

# Improved Word Representations via Summed Target and Context Embeddings

Nancy Fulda, Nathaniel Robinson

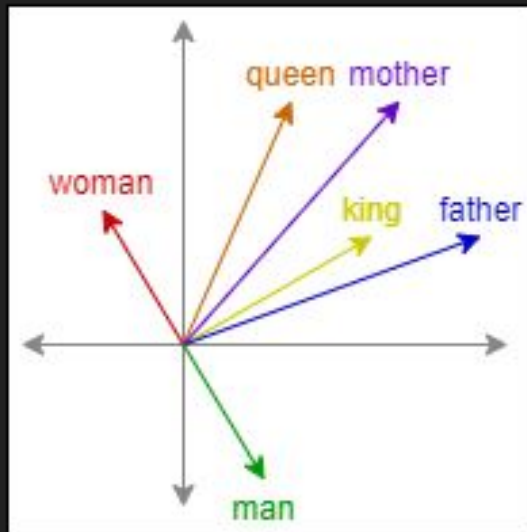


# Word Embedding Vectors

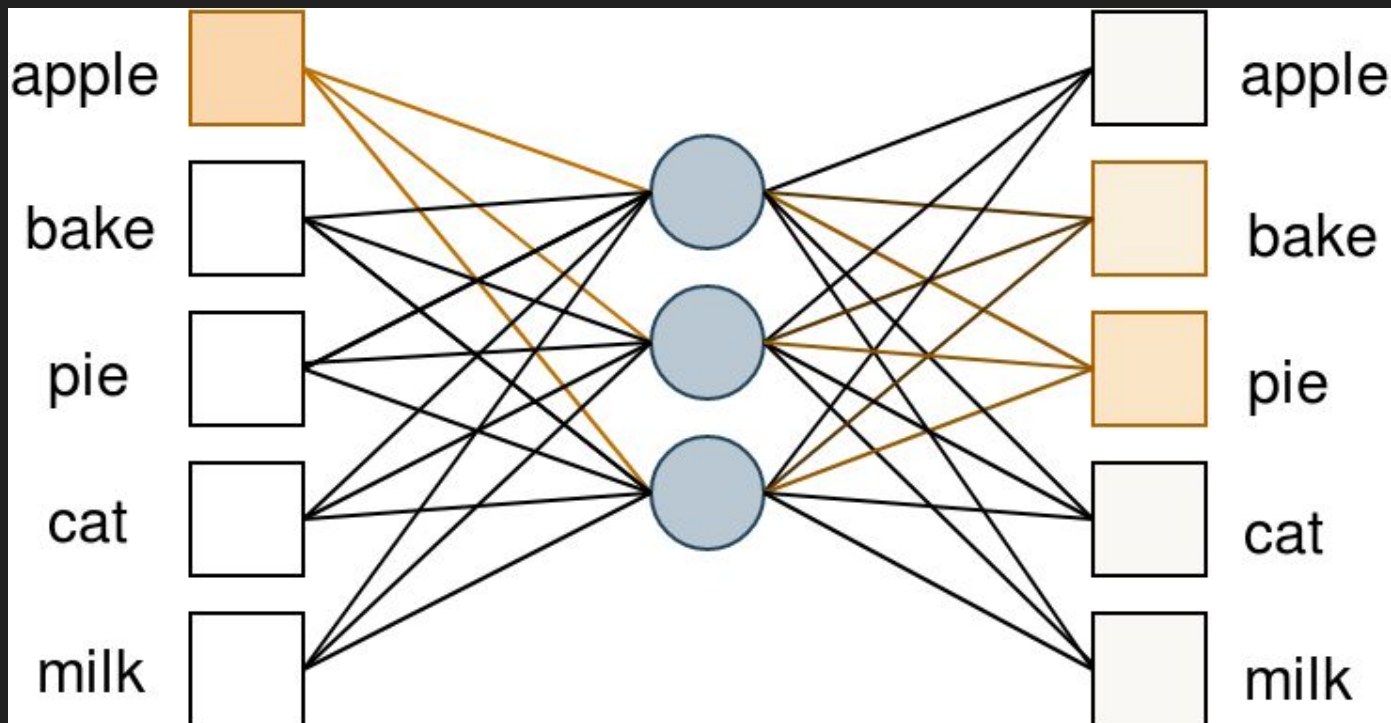
- Word2vec: original skip-gram<sup>[1]</sup>
- GloVe: log bi-linear regression<sup>[2]</sup>
- FastText: subword information<sup>[3]</sup>
- BERT: contextualized embeddings<sup>[4]</sup>
- Sentence embeddings

# Word Embedding Vectors

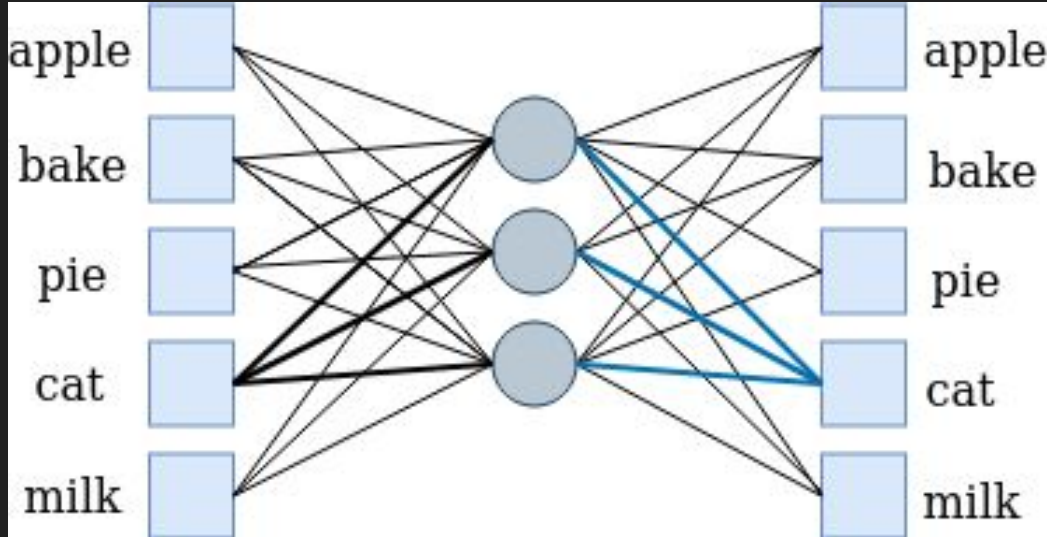
- Word2vec: original skip-gram<sup>[1]</sup>
- GloVe: log bi-linear regression<sup>[2]</sup>
- FastText: subword information<sup>[3]</sup>
- BERT: contextualized embeddings<sup>[4]</sup>
- Sentence embeddings



# Word2vec Skip-gram Model



# Word2vec-PLUS: sum target and context weights



$$\text{vec}(i) = W_{in}[i] + W_{out}^T[i]$$

# Word2vec-PLUS Training Corpora

Corpus	Size	Token count
Scraped articles	59.0 GB	9.6B
Wikipedia text	16.7 GB	2.8B
Toronto Book Corpus	4.6 GB	984M
Gutenberg classic books	1.2 GB	82M
Classic books	20.3 MB	

# How do we evaluate embedding quality?

Analogy sets:

- Google Analogy Test Set<sup>[5]</sup>



- SAT Questions<sup>[6]</sup>



- SemEval 2013<sup>[7]</sup>



# Google Analogy Test Set

- Common evaluation for word embeddings
  - GloVe
  - FastText
  - Word2vec
- 19,544 questions in 14 categories
  - *Athens : Greece :: Baghdad : Iraq* (capital-country)
  - *boy : girl :: brother : sister* (family relationships)
  - *acceptable : unacceptable :: aware : unaware* (opposites)
  - *bad : worse :: big : bigger* (comparatives)

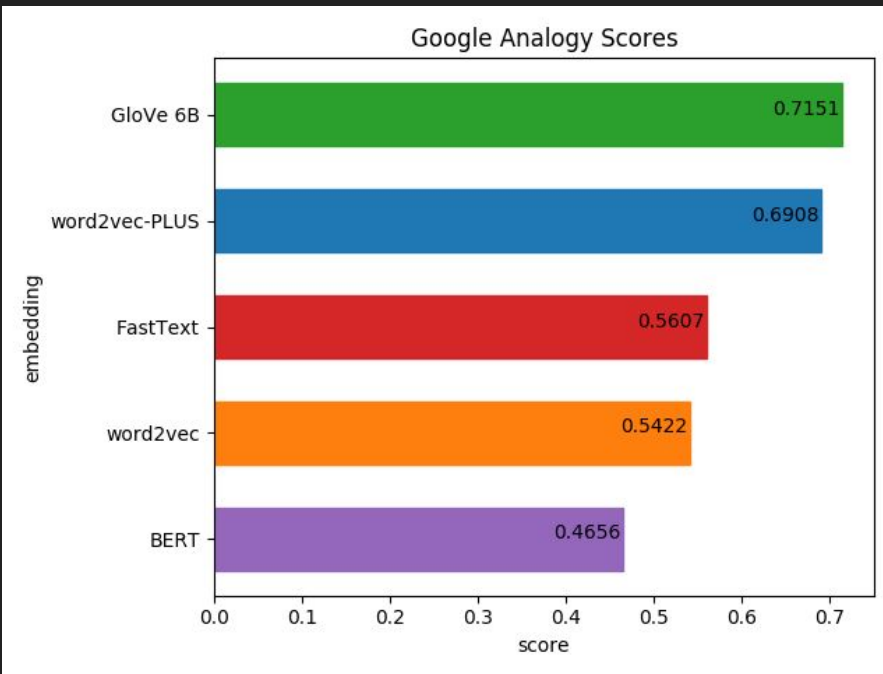
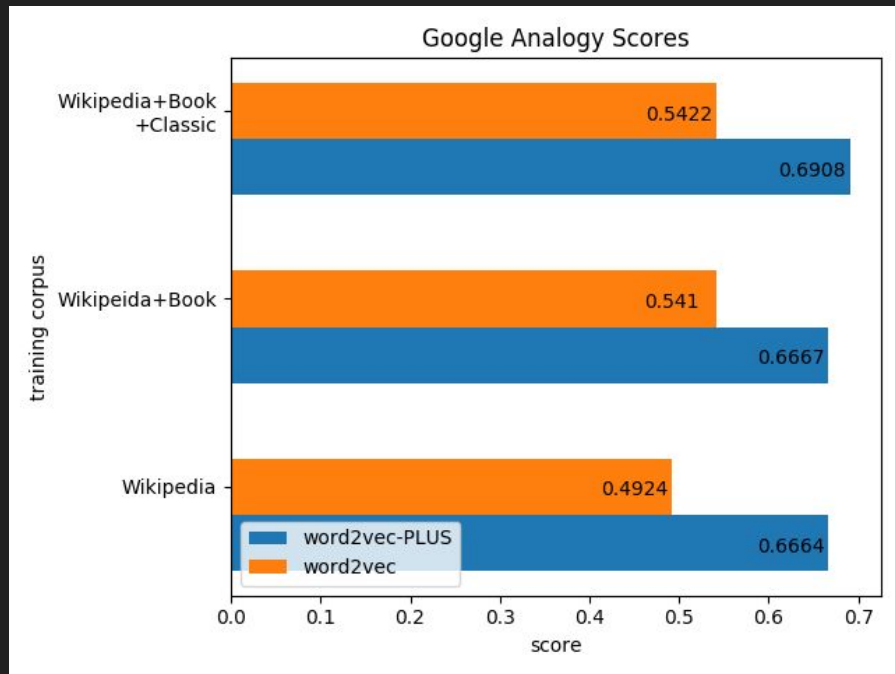


# Google Analogy Test Set



- Cognitively simple questions
  - ... but no multiple choice
- Method:
  - Given  $a : b :: c : d$ , compute  $d' = c + b - a$  and find  $\operatorname{argmin}_{s \in \text{vocab}} \operatorname{dist}_{\cos}(d', s)$
  - Equivalent to  $\operatorname{argmax}_{s \in \text{vocab}} \operatorname{sim}_{\cos}(d', s)$

# Google Analogy Test Set

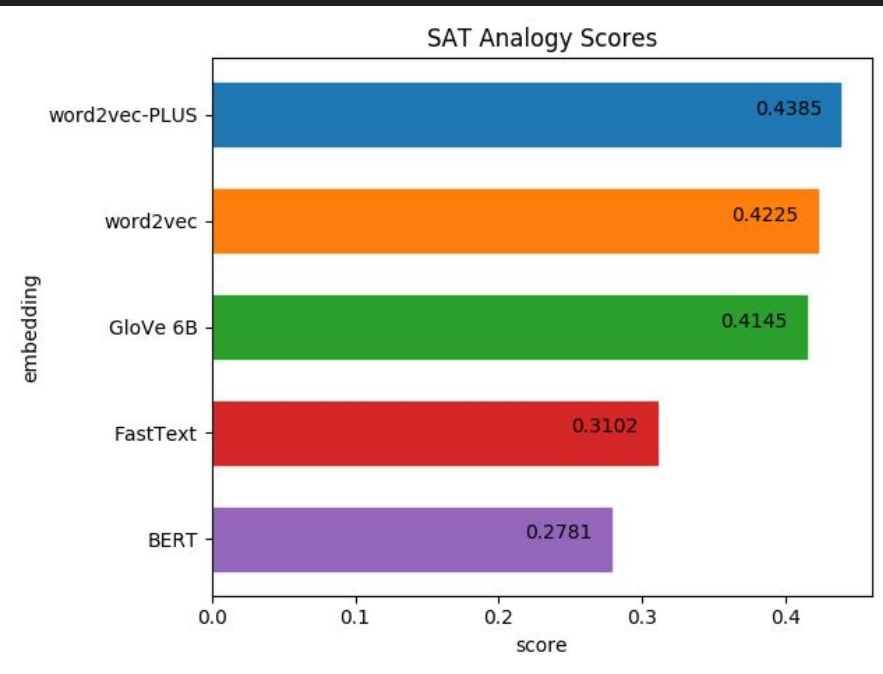
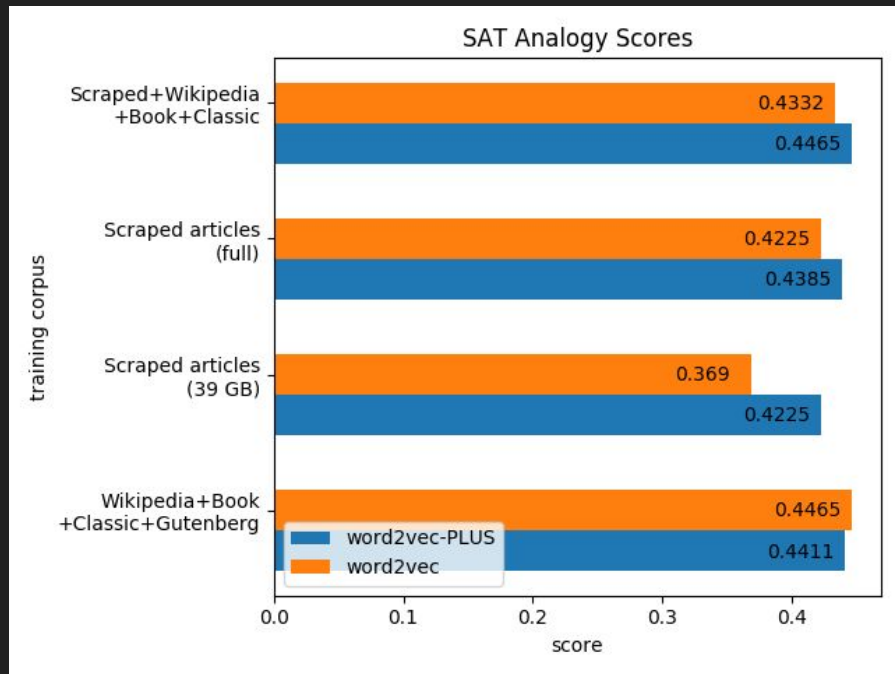


# SAT Questions

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

Compute  $\operatorname{argmin}_{s \in [A, B, C, D, E]} \operatorname{dist}_{\cos}(d', s)$

# SAT Questions

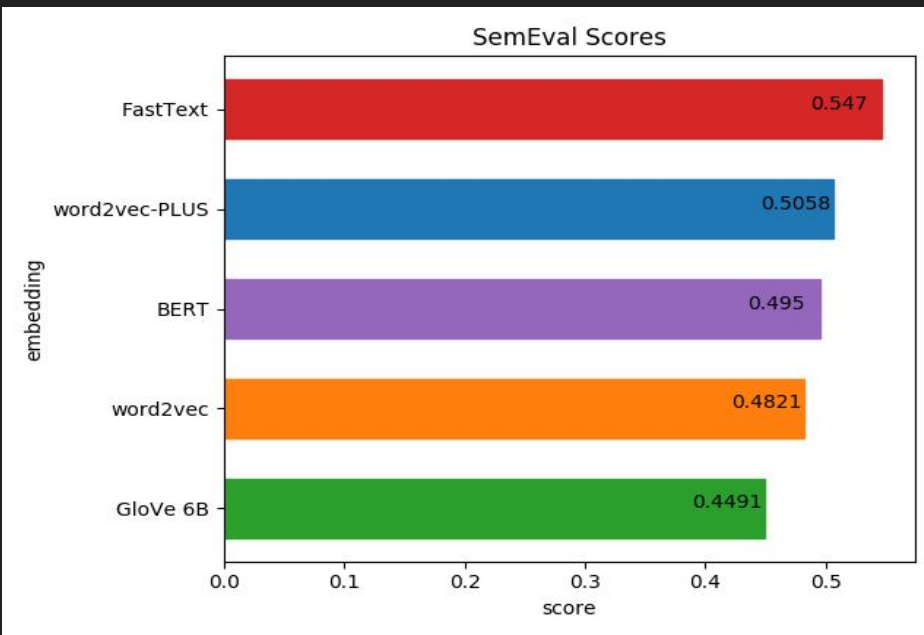
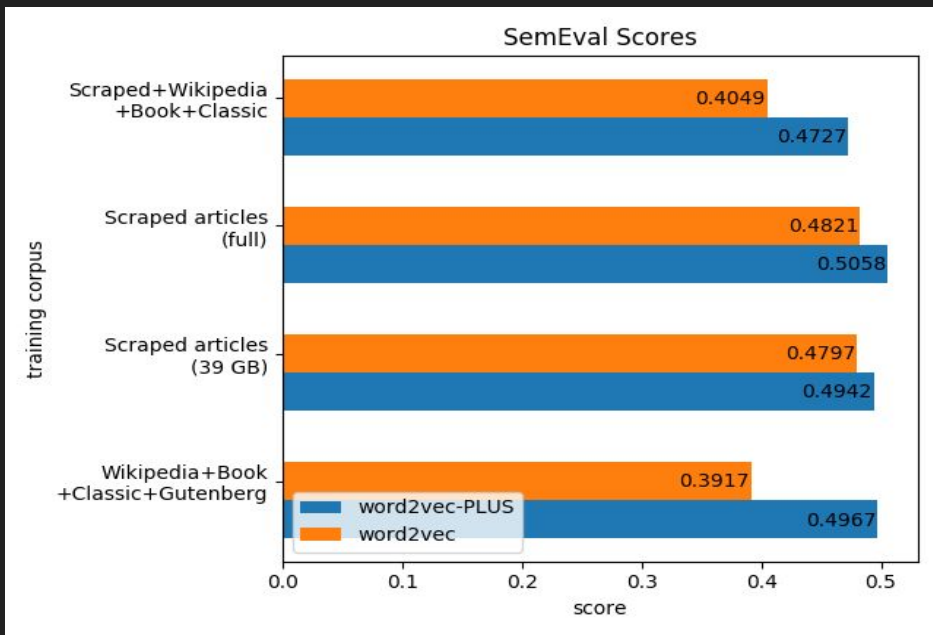


# SemEval 2013 Sentence Similarity



- Sentence pairs with human-given similarity scores
  - E.g. “A woman is cooking.” / “A woman is cooking something.” / score = 3
- Method:
  - Sentence vector  $\leftarrow$  sum of word vectors
  - Find cosine similarity of sentence vectors in each pair
  - Final score: correlation of embedding-given similarity scores with human-given scores

# SemEval 2013 Sentence Similarity



# Analysis: advantage of summed embeddings

$$\text{sim}(a, b) \propto a_{IN}^T b_{IN} \quad (2)$$

$$\text{sim}(a, b) \propto a_{OUT}^T b_{OUT} \quad (3)$$

$$\begin{aligned} \text{sim}(a, b) \propto (a_{IN}^T + a_{OUT}^T)(b_{IN} + b_{OUT}) = \\ a_{IN}^T b_{IN} + a_{IN}^T b_{OUT} + a_{OUT}^T b_{IN} + a_{OUT}^T b_{OUT} \end{aligned} \quad (4)$$

$$\text{sim}(a, b) \propto [a_{IN}^T, a_{OUT}^T] \begin{bmatrix} b_{IN} \\ b_{OUT} \end{bmatrix} = a_{IN}^T b_{IN} + a_{OUT}^T b_{OUT} \quad (5)$$

# Topical and typical similarity (Nalisnick et al.<sup>[8]</sup>)

## Topical

Contexts together

*yale: faculty, alumni, orientation*

$$a_{IN}^T b_{OUT}$$

$$a_{OUT}^T b_{IN}$$

## Typical

Similar contexts

*yale: harvard, nyu, cornell*

$$a_{IN}^T b_{IN}$$

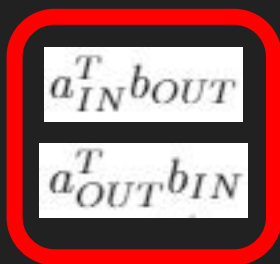


# Topical and typical similarity (Nalisnick et al.<sup>[8]</sup>)

## Topical

Contexts together

*yale: faculty, alumni, orientation*


$$\begin{array}{l} a_{IN}^T b_{OUT} \\ a_{OUT}^T b_{IN} \end{array}$$

Objective of network

## Typical

Similar contexts

*yale: harvard, nyu, cornell*

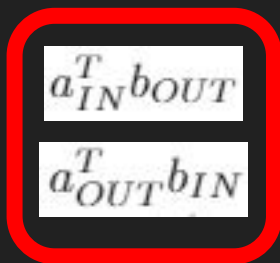
$$a_{IN}^T b_{IN}$$

# Topical and typical similarity (Nalisnick et al.<sup>[8]</sup>)

## Topical

Contexts together

*yale: faculty, alumni, orientation*

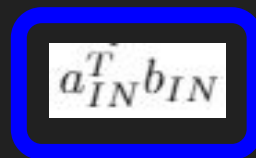

$$\begin{array}{l} a_{IN}^T b_{OUT} \\ a_{OUT}^T b_{IN} \end{array}$$

Objective of network

## Typical

Similar contexts

*yale: harvard, nyu, cornell*


$$a_{IN}^T b_{IN}$$

Approach the same context  
word embeddings

# Analysis: advantage of summed embeddings

$$sim(a, b) \propto a_{IN}^T b_{IN} \quad (2)$$

$$sim(a, b) \propto a_{OUT}^T b_{OUT} \quad (3)$$

$$\begin{aligned} sim(a, b) &\propto (a_{IN}^T + a_{OUT}^T)(b_{IN} + b_{OUT}) = \\ &\quad a_{IN}^T b_{IN} + a_{IN}^T b_{OUT} + a_{OUT}^T b_{IN} + a_{OUT}^T b_{OUT} \end{aligned} \quad (4)$$

$$sim(a, b) \propto [a_{IN}^T, a_{OUT}^T] \begin{bmatrix} b_{IN} \\ b_{OUT} \end{bmatrix} = a_{IN}^T b_{IN} + a_{OUT}^T b_{OUT} \quad (5)$$

# Analysis: advantage of summed embeddings

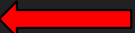
$$sim(a, b) \propto a_{IN}^T b_{IN} \quad (2)$$

$$sim(a, b) \propto a_{OUT}^T b_{OUT} \quad (3)$$

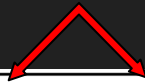
$$sim(a, b) \propto (a_{IN}^T + a_{OUT}^T)(b_{IN} + b_{OUT}) =$$
$$a_{IN}^T b_{IN} + a_{IN}^T b_{OUT} + a_{OUT}^T b_{IN} + a_{OUT}^T b_{OUT} \quad (4)$$

$$sim(a, b) \propto [a_{IN}^T, a_{OUT}^T] \begin{bmatrix} b_{IN} \\ b_{OUT} \end{bmatrix} = a_{IN}^T b_{IN} + a_{OUT}^T b_{OUT} \quad (5)$$


Summed  
embeddings



Solving analogies requires knowledge of both varieties of similarity

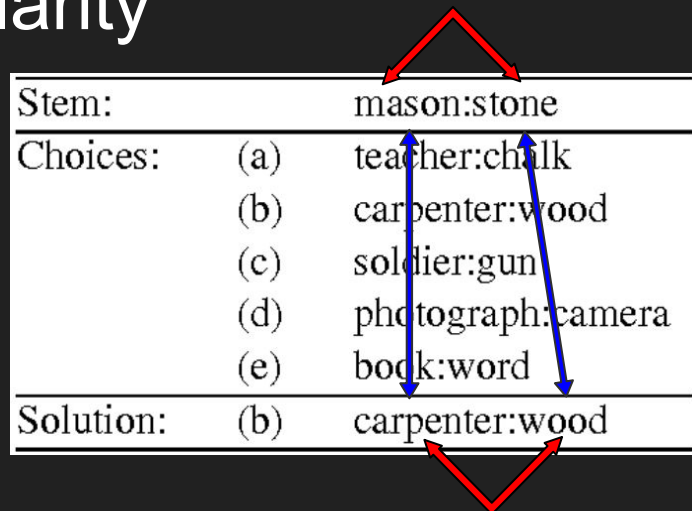


Stem:	mason:stone
Choices:	(a) teacher:chalk
	(b) carpenter:wood
	(c) soldier:gun
	(d) photograph:camera
	(e) book:word
Solution:	(b) carpenter:wood



The diagram illustrates the relationship between the stem and the solution. Red arrows point from the stem 'mason:stone' to the solution '(b) carpenter:wood', indicating the analogy being solved.

# Solving analogies requires knowledge of both varieties of similarity



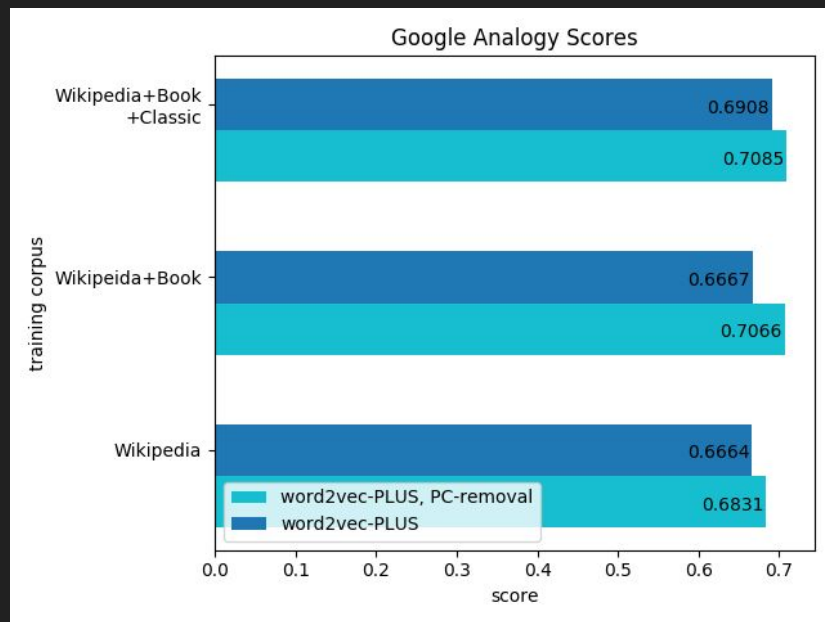
Stem:		mason:stone
Choices:	(a)	teacher:chalk
	(b)	carpenter:wood
	(c)	soldier:gun
	(d)	photograph:camera
	(e)	book:word
Solution:	(b)	carpenter:wood

Recall:  $\operatorname{argmax}_{s \in \text{vocab}} \operatorname{sim}_{\cos}(d', s)$

$$\underbrace{V_{\text{wood}}^T}_{\text{S}} (\underbrace{V_{\text{stone}} + V_{\text{carpenter}} - V_{\text{mason}}}_{d'}) = \underbrace{V_{\text{wood}}^T V_{\text{stone}}}_{\text{blue box}} + \underbrace{V_{\text{wood}}^T V_{\text{carpenter}}}_{\text{red box}} - V_{\text{wood}}^T V_{\text{mason}}$$

# Principal Component Removal (Arora et al.<sup>[9]</sup>)

- $v \leftarrow uu^T v$ 
  - where  $u$  is the first singular vector of the embedding matrix



# Conclusion

- Summing target and context vectors to produce embeddings in a word2vec skip-gram model yields advantages in some analogy tasks
- Benefits in other NLP tasks and other embedding algorithms is an area of future research
- Principle component removal may be a viable method to improve embedding quality in some applications



# Thank you

[1] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

[2] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, pages 1532–1543, 2014.

[3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5:135–146, 2017.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[5] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. Association for Computational Linguistics, 2013.

[6] Peter D Turney. Similarity of semantic relations. Computational Linguistics, 32(3):379–416, 2006.

[7] Theresa Wilson, Zornitsa Kozareva, Preslav Nakov, Alan Ritter, Sara Rosenthal, and Veselin Stoyanov. Sentiment analysis in twitter. <http://www.cs.york.ac.uk/semeval-2013/task2/>, 2013.

[8] Eric Nalisnick, Mitra Bhaskar, Nick Craswell, and Rich Caruana. Improving document ranking with dual word embeddings. In WWW '16 Companion: Proceedings of the 25th International Conference Companion on World Wide Web, pages 83–84, 2016.

[9] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. International Conference on Learning Representations, 2017.

Google icon:

[https://www.pikpng.com/pngvi/ihjJRim\\_leave-a-google-review-google-clipart/](https://www.pikpng.com/pngvi/ihjJRim_leave-a-google-review-google-clipart/)

SAT College Board Acorn icon:

[https://www.pinclipart.com/pindetail/Tbhmio\\_teen-center-sat-600600-college-board-acorn/](https://www.pinclipart.com/pindetail/Tbhmio_teen-center-sat-600600-college-board-acorn/)

University of York icon:

<https://www.hiclipart.com/free-transparent-background-png-clipart-nmfih>