Point-wise Mutual Information (PPMI) normalization
  – Personalized Page-rank like smoothing
Experiments

• More about the $\mathbf{SP}^+$ model
  – Set of symmetric pattern types is extracted from plain text using the 
    (Davidov & Rappoport, 2006) algorithm
  – Positive Point-wise Mutual Information (PPMI) normalization
  – Personalized Page-rank like smoothing

• Embeddings are generated using an 8G words corpus

• Evaluation: Word similarity task
  – SimLex999 dataset (Hill et al., 2014)
  – Compute a ranking based on the $\mathbf{SP}^+$ model’s prediction of the degree 
    of similarity between pairs of word
  – Compare this ranking to the one generated by human judgments
## Results

### SimLex999 Dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>Spearman’s $\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>GloVe (Pennington et al., 2014)</td>
<td>0.35</td>
</tr>
<tr>
<td>PPMI-Bag-of-words</td>
<td>0.423</td>
</tr>
<tr>
<td>word2vec CBO (Mikolov et al., 2013)</td>
<td>0.43</td>
</tr>
<tr>
<td>word2vec Dep (Levy and Goldberg, 2014)</td>
<td>0.436</td>
</tr>
<tr>
<td>NNSE (Murphy et al., 2012)</td>
<td>0.455</td>
</tr>
<tr>
<td>word2vec skip-gram (Mikolov et al., 2013)</td>
<td>0.462</td>
</tr>
<tr>
<td>$SP^+$ (Schwartz et al., 2015)</td>
<td><strong>0.517</strong></td>
</tr>
</tbody>
</table>
## Results

### SimLex999 Dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>Spearman’s $\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>GloVe (Pennington et al., 2014)</td>
<td>0.35</td>
</tr>
<tr>
<td>PPMI-Bag-of-words</td>
<td>0.423</td>
</tr>
<tr>
<td>word2vec CBOW (Mikolov et al., 2013)</td>
<td>0.43</td>
</tr>
<tr>
<td>word2vec Dep (Levy and Goldberg, 2014)</td>
<td>0.436</td>
</tr>
<tr>
<td>NNSE (Murphy et al., 2012)</td>
<td>0.455</td>
</tr>
<tr>
<td>word2vec skip-gram (Mikolov et al., 2013)</td>
<td>0.462</td>
</tr>
<tr>
<td>$\text{SP}^+$ (Schwartz et al., 2015)</td>
<td>0.517</td>
</tr>
</tbody>
</table>

$\uparrow$ 5.5%
# Results

## SimLex999 Dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>Spearman’s $\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>GloVe (Pennington et al., 2014)</td>
<td>0.35</td>
</tr>
<tr>
<td>PPMI-Bag-of-words</td>
<td>0.423</td>
</tr>
<tr>
<td>word2vec CBOB (Mikolov et al., 2013)</td>
<td>0.43</td>
</tr>
<tr>
<td>word2vec Dep (Levy and Goldberg, 2014)</td>
<td>0.436</td>
</tr>
<tr>
<td>NNSE (Murphy et al., 2012)</td>
<td>0.455</td>
</tr>
<tr>
<td>word2vec skip-gram (Mikolov et al., 2013)</td>
<td>0.462</td>
</tr>
<tr>
<td>$SP^+$(Schwartz et al., 2015)</td>
<td>0.517</td>
</tr>
</tbody>
</table>

5.5% Improvement

$$f_{joint}(w_i, w_j) = \alpha \cdot f_{SP^+}(w_i, w_j) + (1 - \alpha) \cdot f_{skip-gram}(w_i, w_j)$$
# Results

SimLex999 Dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>Spearman's $\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>GloVe (Pennington et al., 2014)</td>
<td>0.35</td>
</tr>
<tr>
<td>PPMI-Bag-of-words</td>
<td>0.423</td>
</tr>
<tr>
<td>word2vec CBO (Mikolov et al., 2013)</td>
<td>0.43</td>
</tr>
<tr>
<td>word2vec Dep (Levy and Goldberg, 2014)</td>
<td>0.436</td>
</tr>
<tr>
<td>NNSE (Murphy et al., 2012)</td>
<td>0.455</td>
</tr>
<tr>
<td>word2vec skip-gram (Mikolov et al., 2013)</td>
<td>0.462</td>
</tr>
<tr>
<td>$SP^+$ (Schwartz et al., 2015)</td>
<td>0.517</td>
</tr>
<tr>
<td>Joint</td>
<td>0.563</td>
</tr>
</tbody>
</table>

$$f_{joint}(w_i, w_j) = \alpha \cdot f_{SP^+}(w_i, w_j) + (1-\alpha) \cdot f_{skip-gram}(w_i, w_j)$$

Pattern-based Solutions to Limitations of Leading Word Embeddings @ Roy Schwartz
# Part-of-Speech Analysis

Spearman’s $\rho$ on the SimLex999 Dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>Adjective</th>
</tr>
</thead>
<tbody>
<tr>
<td>GloVe (Pennington et al., 2014)</td>
<td>0.571</td>
</tr>
<tr>
<td>PPMI-Bag-of-words</td>
<td>0.548</td>
</tr>
<tr>
<td>word2vec CBOW (Mikolov et al., 2013)</td>
<td>0.579</td>
</tr>
<tr>
<td>word2vec Dep (Levy and Goldberg, 2014)</td>
<td>0.54</td>
</tr>
<tr>
<td>NNSE (Murphy et al., 2012)</td>
<td>0.594</td>
</tr>
<tr>
<td>word2vec skip-gram (Mikolov et al., 2013)</td>
<td>0.604</td>
</tr>
<tr>
<td>\textbf{SP$^+$ (Schwartz et al., 2015)}</td>
<td>\textbf{0.663}</td>
</tr>
</tbody>
</table>
## Part-of-Speech Analysis

Spearman’s $\rho$ on the SimLex999 Dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>Adjective</th>
<th>Nouns</th>
</tr>
</thead>
<tbody>
<tr>
<td>GloVe (Pennington et al., 2014)</td>
<td>0.571</td>
<td>0.377</td>
</tr>
<tr>
<td>PPMI-Bag-of-words</td>
<td>0.548</td>
<td>0.451</td>
</tr>
<tr>
<td>word2vec CBO (Mikolov et al., 2013)</td>
<td>0.579</td>
<td>0.48</td>
</tr>
<tr>
<td>word2vec Dep (Levy and Goldberg, 2014)</td>
<td>0.54</td>
<td>0.449</td>
</tr>
<tr>
<td>NNSE (Murphy et al., 2012)</td>
<td>0.594</td>
<td>0.487</td>
</tr>
<tr>
<td>word2vec skip-gram (Mikolov et al., 2013)</td>
<td>0.604</td>
<td><strong>0.501</strong></td>
</tr>
<tr>
<td>SP$^+$ (Schwartz et al., 2015)</td>
<td><strong>0.663</strong></td>
<td>0.497</td>
</tr>
</tbody>
</table>
# Part-of-Speech Analysis

Spearman’s $\rho$ on the SimLex999 Dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>Adjective</th>
<th>Nouns</th>
<th>Verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>GloVe (Pennington et al., 2014)</td>
<td>0.571</td>
<td>0.377</td>
<td>0.163</td>
</tr>
<tr>
<td>PPMI-Bag-of-words</td>
<td>0.548</td>
<td>0.451</td>
<td>0.276</td>
</tr>
<tr>
<td>word2vec CBOW (Mikolov et al., 2013)</td>
<td>0.579</td>
<td>0.48</td>
<td>0.252</td>
</tr>
<tr>
<td>word2vec Dep (Levy and Goldberg, 2014)</td>
<td>0.54</td>
<td>0.449</td>
<td>0.376</td>
</tr>
<tr>
<td>NNSE (Murphy et al., 2012)</td>
<td>0.594</td>
<td>0.487</td>
<td>0.318</td>
</tr>
<tr>
<td>word2vec skip-gram (Mikolov et al., 2013)</td>
<td>0.604</td>
<td><strong>0.501</strong></td>
<td>0.307</td>
</tr>
<tr>
<td>\textbf{SP$^+$ (Schwartz et al., 2015)}</td>
<td><strong>0.663</strong></td>
<td>0.497</td>
<td><strong>0.578</strong></td>
</tr>
<tr>
<td>Model</td>
<td>Adjective</td>
<td>Nouns</td>
<td>Verbs</td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>-----------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>GloVe (Pennington et al., 2014)</td>
<td>0.571</td>
<td>0.377</td>
<td>0.163</td>
</tr>
<tr>
<td>PPMI-Bag-of-words</td>
<td>0.548</td>
<td>0.451</td>
<td>0.276</td>
</tr>
<tr>
<td>word2vec CBOW (Mikolov et al., 2013)</td>
<td>0.579</td>
<td>0.48</td>
<td>0.252</td>
</tr>
<tr>
<td>word2vec Dep (Levy and Goldberg, 2014)</td>
<td>0.54</td>
<td>0.449</td>
<td><strong>0.376</strong></td>
</tr>
<tr>
<td>NNSE (Murphy et al., 2012)</td>
<td>0.594</td>
<td>0.487</td>
<td>0.318</td>
</tr>
<tr>
<td>word2vec skip-gram (Mikolov et al., 2013)</td>
<td>0.604</td>
<td><strong>0.501</strong></td>
<td>0.307</td>
</tr>
<tr>
<td>SP+ (Schwartz et al., 2015)</td>
<td><strong>0.663</strong></td>
<td>0.497</td>
<td><strong>0.578</strong></td>
</tr>
</tbody>
</table>
## Part-of-Speech Analysis

Spearman’s ρ on the SimLex999 Dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>Adjective</th>
<th>Nouns</th>
<th>Verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>GloVe (Pennington et al., 2014)</td>
<td>0.571</td>
<td>0.377</td>
<td>0.163</td>
</tr>
<tr>
<td>PPMI-Bag-of-words</td>
<td>0.548</td>
<td>0.451</td>
<td>0.276</td>
</tr>
<tr>
<td>word2vec CBOB (Mikolov et al., 2013)</td>
<td>0.579</td>
<td>0.48</td>
<td>0.252</td>
</tr>
<tr>
<td>word2vec Dep (Levy and Goldberg, 2014)</td>
<td>0.54</td>
<td>0.449</td>
<td>0.376</td>
</tr>
<tr>
<td>NNSE (Murphy et al., 2012)</td>
<td>0.594</td>
<td>0.487</td>
<td>0.318</td>
</tr>
<tr>
<td>word2vec skip-gram (Mikolov et al., 2013)</td>
<td>0.604</td>
<td><strong>0.501</strong></td>
<td><strong>0.307</strong></td>
</tr>
<tr>
<td><strong>SP+ (Schwartz et al., 2015)</strong></td>
<td><strong>0.663</strong></td>
<td>0.497</td>
<td><strong>0.578</strong></td>
</tr>
</tbody>
</table>
Symmetric Patterns are Useful for Capturing Word Similarity

• Symmetric patterns overcome three of the limitations of general word embeddings
  – They capture similarity rather than relatedness
  – They distinguish between similar and opposite pairs
  – They capture verb similarity

• In our experiments on SimLex999
  – 5.5% improvement over six leading models
  – 10% improvement with a joint model
  – 20% improvement on verbs
Word Embeddings that Identify Antonyms
ACL 2015 Papers

• Revisiting Word Embedding for Contrasting Meaning (Chen et al.)

• Learning Semantic Word Embeddings based on Ordinal Knowledge Constraints (Liu et al.)

• A Multitask Objective to Inject Lexical Contrast into Distributional Semantics (Pham et al.)

• AutoExtend: Extending Word Embeddings to Embeddings for Synsets and Lexemes (Rothe and Schutze)
Word Embeddings that Identify Antonyms
ACL 2015 Papers

- Revisiting Word Embedding for Contrasting Meaning (Chen et al.)
- Learning Semantic Word Embeddings based on Ordinal Knowledge Constraints (Liu et al.)
- A Multitask Objective to Inject Lexical Contrast into Distributional Semantics (Pham et al.)
- AutoExtend: Extending Word Embeddings to Embeddings for Synsets and Lexemes (Rothe and Schutze)

Our $SP^+$ model is the only corpus-based model to identify antonyms (w/o using a dictionary or a thesaurus)
• **Background**
  – *Word embeddings are great!*

• **Problem**
  – *They also suffer from major limitations*

• **Solution**
  – *Pattern-based methods overcome many of these limitations*
The Skig-gram model’s Performance on Verb Similarity
(Schwartz et al., in review)

• The word2vec skip-gram model (Mikolov et al., 2013) verb similarity scores are particularly low

<table>
<thead>
<tr>
<th>Model</th>
<th>Nouns</th>
<th>Verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>word2vec skip-gram (Mikolov et al., 2013)</td>
<td>0.501</td>
<td>0.307</td>
</tr>
<tr>
<td>SP⁺ (Schwartz et al., 2015)</td>
<td>0.497</td>
<td>0.578</td>
</tr>
</tbody>
</table>

• We set to isolate the role of the context type in the performance of this model
Controlled Experiments

• We train the word2vec skip-gram model three times, each time with a different type of context
  – Bag-of-words contexts (Mikolov et al., 2013)
  – Dependency contexts (Levy & Goldberg, 2014)
  – Symmetric pattern contexts (Schwartz et al., 2015)

• All other modeling decisions are identical

• Experiments with the *verb* portion of SimLex999
Context Type Matters
Symmetric Patterns >> Bag-of-words

• Results on the verb portion of the SimLex999 Dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>Context Type</th>
<th>Spearman’s $\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>skip-gram</td>
<td>Bag-of-Words</td>
<td>0.307</td>
</tr>
<tr>
<td></td>
<td>Dependency Links</td>
<td>0.386</td>
</tr>
<tr>
<td></td>
<td>Symmetric Patterns</td>
<td>0.459</td>
</tr>
</tbody>
</table>
## Compact Model

<table>
<thead>
<tr>
<th>Model</th>
<th>Context Type</th>
<th>Verbs</th>
<th>#Contexts</th>
<th>Train Time (Mins)</th>
</tr>
</thead>
<tbody>
<tr>
<td>skip-gram</td>
<td>Bag-of-Words</td>
<td>0.307</td>
<td>13000M</td>
<td>320</td>
</tr>
<tr>
<td></td>
<td>Dependency Links</td>
<td>0.386</td>
<td>14500M</td>
<td>551</td>
</tr>
<tr>
<td></td>
<td><strong>Symmetric Patterns</strong></td>
<td><strong>0.459</strong></td>
<td><strong>270M</strong></td>
<td><strong>11</strong></td>
</tr>
</tbody>
</table>
Additive Value of *Symmetric Patterns* and Negative Weighting

<table>
<thead>
<tr>
<th>Model</th>
<th>Context Type</th>
<th>Verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>skip-gram</td>
<td>Bag-of-Words</td>
<td>0.307</td>
</tr>
<tr>
<td></td>
<td>Dependency Links</td>
<td>0.386</td>
</tr>
<tr>
<td></td>
<td><strong>Symmetric Patterns</strong></td>
<td><strong>0.459</strong></td>
</tr>
<tr>
<td>SP+ <em>(Schwartz et al., 2015)</em></td>
<td>Symmetric Patterns</td>
<td>0.578</td>
</tr>
</tbody>
</table>
Additive Value of Symmetric Patterns and Negative Weighting

<table>
<thead>
<tr>
<th>Model</th>
<th>Context Type</th>
<th>Verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>skip-gram</td>
<td>Bag-of-Words</td>
<td>0.307</td>
</tr>
<tr>
<td></td>
<td>Dependency Links</td>
<td>0.386</td>
</tr>
<tr>
<td><strong>Symmetric Patterns</strong></td>
<td><strong>Symmetric Patterns</strong></td>
<td><strong>0.459</strong></td>
</tr>
<tr>
<td>SP⁺ <em>(Schwartz et al., 2015)</em></td>
<td>Symmetric Patterns</td>
<td>0.578</td>
</tr>
<tr>
<td>SP⁻NW <em>(Schwartz et al., 2015)</em></td>
<td>Symmetric Patterns</td>
<td>0.441</td>
</tr>
</tbody>
</table>
# Additive Value of Symmetric Patterns and Negative Weighting

<table>
<thead>
<tr>
<th>Model</th>
<th>Context Type</th>
<th>Verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>skip-gram</td>
<td>Bag-of-Words</td>
<td>0.307</td>
</tr>
<tr>
<td></td>
<td>Dependency Links</td>
<td>0.386</td>
</tr>
<tr>
<td></td>
<td><strong>Symmetric Patterns</strong></td>
<td><strong>0.459</strong></td>
</tr>
<tr>
<td>$SP^+$ (Schwartz et al., 2015)</td>
<td>Symmetric Patterns</td>
<td>0.578</td>
</tr>
<tr>
<td>$SP^{-NW}$ (Schwartz et al., 2015)</td>
<td>Symmetric Patterns</td>
<td>0.441</td>
</tr>
</tbody>
</table>
Additive Value of Symmetric Patterns and Negative Weighting

<table>
<thead>
<tr>
<th>Model</th>
<th>Context Type</th>
<th>Verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>skip-gram</td>
<td>Bag-of-Words</td>
<td>0.307</td>
</tr>
<tr>
<td></td>
<td>Dependency Links</td>
<td>0.386</td>
</tr>
<tr>
<td></td>
<td>Symmetric Patterns</td>
<td>0.459</td>
</tr>
<tr>
<td>$SP^+$ (Schwartz et al., 2015)</td>
<td>Symmetric Patterns</td>
<td>0.578</td>
</tr>
<tr>
<td>$SP^{-NW}$ (Schwartz et al., 2015)</td>
<td>Symmetric Patterns</td>
<td>0.441</td>
</tr>
</tbody>
</table>
Additive Value of Symmetric Patterns and Negative Weighting

<table>
<thead>
<tr>
<th>Model</th>
<th>Context Type</th>
<th>Verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>skip-gram</td>
<td>Bag-of-Words</td>
<td>0.307</td>
</tr>
<tr>
<td></td>
<td>Dependency Links</td>
<td>0.386</td>
</tr>
<tr>
<td></td>
<td>Symmetric Patterns</td>
<td>0.459</td>
</tr>
<tr>
<td>SP⁺ (Schwartz et al., 2015)</td>
<td>Symmetric Patterns</td>
<td>0.578</td>
</tr>
<tr>
<td>SP⁻NW (Schwartz et al., 2015)</td>
<td>Symmetric Patterns</td>
<td>0.441</td>
</tr>
</tbody>
</table>

0 +~15% +~27% +~15%
Summary

• Patterns provide strong answers to the shortcomings of word embeddings

• They capture fine grained distinctions of word relatedness (similarity, dissimilarity, ...)

• They are particularly useful for modeling verb similarity
  – 15-27% improvement on a verb similarity task

• They are much more compact than other types of context
  – Training with pattern contexts takes ~2-3% of the training time with other types of context
Ongoing Work

• Negative weighting vs. negative *sampling*

• Use patterns to identify multiword expressions

• Experiment with symmetric patterns in a multilingual setup

• Semantics of prepositions

• Word analogies: patterns vs. vector operations

• Does order count? The asymmetry of symmetric patterns
  – now or never > *never or now*
Acknowledgments

• Many thanks to:
• Ari Rappoport
• Roi Reichart
• Dana Rubinstein
• Effi Levi
Acknowledgments

• Many thanks to:
  • Ari Rappoport
  • Roi Reichart
  • Dana Rubinstein
  • Effi Levi

• Surprise!
John and Mary are friends. They hang out together. Last night John moved out of town without telling Mary.
Surprise – why?

• surprising $\approx$ interesting

• Useful for NLP
  – Text summarization
  – Text search
  – News feed
  – Dialogue systems
  – Essay scoring
  – Detection of sarcasm/humor
  – ...

• Interesting from a cognitive perspective
Thank you!

• **Background**
  – *Word embeddings are great!*

• **Problem**
  – *They also suffer from major limitations*

• **Solution**
  – *Pattern-based methods overcome many of these limitations*