What analogies reveal about word vectors and their compositionality

Gregory P. Finley
Stephanie Farmer
Serguei V.S. Pakhomov

The Sixth Joint Conference on Lexical and Computational Semantics

August 3, 2017

Computational approaches to lexical semantics commonly rely on the *distributional hypothesis*: that a word's meaning can be approximated based upon the words occurring near it.

Computational approaches to lexical semantics commonly rely on the *distributional hypothesis*: that a word's meaning can be approximated based upon the words occurring near it.

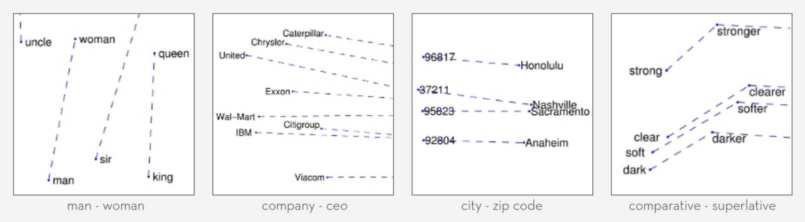
By modeling word meanings with co-occurrence statistics, we unlock linear algebra as a tool for linguistic computation.

- Cosine similarity and human judgments
- Average, add, subtract meaning between words
- Lexical $\rightarrow$ compositional?

Analogy questions

$$dog : puppy :: cat : ?$$

$$w_2 - w_1 \approx w_4 - w_3$$



Pennington *et al.*, https://nlp.stanford.edu/projects/glove/

$$w_2 - w_1 \approx w_4 - w_3$$

$$w_2 - w_1 \approx w_4 - w_3$$

$$w_4 \approx w_3 + w_2 - w_1$$

$$w_2 - w_1 \approx w_4 - w_3$$

$$w_4 \approx w_3 + w_2 - w_1$$

$$dog : puppy :: cat : ?$$

$$\text{cat} + \text{puppy} - \text{dog}$$

$$w_2 - w_1 \approx w_4 - w_3$$

$$w_4 \approx w_3 + w_2 - w_1$$

$$\textit{dog} : \textit{puppy} :: \textit{cat} : \textit{?}$$

$$\begin{array}{lcl} \text{cat} + \text{puppy} - \text{dog} & \approx & \text{kitten} \\ & & \text{puppies} \\ & & \text{pet} \\ & & \text{beagle} \\ & & \quad\vdots \\ & & \text{angiography} \end{array}$$

Prior results

Given the simplicity of the solving method, surprisingly high accuracy for some types of analogy questions.

The most-used test set is probably the Google set (distributed with `word2vec`).

- Rather low diversity of categories—mostly geography and inflection
- Results often reported on "syntactic" and "semantic" subsets; this division is too coarse to be useful

Word similarity

High performance on many categories may be driven by **prior similarity** rather than successfully isolating components of meaning.

Words in any relationship are usually fairly similar.

Word similarity

High performance on many categories may be driven by **prior similarity** rather than successfully isolating components of meaning.

Words in any relationship are usually fairly similar.

$$\textit{horse} : \textit{horses} :: \textit{sailboat} : \textit{sailboats}$$

$$\text{horse} \approx \text{horses},$$
$$\text{sailboat} \approx \text{sailboats}$$

Word similarity

High performance on many categories may be driven by **prior similarity** rather than successfully isolating components of meaning.

Words in any relationship are usually fairly similar.

$$\textit{horse} : \textit{horses} :: \textit{sailboat} : \textit{sailboats}$$

$$\text{horse} \approx \text{horses},$$
$$\text{sailboat} \approx \text{sailboats}$$

$$\text{hypothesis} = \text{sailboat} + \text{horses} - \text{horse} \approx \text{sailboat (!)}$$

Goal

We designed a study that:

1. addresses a wide variety of categories, and
2. controls for prior similarity.

We want to **describe** and **explain** inter-category differences.

Vectors

- `word2vec`
- Wikipedia
- no case or punctuation
- $d = 200$, CBOW

(Also experimented with GloVe, skip-gram, etc.)

Test sets

Test sets

Microsoft Research (Mikolov *et al.*, 2013a): inflectional relationships

- *cheap* : *cheaper* :: *mighty* : *mightier*
- *learn* : *learned* :: *think* : *thought*

  (etc.)

Test sets

Microsoft Research (Mikolov *et al.*, 2013a): inflectional relationships

- *cheap* : *cheaper* :: *mighty* : *mightier*
- *learn* : *learned* :: *think* : *thought*

  (etc.)

Google (`word2vec`; Mikolov *et al.*, 2013b): adds “semantic” categories

- *paris* : *france* :: *havana* : *cuba*
- *austin* : *texas* :: *minneapolis* : *minnesota*
- *king* : *queen* :: *man* : *woman*

  (etc.)

Test sets

Better Analogy Test Set (BATS; Gladkova *et al.*, 2016): more derivational and semantic categories

- *helpful* : *helpfulness* :: *righteous* : *righteousness*
- *bottle* : *glass* :: *clothing* : *fabric*

  (etc.)

Test sets

Better Analogy Test Set (BATS; Gladkova *et al.*, 2016): more derivational and semantic categories

- *helpful* : *helpfulness* :: *righteous* : *righteousness*
- *bottle* : *glass* :: *clothing* : *fabric*

(etc.)

SemEval 2012 (Jurgens *et al.*, 2012): many, many more semantic categories

- *candy* : *sweet* :: *snow* : *cold*
- *boy* : *man* :: *gosling* : *goose*
- *bar* : *drinking* :: *church* : *worship*

(etc.)

Test sets

Source	Categories	Analogies
Microsoft Research	14	7,000
Google (`word2vec`)	14	19,544
BATS	40	95,625
SemEval2012	79	30,082
Total	147	152,251

Table 1: Summary of test data sources.

Metrics: Reciprocal rank

Measure the **rank** of the correct answer in the entire vocabulary, ordered by similarity to hypothesis vector.
(Accuracy only measures if the correct answer is top-ranked.)

Metrics: Reciprocal rank

Measure the **rank** of the correct answer in the entire vocabulary, ordered by similarity to hypothesis vector. (Accuracy only measures if the correct answer is top-ranked.)

Reciprocal of rank (RR) is more sensitive and forgiving than accuracy:

rank	acc	RR
1	1	1
2	0	.5
3	0	.3333
4	0	.25
⋮	⋮	⋮
10526	0	.0001

Metrics: Baseline

hypothesis vector := w_2 or w_3, whichever is better

walk : *walked* :: ***fly*** : *flew*

banana : ***yellow*** :: *cherry* : *red*

- (w_3 is better than w_2 in about 85% of cases)

Metrics: Baseline

hypothesis vector := w_2 or w_3, whichever is better

$walk : walked :: \boldsymbol{fly} : flew$

$banana : \boldsymbol{yellow} :: cherry : red$

- (w_3 is better than w_2 in about 85% of cases)

For each category of analogy questions, measure:

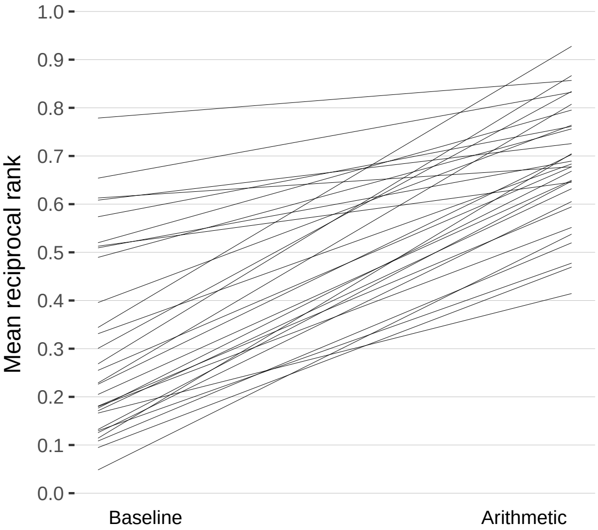
- mean RR using vector arithmetic hypothesis,
- mean RR of the baseline hypothesis,
- the difference between them.

Analogy supercategories

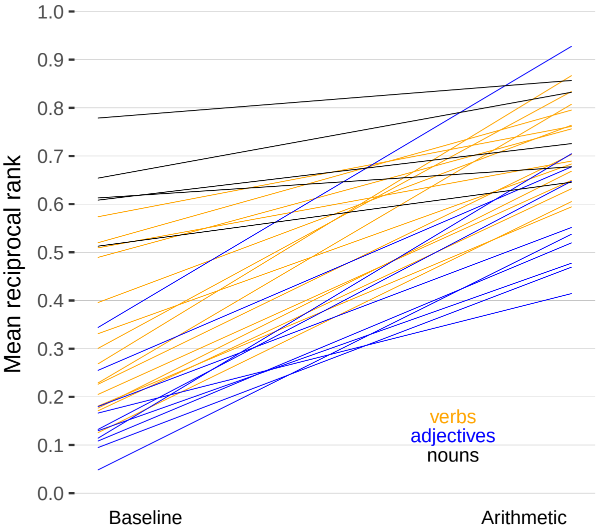
We have 147 distinct categories of analogical relationships.
For visualization and analysis, consider supercategories:

- **inflection:** inflectional morphological relationships (noun plural, adjective degree, verb tense)
- **derivation:** derivational morphology (*-tion*, *un-*)
- **named entity semantics:** meanings of words with a single real-world referent (*Vancouver, Beethoven*)
- **lexical semantics:** meanings of common nouns, adjectives, verbs, etc.

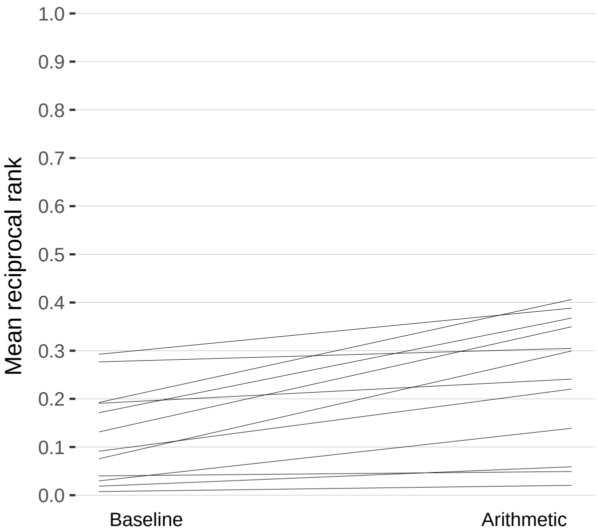
Results: Inflection



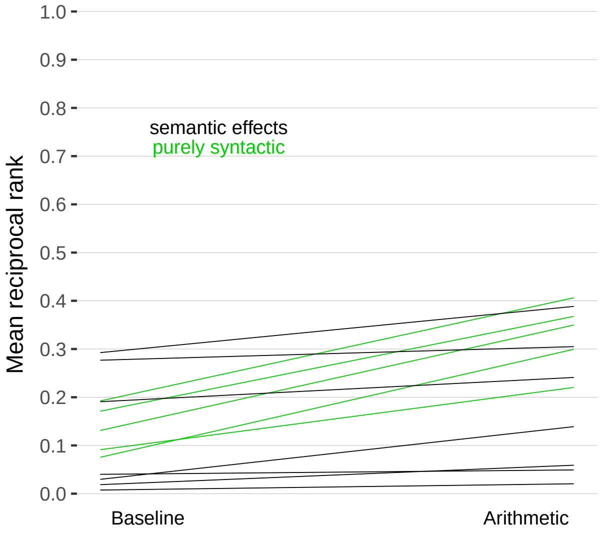
Results: Inflection



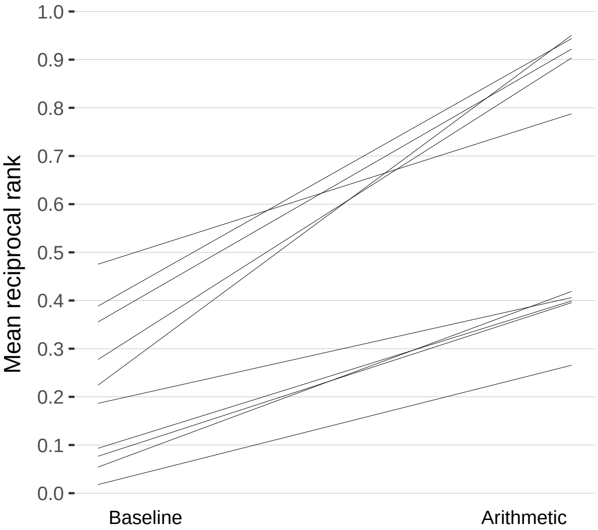
Results: Derivation



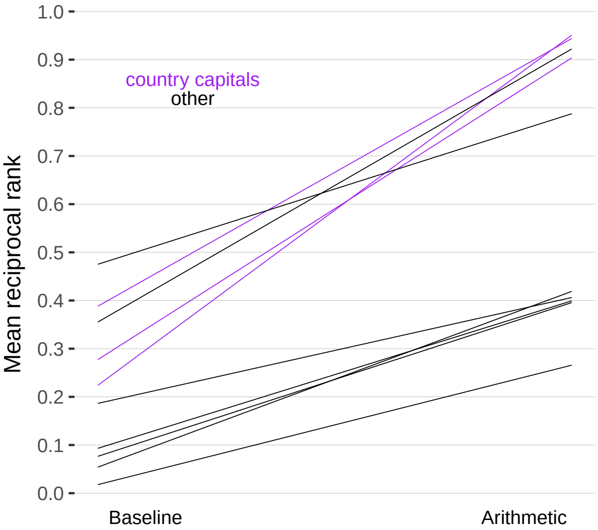
Results: Derivation



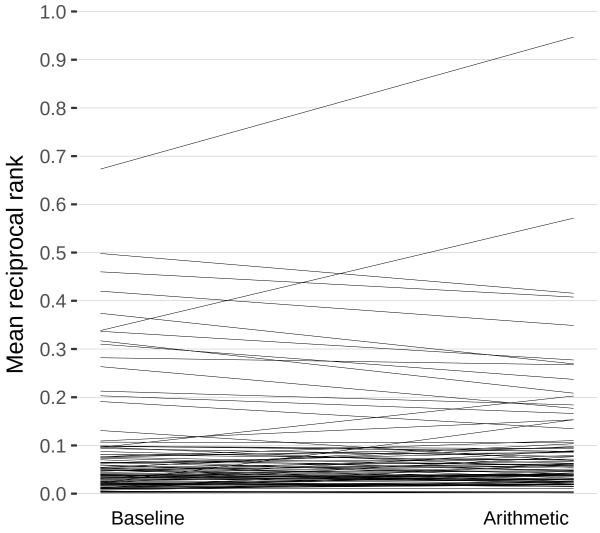
Results: Named entities



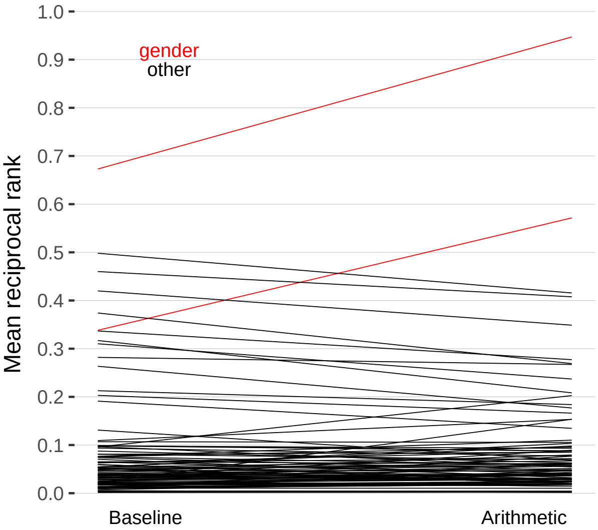
Results: Named entities



Results: Lexical



Results: Lexical



Discussion: Derivation

Linguists have proposed an inflection–derivation continuum.
Is performance worse with “more derivational” affixes?

Discussion: Derivation

Linguists have proposed an inflection–derivation continuum.
Is performance worse with "more derivational" affixes?

One hallmark of derivation is changing a word's syntactic class.
But for our results, a better continuum seems to be from morphemes with purely **syntactic** to purely **semantic** effects.

Discussion: Derivation

Linguists have proposed an inflection–derivation continuum.
Is performance worse with "more derivational" affixes?

One hallmark of derivation is changing a word's syntactic class.
But for our results, a better continuum seems to be from morphemes with purely **syntactic** to purely **semantic** effects.

affix	syntactic?	semantic?	RR gain
-ment	V $\rightarrow$ N	minimal	.224
-tion	V $\rightarrow$ N	minimal	.218
-ly	A $\rightarrow$ Adv	minimal	.205
-ness	A $\rightarrow$ N	minimal	.129
-er	V $\rightarrow$ N	some?	.109
un-	no (A)	yes	.062
re-	no (V)	yes	.050
-able	V $\rightarrow$ A	yes	.040
-less	N $\rightarrow$ A	yes	.013
-over	no (V)	yes	.009

Discussion: Named entities

Why the stark difference between named entities and other semantic relationships?

Semantic theory supports differentiating common from named nouns. E.g., in Montagovian semantics:

- proper nouns denote **individuals** (type e)
- common nouns denote **sets of individuals** (one-place predicates of type $\langle e, t \rangle$)

Discussion: Named entities

Why the stark difference between named entities and other semantic relationships?

Semantic theory supports differentiating common from named nouns. E.g., in Montagovian semantics:

- proper nouns denote **individuals** (type e)
- common nouns denote **sets of individuals** (one-place predicates of type $\langle e, t \rangle$)

Polysemy/ambiguity is a known problem for distributional approaches. If every **referent** is a sense, common nouns are extremely polysemous!

Concretely: vector must *simultaneously* model the word co-occurrences for every individual in the set.

A unified account

A relationship can be captured effectively by vector subtraction if it has **predictable distributional consequences**.

A unified account

A relationship can be captured effectively by vector subtraction if it has **predictable distributional consequences**.

Sets of co-occurrence differences between terms in a pair should be **regular** and **small**.

A unified account

A relationship can be captured effectively by vector subtraction if it has **predictable distributional consequences**.

Sets of co-occurrence differences between terms in a pair should be **regular** and **small**.

- Inflection has predictable effects with agreement and syntax. Adjectives especially:

  *that is a **cheap** tuxedo*
  *that is a **cheaper** tuxedo than ...*
  *that is the **cheapest** tuxedo*

A unified account

A relationship can be captured effectively by vector subtraction if it has **predictable distributional consequences**.

Sets of co-occurrence differences between terms in a pair should be **regular** and **small**.

- Inflection has predictable effects with agreement and syntax. Adjectives especially:

  that is a ***cheap*** *tuxedo*
  that is a ***cheaper*** *tuxedo than ...*
  that is the ***cheapest*** *tuxedo*

  ... and verbs too:

  she ***ran*** *out of time*
  she is ***running*** *out of time*
  she has ***run*** *out of time*

A unified account

- Derivation is less regular than inflection. More importantly, its distributional effects are less automatic and predictable:

  Billy was a ***slow*** *runner*
  Billy ran ***slowly***
  vs.
  their investments have been very ***prudent*** *this year*
  they invested very ***prudently*** *this year*

- Adverbs tend to co-occur with verbs and adjectives with nouns, but these words do not belong to **closed classes** as they tend to with inflection.

A unified account

What about semantics?

- Less polysemous nouns will have “tighter” distributions: lower diversity of co-occurrences, thus smaller sets of differences. Named entities are especially non-polysemous.
  - The relationship between every *dog* and every *puppy* is less consistent than the relationship between every *Netherlands* and every *Amsterdam*.

A unified account

What about semantics?

- Less polysemous nouns will have "tighter" distributions: lower diversity of co-occurrences, thus smaller sets of differences. Named entities are especially non-polysemous.
  - The relationship between every *dog* and every *puppy* is less consistent than the relationship between every *Netherlands* and every *Amsterdam*.
- Gendered nouns agree with pronouns—a closed class, as seen with inflectional relationships.

*when the **boy** dropped his ice cream, he cried*
*when the **girl** dropped her ice cream, she cried*

Conclusion

We have arrived at an explanation grounded in linguistic and distributional theory that accounts for the effects observed.

- Should work further to verify the claim that certain distributional differences are more regular (although the analogy task *does* measure that directly).

Recommend: Test a wide variety of questions. Use a baseline. Don't rely on coarse splits like "syntactic/semantic."

All code, results, and figures will be available at:
https://github.com/gpfinley/analogies

This work was partially supported by the National Institute of General Medical Sciences (GM102282).