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

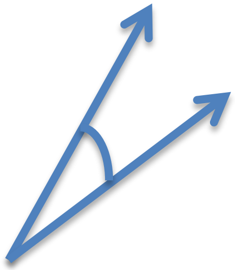
- *Background*
  - *Word embeddings are great!*
- *Problem*
  - *They also suffer from major limitations*
- *Solution*
  - *Pattern-based methods overcome many of these limitations*

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*)
- *How Well Do Distributional Models Capture Different Types of Semantic Knowledge?* (Rubinstein, Levi, **Schwartz** & Rappoport, *ACL 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

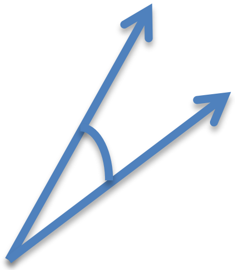
A.K.A Vector Space Models



- 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

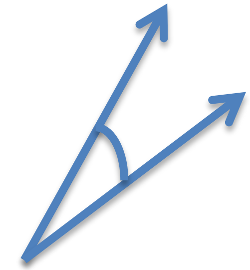
A.K.A Vector Space Models



- 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

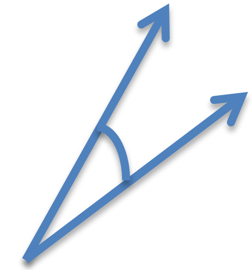
A.K.A Vector Space Models



- 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*

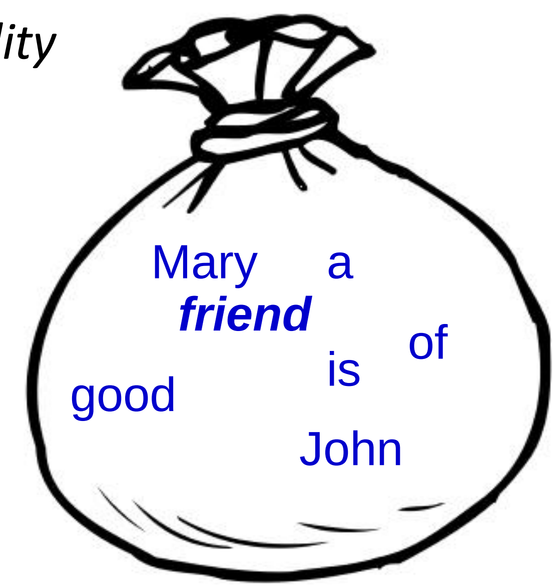
Word Embedding Models

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**

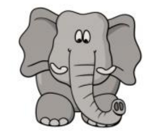
- Failure to distinguish between **similarity** and (**dis**)**similarity** (**Schwartz** et al., CoNLL 2015)
  - good/**great** vs. good/**bad**
  - big/**large** vs. big/**small**

- Failure to capture hyponyms and entailment (Levy et al., NAACL 2015)
  - dog/animal, flu/fever

Limitations of Word Embeddings

No ***Attributive*** Knowledge

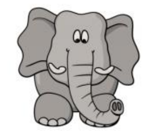
- Word embeddings are very good at capturing taxonomic properties
  - *cat, dog* and *elephant* belong to the same class (*animals*)



Limitations of Word Embeddings

No ***<u>Attributive</u>*** Knowledge

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

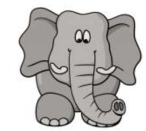


- They are much worse at capturing **attributive** properties (Rubinstein, Levi, **Schwartz** and Rappoport, ACL 2015)
  - *bananas, the sun* and *school buses* share the same color (*yellow*)

Limitations of Word Embeddings

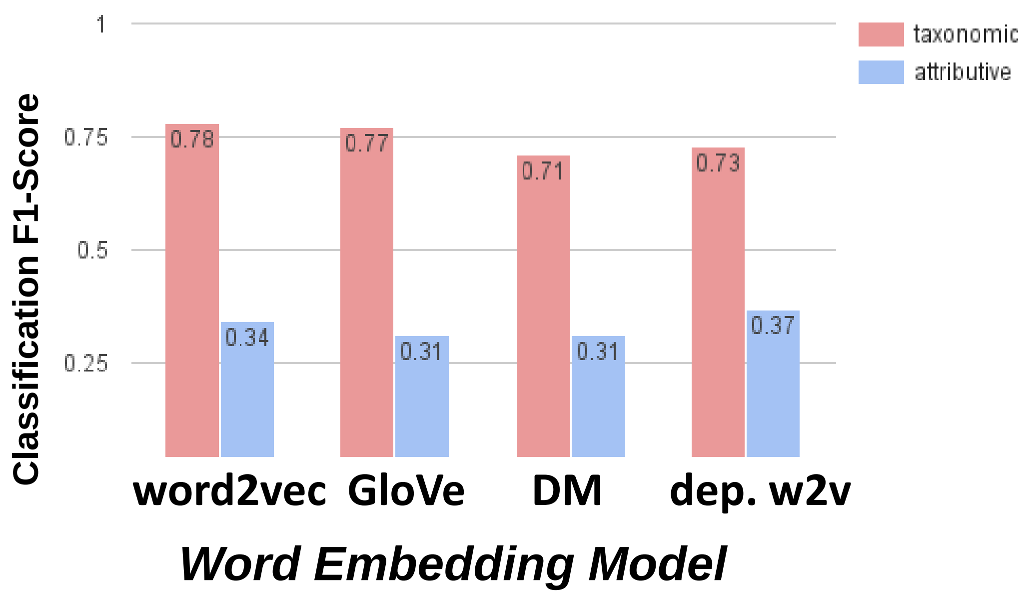
No ***Attributive*** Knowledge

- 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 ρ scores on SimLex999 (Hill et al., 2014):

Model	Nouns	Verbs
GloVe (Pennington et al., 2014)	0.377	**0.163**
word2vec skip-gram (Mikolov et al., 2013)	0.501	**0.307**

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
  - Entertainment: *stars-in-film* (Etzioni et al., Artificial Intelligence 2005)
  - 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, *Chklovski and Pantel, 2004*)
- Patterns can also represent a wide range of semantic relations from different domains
  - Entertainment: *stars-in-film* (Etzioni et al., Artificial Intelligence 2005)
  - 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	X is a Y
beds and sofas	Rihanna is a singer
sofas and beds	*singer is a Rihanna

Symmetric Patterns

X and Y

beds and sofas

sofas and beds

- 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 ***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 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

Type	Example	#instances	
		Bag-of-words	Symmetric Patterns
similar	(car,**train**)	2418	145
	(coffee,**tea**)	6324	1857
	(dog,**cat**)	3645	2090

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

Type	Example	#instances	
		Bag-of-words	Symmetric Patterns
similar	(car,**train**)	2418	145
	(coffee,**tea**)	6324	1857
	(dog,**cat**)	3645	2090
related	(car,**wheel**)	333	3
	(coffee,**cup**)	7247	6
	(dog,**walking**)	2837	4

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

Type	Example	#instances	
		Bag-of-words	Symmetric Patterns
similar	(car,**train**)	2418	145
	(coffee,**tea**)	6324	1857
	(dog,**cat**)	3645	2090
related	(car,**wheel**)	333	3
	(coffee,**cup**)	7247	6
	(dog,**walking**)	2837	4

Symmetric Patterns as Word Embeddings Contexts

Schwartz, Reichart and Rappoport, CoNLL 2015

$$V_{dog} = \begin{pmatrix} \cdot \\ \cdot \\ \cdot \\ \text{count}(\textit{dog}, w_i) \\ \cdot \\ \cdot \\ \cdot \end{pmatrix}$$

Symmetric Patterns as Word Embeddings Contexts

Schwartz, Reichart and Rappoport, CoNLL 2015

$$V_{dog} = \begin{pmatrix} \vdots \\ \text{count}(dog, w_i) \\ \vdots \end{pmatrix}$$

$$V_{dog}^{SP} = \begin{pmatrix} \vdots \\ \textbf{symmatric-pattern\_count}(dog, \text{w}_\text{i}) \\ \vdots \end{pmatrix}$$

Symmetric Patterns as Word Embeddings Contexts

Schwartz, Reichart and Rappoport, CoNLL 2015

$$V_{dog} = \begin{pmatrix} \cdot \\ \cdot \\ \cdot \\ \text{count}(\mathit{dog}, w_i) \\ \cdot \\ \cdot \\ \cdot \end{pmatrix}$$

$$V_{dog}^{SP} = \begin{pmatrix} \cdot \\ \cdot \\ \cdot \\ \textbf{symmatric-pattern\_count}(\mathit{dog}, w_i) \\ \cdot \\ \cdot \\ \cdot \end{pmatrix}$$

The goal:

Distinguish between ***similarity*** and ***relatedness***

Similar Contexts

dog,**cat**

$$V_{dog} = \begin{pmatrix} \cdot \\ \cdot \\ \cdot \\ \text{count}(\textit{dog},\textbf{cat}) \\ \cdot \\ \cdot \\ \cdot \end{pmatrix} \qquad V_{dog}^{SP} = \begin{pmatrix} \cdot \\ \cdot \\ \cdot \\ \textbf{symmatric-pattern}\text{\_count}(\textit{dog},\textbf{cat}) \\ \cdot \\ \cdot \\ \cdot \end{pmatrix}$$

positive

small/zero

Similar Contexts

dog,**cat**

$$V_{dog} = \begin{pmatrix} \vdots \\ \boxed{\text{count}(\textit{dog},\textbf{cat})} \\ \vdots \end{pmatrix} \qquad V_{dog}^{SP} = \begin{pmatrix} \vdots \\ \boxed{\textbf{symmatric-pattern}\text{\_count}(\textit{dog},\textbf{cat})} \\ \vdots \end{pmatrix}$$

positive

small/zero

Related Contexts

dog,**leash**

$$V_{dog} = \begin{pmatrix} \vdots \\ \text{count}(\textit{dog},\textbf{leash}) \\ \vdots \end{pmatrix} \qquad V_{dog}^{SP} = \begin{pmatrix} \vdots \\ \textbf{symmatric-pattern}\text{\_count}(\textit{dog},\textbf{leash}) \\ \vdots \end{pmatrix}$$

positive

small/zero

Related Contexts

dog,**leash**

$$V_{dog} = \begin{pmatrix} \vdots \\ \boxed{\text{count}(\textit{dog},\textbf{leash})} \\ \vdots \end{pmatrix} \qquad V_{dog}^{SP} = \begin{pmatrix} \vdots \\ \boxed{\textbf{symmatric-pattern}\text{\_count}(\textit{dog},\textbf{leash})} \\ \vdots \end{pmatrix}$$

positive

small/zero

Related Contexts

dog,**leash**

$$V_{dog} = \begin{pmatrix} \vdots \\ \boxed{\text{count}(dog,\textbf{leash})} \\ \vdots \end{pmatrix} \qquad V_{dog}^{SP} = \begin{pmatrix} \vdots \\ \boxed{\textbf{symmatric-pattern}\text{\_count}(dog,\textbf{leash})} \\ \vdots \end{pmatrix}$$

Symmetric-pattern embeddings distinguish between ***similarity*** and ***relatedness***

positive

small/zero

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**

Type	Example	#instances		
		Bag-of-words	Symmetric Patterns	Antonym Patterns
related	(bad,**dream**)	1208	0	0
similar	(bad,**evil**)	561	114	0
opposite	(bad,**good**)	23532	806	80

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)

$$V_w^{\mathrm{SP}^+} = V_w^{\mathrm{SP}} - \beta \cdot V_w^{\mathrm{AP}}$$

❖ β is tuned using a development set

Values for ***Related*** Contexts are **small**

bad, **dream**

$$V_{bad}^{SP^+} = \begin{pmatrix} \cdot \\ \cdot \\ \cdot \\ \textbf{symmatric-pattern}\text{\_count}(\textit{bad},\textbf{dream}) \\ \cdot \\ \cdot \\ \cdot \end{pmatrix} - \beta \cdot \begin{pmatrix} \cdot \\ \cdot \\ \cdot \\ \underline{\textbf{antonym}}\textbf{-pattern}\text{\_count}(\textit{bad},\textbf{dream}) \\ \cdot \\ \cdot \\ \cdot \end{pmatrix}$$

positive

small/zero

Values for ***Related*** Contexts are **small**

bad, **dream**

$$V_{bad}^{SP^+} = \begin{pmatrix} \vdots \\ \boxed{\textbf{symmatric-pattern}\_\text{count}(\textit{bad},\textbf{dream})} \\ \vdots \end{pmatrix} - \beta \cdot \begin{pmatrix} \vdots \\ \boxed{\underline{\textbf{antonym}}\textbf{-pattern}\_\text{count}(\textit{bad},\textbf{dream})} \\ \vdots \end{pmatrix}$$

positive

small/zero

Values for ***Similar*** Contexts are **large**

bad, **evil**

$$V_{bad}^{SP^+} = \begin{pmatrix} \vdots \\ \textbf{symmatric-pattern}\text{\_count}(\textit{bad},\textbf{evil}) \\ \vdots \end{pmatrix} - \beta \cdot \begin{pmatrix} \vdots \\ \underline{\textbf{antonym}}\textbf{-pattern}\text{\_count}(\textit{bad},\textbf{evil}) \\ \vdots \end{pmatrix}$$

positive

small/zero

Values for *Similar* Contexts are **large**

bad, **evil**

$$V_{bad}^{SP^{+}} = \begin{pmatrix} \vdots \\ \boxed{\textbf{symmatric-pattern}\_\text{count}(\textit{bad},\textbf{evil})} \\ \vdots \end{pmatrix} - \beta \cdot \begin{pmatrix} \vdots \\ \boxed{\underline{\textbf{antonym}}\textbf{-pattern}\_\text{count}(\textit{bad},\textbf{evil})} \\ \vdots \end{pmatrix}$$

positive

small/zero

Values for ***Opposite*** Contexts are **small**

bad, **good**

$$V_{bad}^{SP^+} = \begin{pmatrix} \vdots \\ \textbf{symmatric-pattern}\_\text{count}(\textit{bad},\textbf{good}) \\ \vdots \end{pmatrix} - \beta \cdot \begin{pmatrix} \vdots \\ \underline{\textbf{antonym}}\textbf{-pattern}\_\text{count}(\textit{bad},\textbf{good}) \\ \vdots \end{pmatrix}$$

positive

small/zero

Values for *Opposite* Contexts are **small**

bad, **good**

$$V_{bad}^{SP^+} = \begin{pmatrix} \vdots \\ \boxed{\textbf{symmatric-pattern}\_\text{count}(\textit{bad},\textbf{good})} \\ \vdots \end{pmatrix} - \beta \cdot \begin{pmatrix} \vdots \\ \boxed{\underline{\textbf{antonym}}\textbf{-pattern}\_\text{count}(\textit{bad},\textbf{good})} \\ \vdots \end{pmatrix}$$

positive

small/zero

Values for ***Opposite*** Contexts are **small**

bad, **good**

$$V_{bad}^{SP^+} = \begin{pmatrix} \vdots \\ \boxed{\textbf{symmatric-pattern}\text{\_count}(\textit{bad},\textbf{good})} \\ \vdots \end{pmatrix} - \beta \cdot \begin{pmatrix} \vdots \\ \boxed{\underline{\textbf{antonym}}\textbf{-pattern}\text{\_count}(\textit{bad},\textbf{good})} \\ \vdots \end{pmatrix}$$

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 **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 **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

Model	Spearman's ρ
GloVe (Pennington et al., 2014)	0.35
PPMI-Bag-of-words	0.423
word2vec CBOW (Mikolov et al,. 2013)	0.43
word2vec Dep (Levy and Goldberg, 2014)	0.436
NNSE (Murphy et al., 2012)	0.455
word2vec skip-gram (Mikolov et al., 2013)	0.462
SP$^+$ (Schwartz et al., 2015)	**0.517**

Results

SimLex999 Dataset

Model	Spearman's ρ
GloVe (Pennington et al., 2014)	0.35
PPMI-Bag-of-words	0.423
word2vec CBOW (Mikolov et al,. 2013)	0.43
word2vec Dep (Levy and Goldberg, 2014)	0.436
NNSE (Murphy et al., 2012)	0.455
word2vec skip-gram (Mikolov et al., 2013)	0.462
SP$^+$ (Schwartz et al., 2015)	**0.517**

5.5%

Results

SimLex999 Dataset

Model	Spearman's ρ
GloVe (Pennington et al., 2014)	0.35
PPMI-Bag-of-words	0.423
word2vec CBOW (Mikolov et al,. 2013)	0.43
word2vec Dep (Levy and Goldberg, 2014)	0.436
NNSE (Murphy et al., 2012)	0.455
word2vec skip-gram (Mikolov et al., 2013)	0.462
SP⁺ (Schwartz et al., 2015)	**0.517**

5.5%

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

Results

SimLex999 Dataset

Model	Spearman's ρ
GloVe (Pennington et al., 2014)	0.35
PPMI-Bag-of-words	0.423
word2vec CBOW (Mikolov et al,. 2013)	0.43
word2vec Dep (Levy and Goldberg, 2014)	0.436
NNSE (Murphy et al., 2012)	0.455
word2vec skip-gram (Mikolov et al., 2013)	0.462
$\mathbf{SP^+}$ (Schwartz et al., 2015)	**0.517**
Joint	**0.563**

10.1%

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

Part-of-Speech Analysis

Spearman's ρ on the SimLex999 Dataset

Model	Adjective
GloVe (Pennington et al., 2014)	0.571
PPMI-Bag-of-words	0.548
word2vec CBOW (Mikolov et al,. 2013)	0.579
word2vec Dep (Levy and Goldberg, 2014)	0.54
NNSE (Murphy et al., 2012)	0.594
word2vec skip-gram (Mikolov et al., 2013)	0.604
SP^+ (Schwartz et al., 2015)	**0.663**

Part-of-Speech Analysis

Spearman's ρ on the SimLex999 Dataset

Model	Adjective	Nouns
GloVe (Pennington et al., 2014)	0.571	0.377
PPMI-Bag-of-words	0.548	0.451
word2vec CBOW (Mikolov et al,. 2013)	0.579	0.48
word2vec Dep (Levy and Goldberg, 2014)	0.54	0.449
NNSE (Murphy et al., 2012)	0.594	0.487
word2vec skip-gram (Mikolov et al., 2013)	0.604	**0.501**
SP^+ (Schwartz et al., 2015)	**0.663**	0.497

Part-of-Speech Analysis

Spearman's ρ on the SimLex999 Dataset

Model	Adjective	Nouns	Verbs
GloVe (Pennington et al., 2014)	0.571	0.377	0.163
PPMI-Bag-of-words	0.548	0.451	0.276
word2vec CBOW (Mikolov et al,. 2013)	0.579	0.48	0.252
word2vec Dep (Levy and Goldberg, 2014)	0.54	0.449	0.376
NNSE (Murphy et al., 2012)	0.594	0.487	0.318
word2vec skip-gram (Mikolov et al., 2013)	0.604	**0.501**	0.307
SP^+ (Schwartz et al., 2015)	**0.663**	0.497	**0.578**

Part-of-Speech Analysis

Spearman's ρ on the SimLex999 Dataset

Model	Adjective	Nouns	Verbs
GloVe (Pennington et al., 2014)	0.571	0.377	0.163
PPMI-Bag-of-words	0.548	0.451	0.276
word2vec CBOW (Mikolov et al,. 2013)	0.579	0.48	0.252
word2vec Dep (Levy and Goldberg, 2014)	0.54	0.449	0.376
NNSE (Murphy et al., 2012)	0.594	0.487	0.318
word2vec skip-gram (Mikolov et al., 2013)	0.604	**0.501**	0.307
SP^+ (Schwartz et al., 2015)	**0.663**	0.497	**0.578**

Part-of-Speech Analysis

Spearman's ρ on the SimLex999 Dataset

Model	Adjective	Nouns	Verbs
GloVe (Pennington et al., 2014)	0.571	0.377	0.163
PPMI-Bag-of-words	0.548	0.451	0.276
word2vec CBOW (Mikolov et al,. 2013)	0.579	0.48	0.252
word2vec Dep (Levy and Goldberg, 2014)	0.54	0.449	0.376
NNSE (Murphy et al., 2012)	0.594	0.487	0.318
word2vec skip-gram (Mikolov et al., 2013)	0.604	**0.501**	$\boxed{0.307}$
SP$^+$ (Schwartz et al., 2015)	**0.663**	0.497	**0.578**

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

Model	Nouns	Verbs
word2vec skip-gram (Mikolov et al., 2013)	**0.501**	0.307
SP^{+} (Schwartz et al., 2015)	0.497	**0.578**

- 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

Model	Context Type	Spearman's ρ
skip-gram	Bag-of-Words	0.307
	Dependency Links	0.386
	Symmetric Patterns	**0.459**

Compact Model

Model	Context Type	Verbs	#Contexts	Train Time (Mins)
skip-gram	Bag-of-Words	0.307	13000M	320
	Dependency Links	0.386	14500M	551
	Symmetric Patterns	**0.459**	**270M**	**11**

Additive Value of *Symmetric Patterns* and *Negative Weighting*

Model	Context Type	Verbs
skip-gram	Bag-of-Words	0.307
	Dependency Links	0.386
	Symmetric Patterns	**0.459**
SP+ (**Schwartz** et al., 2015)	Symmetric Patterns	0.578

Additive Value of *Symmetric Patterns* and *Negative Weighting*

Model	Context Type	Verbs
skip-gram	Bag-of-Words	0.307
	Dependency Links	0.386
	Symmetric Patterns	**0.459**
SP^{+} (**Schwartz** et al., 2015)	Symmetric Patterns	0.578
SP^{-NW} (**Schwartz** et al., 2015)	Symmetric Patterns	0.441

Additive Value of *Symmetric Patterns* and *Negative Weighting*

Model	Context Type	Verbs
skip-gram	Bag-of-Words	0.307
	Dependency Links	0.386
	Symmetric Patterns	**0.459**
SP^{+} (**Schwartz** et al., 2015)	Symmetric Patterns	0.578
SP^{-NW} (**Schwartz** et al., 2015)	Symmetric Patterns	0.441

0

Additive Value of *Symmetric Patterns* and *Negative Weighting*

Model	Context Type	Verbs	
skip-gram	Bag-of-Words	0.307	0
	Dependency Links	0.386	
	Symmetric Patterns	**0.459**	+~15%
SP^{+} (**Schwartz** et al., 2015)	Symmetric Patterns	0.578	
SP^{-NW} (**Schwartz** et al., 2015)	Symmetric Patterns	0.441	+~15%

Additive Value of *Symmetric Patterns* and *Negative Weighting*

Model	Context Type	Verbs	
skip-gram	Bag-of-Words	0.307	0
	Dependency Links	0.386	
	Symmetric Patterns	**0.459**	+~15%
SP^+ (**Schwartz** et al., 2015)	Symmetric Patterns	0.578	+~27%
SP^{-NW} (**Schwartz** et al., 2015)	Symmetric Patterns	0.441	+~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!***

Surprise

John and Mary are friends. They hang out together. Last night ***John moved out of town without telling Mary***

Surprise – why?

- *surprising ≈ 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*