

word embeddings what, how and whither

Yoav Goldberg
Bar Ilan University

one morning,
as a parsing researcher woke
from an uneasy dream,
he realized that
he somehow became an expert
in distributional lexical semantics.

and that everybody calls them
"distributed word embeddings" now.

how did this happen?

- People were really excited about word embeddings and their magical properties.
- Specifically, we came back from NAACL, where Mikolov presented the vector arithmetic analogies.
- We got excited too.
- And wanted to **understand what's going on.**

the quest for understanding

- Reading the papers? useless. really.
- Fortunately, Tomas Mikolov released word2vec.
- Read the C code. (dense, but short!)
- Reverse engineer the reasoning behind the algorithm.
- Now it all makes sense.
 - Write it up and post a tech-report on arxiv.

math > magic

the revelation

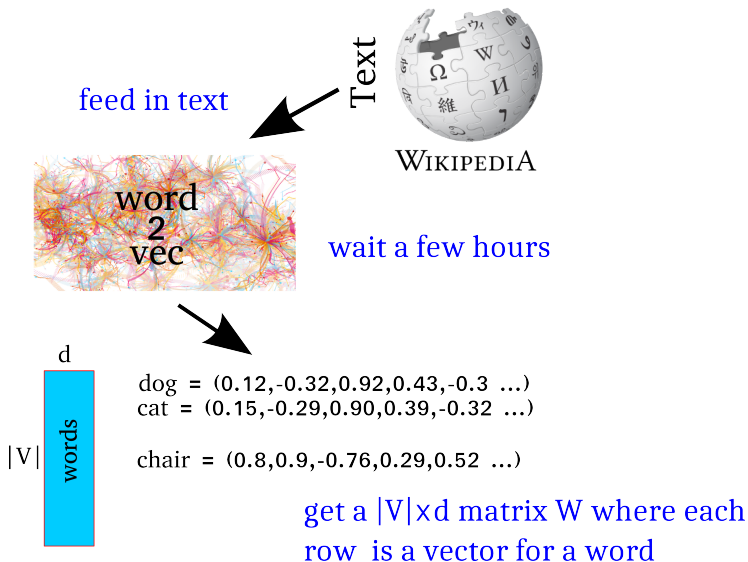
- The math behind word2vec is actually pretty simple.
- Skip-grams with negative sampling are especially easy to analyze.
- Things are really, really similar to what people have been doing in distributional lexical semantics for decades.
 - this is a good thing, as we can re-use a lot of their findings.

this talk

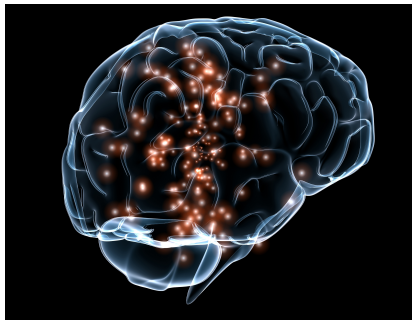
- Understanding word2vec
- Rants:
 - Rants about evaluation.
 - Rants about word vectors in general.
 - Rants about what's left to be done.

understanding
word2vec

word2vec

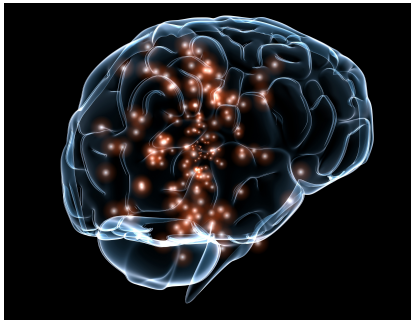


Seems magical.



“Neural computation, just like in the brain!”

Seems magical.



“Neural computation, just like in the brain!”

How does this actually work?

How does word2vec work?

word2vec implements several different algorithms:

Two training methods

- ▶ Negative Sampling
- ▶ Hierarchical Softmax

Two context representations

- ▶ Continuous Bag of Words (CBOW)
- ▶ Skip-grams

How does word2vec work?

word2vec implements several different algorithms:

Two training methods

- ▶ **Negative Sampling**
- ▶ Hierarchical Softmax

Two context representations

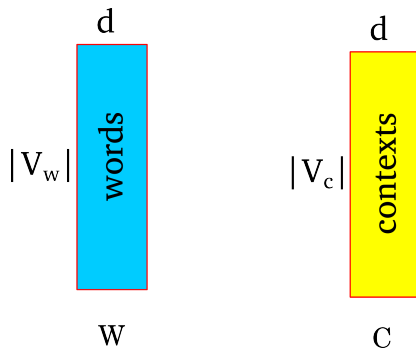
- ▶ Continuous Bag of Words (CBOW)
- ▶ **Skip-grams**

We'll focus on skip-grams with negative sampling.

intuitions apply for other models as well.

How does word2vec work?

- ▶ Represent each word as a d dimensional vector.
- ▶ Represent each context as a d dimensional vector.
- ▶ Initialize all vectors to random weights.
- ▶ Arrange vectors in two matrices, W and C .



How does word2vec work?

While more text:

- ▶ Extract a word window:

A springer is [a cow or **heifer** close to calving].

c_1 c_2 c_3 w c_4 c_5 c_6

- ▶ w is the focus word vector (row in W).
- ▶ c_i are the context word vectors (rows in C).

How does word2vec work?

While more text:

- ▶ Extract a word window:

A springer is [a cow or **heifer** close to calving] .
 c_1 c_2 c_3 w c_4 c_5 c_6

- ▶ Try setting the vector values such that:

$$\sigma(w \cdot c_1) + \sigma(w \cdot c_2) + \sigma(w \cdot c_3) + \sigma(w \cdot c_4) + \sigma(w \cdot c_5) + \sigma(w \cdot c_6)$$

is **high**

How does word2vec work?

While more text:

- ▶ Extract a word window:

A springer is [a cow or **heifer** close to calving] .
 c_1 c_2 c_3 w c_4 c_5 c_6

- ▶ Try setting the vector values such that:

$$\sigma(w \cdot c_1) + \sigma(w \cdot c_2) + \sigma(w \cdot c_3) + \sigma(w \cdot c_4) + \sigma(w \cdot c_5) + \sigma(w \cdot c_6)$$

is **high**

- ▶ Create a corrupt example by choosing a random word w'

[a cow or **comet** close to calving]
 c_1 c_2 c_3 w' c_4 c_5 c_6

- ▶ Try setting the vector values such that:

$$\sigma(w' \cdot c_1) + \sigma(w' \cdot c_2) + \sigma(w' \cdot c_3) + \sigma(w' \cdot c_4) + \sigma(w' \cdot c_5) + \sigma(w' \cdot c_6)$$

is **low**

How does word2vec work?

The training procedure results in:

- ▶ $w \cdot c$ for **good** word-context pairs is **high**.
- ▶ $w \cdot c$ for **bad** word-context pairs is **low**.
- ▶ $w \cdot c$ for **ok-ish** word-context pairs is **neither high nor low**.

As a result:

- ▶ Words that share many contexts get close to each other.
- ▶ Contexts that share many words get close to each other.

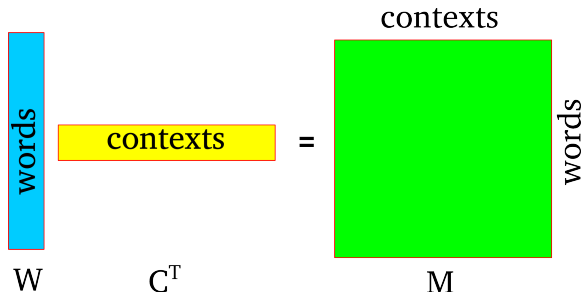
At the end, word2vec throws away C and returns W .

Reinterpretation

Imagine we didn't throw away C . Consider the product WC^T

Reinterpretation

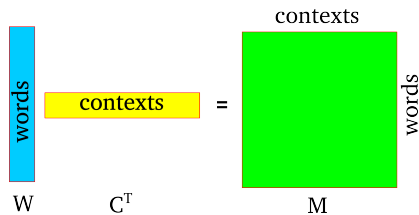
Imagine we didn't throw away C . Consider the product WC^T



The result is a matrix M in which:

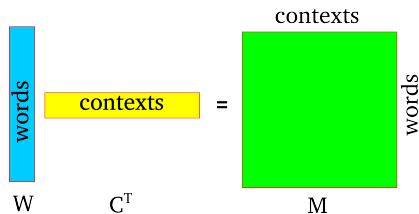
- ▶ Each row corresponds to a word.
- ▶ Each column corresponds to a context.
- ▶ Each cell correspond to $w \cdot c$, an association measure between a word and a context.

Reinterpretation



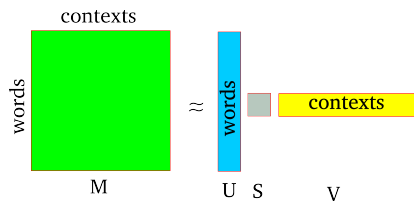
Does this remind you of something?

Reinterpretation



Does this remind you of something?

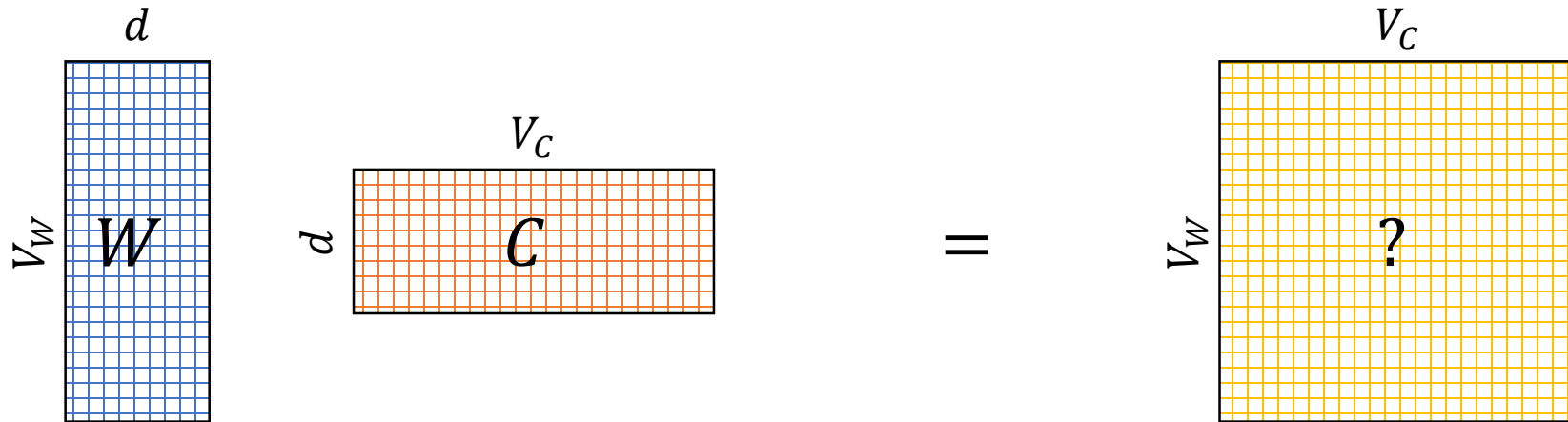
Very similar to SVD over distributional representation:



What is SGNS learning?

- A $V_W \times V_C$ matrix
- Each cell describes the relation between a specific word-context pair

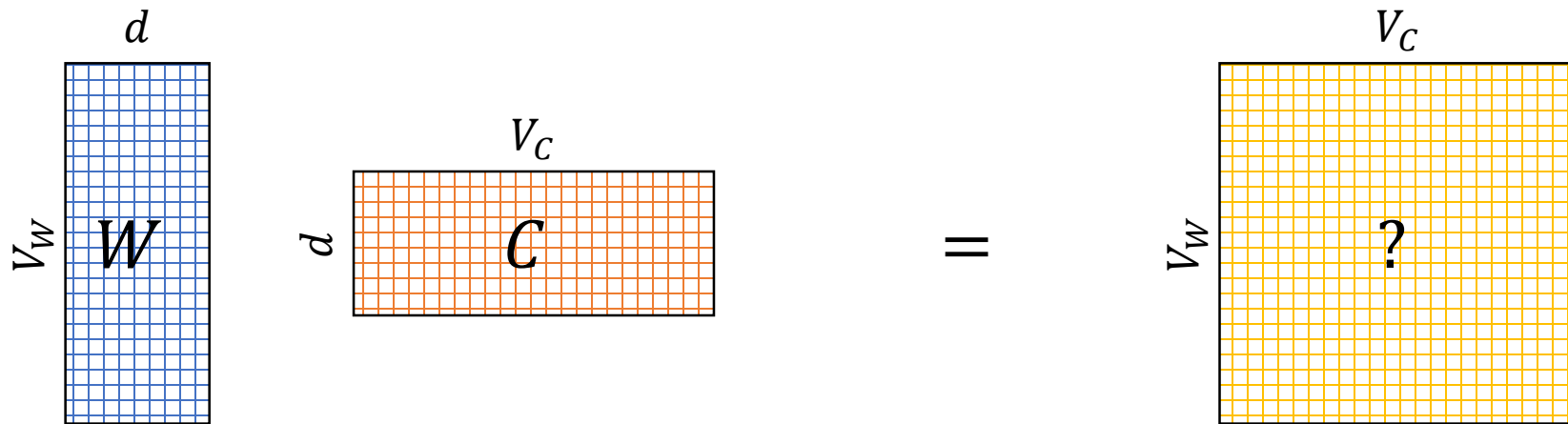
$$\vec{w} \cdot \vec{c} = ?$$



“Neural Word Embeddings as Implicit Matrix Factorization”
Levy & Goldberg, NIPS 2014

What is SGNS learning?

- We **prove** that for large enough d and enough iterations



The diagram illustrates the matrix factorization equation $W \cdot C = ?$. On the left, matrix W is represented by a blue grid with dimensions V_W (vertical) and d (horizontal). Next to it is matrix C , represented by an orange grid with dimensions d (vertical) and V_C (horizontal). An equals sign follows, leading to a yellow grid representing the product matrix, which has dimensions V_W (vertical) and V_C (horizontal) and contains a question mark.

“Neural Word Embeddings as Implicit Matrix Factorization”
Levy & Goldberg, NIPS 2014

What is SGNS learning?

- We **prove** that for large enough d and enough iterations
- We get the word-context PMI matrix

The diagram illustrates the matrix multiplication $W \cdot C = M^{PMI}$. Matrix W is a vertical grid of size $V_W \times d$ with a blue grid pattern. Matrix C is a horizontal grid of size $d \times V_C$ with an orange grid pattern. The resulting matrix M^{PMI} is a vertical grid of size $V_W \times V_C$ with a yellow grid pattern. The dimensions are labeled as follows: V_W for the height of W and M^{PMI} , d for the width of W and the height of C , and V_C for the width of C and the height of M^{PMI} .

“Neural Word Embeddings as Implicit Matrix Factorization”
Levy & Goldberg, NIPS 2014

What is SGNS learning?

- We **prove** that for large enough d and enough iterations
- We get the word-context PMI matrix, shifted by a global constant

$$Opt(\vec{w} \cdot \vec{c}) = PMI(w, c) - \log k$$

The diagram illustrates the SGNS learning process. It shows a vertical blue grid matrix W of size V_W by d , a horizontal orange grid matrix C of size d by V_C , and an equals sign followed by a vertical yellow grid matrix M^{PMI} of size V_W by V_C , with $-\log k$ to its right.

What is SGNS learning?

- SGNS is doing something very similar to the older approaches
- SGNS is factorizing the traditional word-context PMI matrix
- So does SVD!
- Do they capture the same similarity function?

SGNS vs SVD

Target Word	SGNS	SVD
cat	dog rabbit cats poodle pig	dog rabbit pet monkey pig

SGNS vs SVD

Target Word	SGNS	SVD
wine	wines grape grapes winemaking tasting	wines grape grapes varietal vintages

SGNS vs SVD

Target Word	SGNS	SVD
November	October December April January July	October December April June March

But `word2vec` is still better, isn't it?

- Plenty of evidence that `word2vec` outperforms traditional methods
 - In particular: “Don’t count, predict!” (Baroni et al., 2014)
- How does this fit with our story?

The Big Impact of “Small” Hyperparameters

Hyperparameters

- `word2vec` is more than just an algorithm...
- Introduces many **engineering tweaks** and **hyperparameter settings**
 - May seem minor, but **make a big difference** in practice
 - Their impact is often more significant than the embedding algorithm's
- These modifications can be ported to distributional methods!

rant number 1

- ACL sessions this year:

rant number 1

- ACL sessions this year:
 - Semantics: Embeddings
 - Semantics: Distributional Approaches
 - Machine Learning: Embeddings
 - Lexical Semantics
- **ALL THE SAME THING.**

key point

- Nothing magical about embeddings.
- It is just the same old distributional word similarity in a shiny new dress.

what am I going
to talk about
in the remaining time?



yoav goldberg

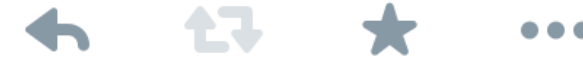
@yoavgo

ok [#acl2015](#), let's pretend this is a democracy. I'm giving a talk at the CVSC workshop. What do you want it to be about? (within topic, obv)

giving a talk at the CVSSC workshop. What do you want it to be about? (within topic, obv)

7/21/15, 7:48 PM

1 RETWEET 1 FAVORITE



Manaal Faruqui @manaalfar

17h

@yoavgo word vecs are semantic or syntactic? why dont they work well for syntactic probs? Should we continue or stop working on them?



Manaal Faruqui @manaalfar

17h

@yoavgo how to best compare two diff vector models? abolish (toy?) tasks like word-sim, word-analogy? how to standardize evaluation?



Manaal Faruqui @manaalfar

17h

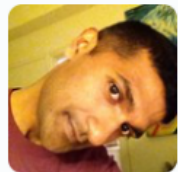
@yoavgo what else is left to be done in word vectors? (apart from word senses) this might help me select my next project :) thanks!



craig pfeifer @aCraigPfeifer

17h

@yoavgo impact of data on vector representations. How much data is enough and where is the knee on the curve?



Delip Rao @deliprao

16h

@manaalfar @yoavgo why abolish any task? Is proving something on a different task, say parsing, some how make it more useful?



Michaël Benesty @pommedeterre33

16h

@yoavgo document vector representation. Not just sentences or paragraph sizes but newspaper article or legal contract size. With use cases?



Michaël Benesty

@pommedeterre33



 Follow

@yoavgo reading other answers, main point is that people really want you to speak about vec representation! No 1 seems to care about parsers?

7:34 PM - 21 Jul 2015



Reply to @pommedeterre33



yoav goldberg @yoavgo · 12h

@pommedeterre33 :((well, but it is a vector-space workshop after all.



sort-of a global trend

Compare Search terms ▼

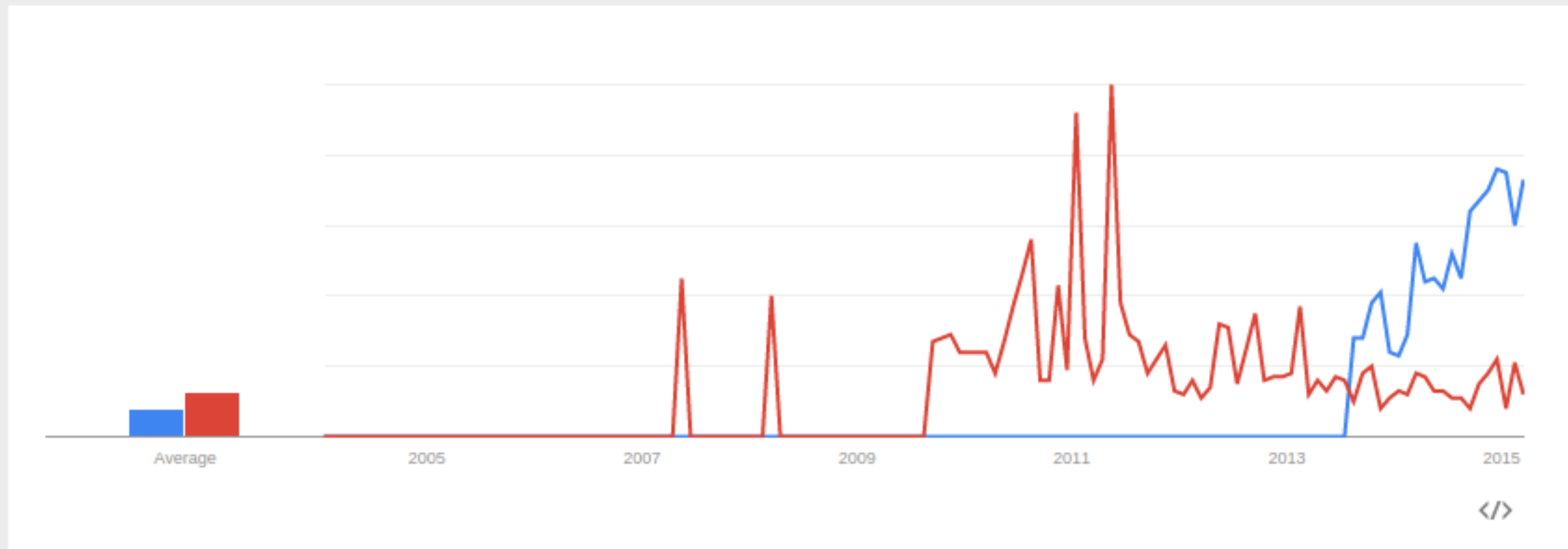
word2vec
Search term

dependency parsing
Search term

+ Add term

Interest over time ?

News headlines ? Forecast ?





craig pfeifer

@aCraigPfeifer



Following

@yoavgo impact of data on vector representations. How much data is enough and where is the knee on the curve?



craig pfeifer
@aCraigPfeifer



Following

@yoavgo impact of data on vector representations. How much data is enough and where is the knee on the curve?

- I have no idea.
- I guess you'd like each word in the vocabulary you care about to get enough examples.
- How much is enough? let's say 100.

turns out I don't have good, definitive
answers for most of the questions.

but boy do I have strong opinions!



Michaël Benesty

@pommedeterre33



 Follow

@yoavgo document vector representation.
Not just sentences or paragraph sizes but
newspaper article or legal contract size.
With use cases?



Michaël Benesty

@pommedeterre33



 Follow

@yoavgo document vector representation.
Not just sentences or paragraph sizes but
newspaper article or legal contract size.
With use cases?

- My first (and last) reaction:
 - Why do you want to do it?
 - No, really, what do you want your document representation to capture?
 - We'll get back to this later.
- But now, let's talk about...

the magic of cbow

the magic of cbow

- Represent a sentence / paragraph / document as a (weighted) average vectors of its words.
- Now we have a single, 100-dim representation of the text.
- Similar texts have similar vectors!
- Isn't this magical? (no)

the math of cbow

the math of cbow

$$\vec{o} \subset A = A \perp B \perp C$$

$$\vec{o} \subset B = X \perp Y \perp Z$$

$$C \subset \left(\vec{o} \subset A, \vec{o} \subset B \right) =$$

$$\vec{o} \subset A \cdot \vec{o} \subset B$$

$$\| \vec{o} \subset A \| \cdot \| \vec{o} \subset B \|^2$$

the math of cbow

$$\frac{\downarrow \circ \subset A \cdot \downarrow \circ \subset B}{\| \downarrow \circ \subset A \| \cdot \| \downarrow \circ \subset B \|} =$$

$$(A + B + C) \cdot (x + y + z)$$

$$\| A + B + C \| \cdot \| x + y + z \|$$

the math of cbow

$$\begin{aligned} (A + B + C)(x + y + z) = & \\ A \cdot x + A \cdot y + A \cdot z & \\ + B \cdot x + B \cdot y + B \cdot z & \\ + C \cdot x + C \cdot y + C \cdot z & \end{aligned}$$

the magic of cbow

- It's all about (weighted) all-pairs similarity
 - ... done in an efficient manner.
- That's it. no more, no less.
- I'm amazed by how few people realize this.

(the math is so simple... even I could do it)

this also explains
king-man+woman

this also explains
king-man+woman

$$\begin{array}{l} \text{argmax}_x \quad \cos(x, k - m + w) = \\ \text{argmax}_x \quad \frac{x \cdot (k - m + w)}{\|x\| \|k - m + w\|} \end{array}$$



1

constant

$$= \text{argmax}_x \quad x \cdot k - x \cdot m + x \cdot w$$

similarity arithmetic!!

and once we understand
we can improve

and once we understand
we can improve

$$XK - XM \rightarrow XL$$

additive.

one term can dominate.

and once we understand
we can improve

$$XK - Xm \rightarrow XW$$

additive.

one term can dominate.

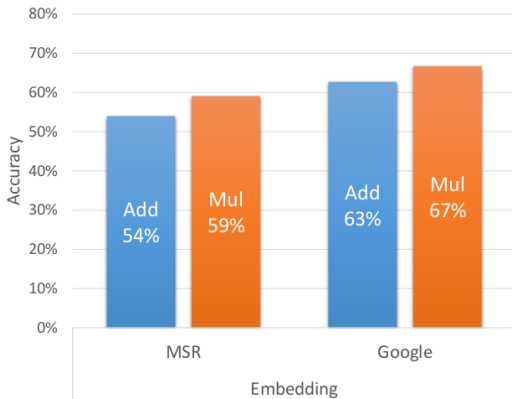


$$\frac{(XK) \cdot (XW)}{(X \cdot m)}$$

multiplicative.

much more balanced.

Multiplication > Addition



math > magic

can we improve analogies
even further?

which brings me to:



Manaal Faruqui

@manaalfar



Follow

@yoavgo how to best compare two diff
vector models? abolish (toy?) tasks like
word-sim, word-analogy? how to
standardize evaluation?

which brings me to:



Manaal Faruqui
@manaalfar



+ Follow

@yoavgo how to best compare two diff
vector models? abolish (toy?) tasks like
word-sim, word-analogy? how to
standardize evaluation?

- **Yes. Please stop evaluating on word analogies.**
- It is an artificial and useless task.
- Worse, it is just a proxy for (a very particular kind of) word similarity.
- Unless you have a good use case, don't do it.
- **Alternatively:** show that it correlates well with a real and useful task.



Manaal Faruqui

@manaalfar



 Follow

@yoavgo how to best compare two diff vector models? abolish (toy?) tasks like word-sim, word-analogy? how to standardize evaluation?



SJ

@TheAshenLight



 Follow

@yoavgo evaluation of vector models - how to comparatively evaluate models. Deciding on a downstream task (and/or domain?) for evaluation

let's take a step back

- We don't really care about the vectors.
- We care about the **similarity function** they induce.
 - (or, maybe we want to use them in an external task)
- We want similar words to have similar vectors.
- So evaluating on word-similarity tasks is great.
- **But what does similar mean?**

many faces of similarity

- dog -- cat
- dog -- poodle
- dog -- animal
- dog -- bark
- dog -- leash

many faces of similarity

- dog -- cat
- dog -- poodle
- dog -- animal
- dog -- bark
- dog -- leash
- dog -- chair
- dog -- dig
- dog -- god
- dog -- fog
- dog -- 6op

many faces of similarity

- dog -- cat
- dog -- poodle
- dog -- animal
- dog -- bark
- dog -- leash

- dog -- chair same POS
- dog -- dig edit distance
- dog -- god same letters
- dog -- fog rhyme
- dog -- 6op shape

some forms of similarity look more useful than they really are

- Almost every algorithm you come up with will be good at capturing:
 - countries
 - cities
 - months
 - person names

some forms of similarity look more useful than they really are

- Almost every algorithm you come up with will be good at capturing:

- countries

useful for tagging/parsing/NER

- cities

- months

- person names

some forms of similarity look more useful than they really are

- Almost every algorithm you come up with will be good at capturing:

- countries

useful for tagging/parsing/NER

- cities

- months

- person names

but do we really want
"John went to China in June"
to be similar to
"Carl went to Italy in February"
??

there is no single downstream task

- Different tasks require different kinds of similarity.
- Different vector-inducing algorithms produce different similarity functions.
- **No single representation for all tasks.**
- If your vectors do great on task X, I don't care that they suck on task Y.

"but my algorithm works great for all these
different word-similarity datasets!
doesn't it mean something?"

"but my algorithm works great for all these
different word-similarity datasets!
doesn't it mean something?"

- Sure it does.
- It means these datasets are not diverse enough.
- They should have been a single dataset.
- (**alternatively**: our evaluation metrics are not discriminating enough.)

which brings us back to:



Michaël Benesty
@pommedeterre33



+ Follow

@yoavgo document vector representation.
Not just sentences or paragraph sizes but
newspaper article or legal contract size.
With use cases?

- This is really, really ill-defined.
- What does it mean for legal contracts to be similar?
- What does it mean for newspaper articles to be similar?
- Think about this before running to design your next super-LSTM-recursive-autoencoding-document-embedder.
- **Start from the use case!!!!**

case in point:

Skip-Thought Vectors

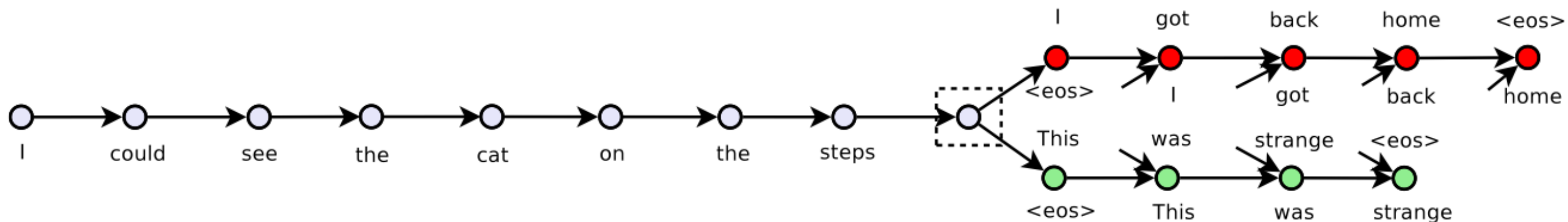
Ryan Kiros¹, Yukun Zhu¹, Ruslan Salakhutdinov^{1,2}, Richard S. Zemel¹

Antonio Torralba³, Raquel Urtasun¹, Sanja Fidler¹

University of Toronto¹

Canadian Institute for Advanced Research²

Massachusetts Institute of Technology³



skip thought vectors

- Terrible name. (really)
- Beautiful idea. (really!)
- Impressive results.

Impressive results:

Query and nearest sentence

he ran his hand inside his coat , double-checking that the unopened letter was still there .
he slipped his hand between his coat and his shirt , where the folded copies lay in a brown envelope .

im sure youll have a glamorous evening , she said , giving an exaggerated wink .
im really glad you came to the party tonight , he said , turning to her .

although she could tell he had n't been too invested in any of their other chitchat , he seemed genuinely curious about this .
although he had n't been following her career with a microscope , he 'd definitely taken notice of her appearances .

an annoying buzz started to ring in my ears , becoming louder and louder as my vision began to swim .
a weighty pressure landed on my lungs and my vision blurred at the edges , threatening my consciousness altogether .

if he had a weapon , he could maybe take out their last imp , and then beat up errol and vanessa .
if he could ram them from behind , send them sailing over the far side of the levee , he had a chance of stopping them .

then , with a stroke of luck , they saw the pair head together towards the portaloos .
then , from out back of the house , they heard a horse scream probably in answer to a pair of sharp spurs digging deep into its

“ i 'll take care of it , ” goodman said , taking the phonebook .
“ i 'll do that , ” julia said , coming in .

he finished rolling up scrolls and , placing them to one side , began the more urgent task of finding ale and tankards .
he righted the table , set the candle on a piece of broken plate , and reached for his flint , steel , and tinder .

Impressive results:

Query and nearest sentence

he ran his hand inside his coat , double-checking that the unopened letter was still there .
he slipped his hand between his coat and his shirt , where the folded copies lay in a brown envelope .

im sure youll have a glamorous evening , she said , giving an exaggerated wink .
im really glad you came to the party tonight , he said , turning to her .

although she could tell he had n't been too invested in any of their other chitchat , he seemed genuinely curious about this .
although he had n't been following her career with a microscope , he 'd definitely taken notice of her appearances .

an annoying buzz started to ring in my ears , becoming louder and louder as my vision began to swim .
a weighty pressure landed on my lungs and my vision blurred at the edges , threatening my consciousness altogether .

if he had a weapon , he could maybe take out their last imp , and then beat up errol and vanessa .
if he could ram them from behind , send them sailing over the far side of the levee , he had a chance of stopping them .

then , with a stroke of luck , they saw the pair head together towards the portaloos .
then , from out back of the house , they heard a horse scream probably in answer to a pair of sharp spurs digging deep into its

“ i 'll take care of it , ” goodman said , taking the phonebook .
“ i 'll do that , ” julia said , coming in .

he finished rolling up scrolls and , placing them to one side , began the more urgent task of finding ale and tankards .
he righted the table , set the candle on a piece of broken plate , and reached for his flint , steel , and tinder .

- **Is this actually useful? what for?**
- **Is this the kind of similarity we need?**

so how to evaluate?

- Define the similarity / task you care about.
- **Score on this particular similarity / task.**
- Design your vectors to match this similarity
- ...and since the methods we use are distributional and unsupervised...
- ...design has less to do with the fancy math (= objective function, optimization procedure) and more with what you feed it.

context matters

What's in a Context?

- Importing ideas from embeddings improves distributional methods
- Can distributional ideas also improve embeddings?
- **Idea:** change SGNS's default **BoW contexts** into **dependency contexts**

Example

Australian scientist discovers star with telescope

Target Word

Australian scientist discovers star with telescope

Bag of Words (BoW) Context

Australian scientist discovers star with telescope

Bag of Words (BoW) Context

Australian scientist discovers star with telescope

Bag of Words (BoW) Context

Australian scientist discovers star with telescope

Syntactic Dependency Context

Australian scientist discovers star with telescope

Syntactic Dependency Context



"Dependency-Based Word Embeddings"
Levy & Goldberg, ACL 2014

Syntactic Dependency Context



“Dependency-Based Word Embeddings”
Levy & Goldberg, ACL 2014

Embedding Similarity with Different Contexts

Target Word	Bag of Words (k=5)	Dependencies
Hogwarts (Harry Potter's school)	Dumbledore hallows half-blood Malfoy Snape	Sunnydale Collinwood Calarts Greendale Millfield

**Related to
Harry Potter**

Schools

“Dependency-Based Word Embeddings”
Levy & Goldberg, ACL 2014

Embedding Similarity with Different Contexts

Target Word	Bag of Words (k=5)	Dependencies
Turing (computer scientist)	nondeterministic non-deterministic computability deterministic finite-state	Pauling Hotelling Heting Lessing Hamming

**Related to
computability**

Scientists

“Dependency-Based Word Embeddings”
Levy & Goldberg, ACL 2014

Embedding Similarity with Different Contexts

Target Word	Bag of Words (k=5)	Dependencies
dancing (dance gerund)	singing dance dances dancers tap-dancing	singing rapping breakdancing miming busking

**Related to
dance**

Gerunds

“Dependency-Based Word Embeddings”
Levy & Goldberg, ACL 2014

What is the effect of different context types?

- Thoroughly studied in distributional methods
 - Lin (1998), Padó and Lapata (2007), and many others...

General Conclusion:

- Bag-of-words contexts induce *topical* similarities
- Dependency contexts induce *functional* similarities
 - Share the same semantic type
 - Cohyponyms
- Holds for **embeddings** as well

- Same algorithm, different inputs -- very different kinds of similarity.
- Inputs matter much more than algorithm.
- **Think about your inputs.**



Manaal Faruqui

@manaalfar



Follow

@yoavgo word vecs are semantic or syntactic? why dont they work well for syntactic probs? Should we continue or stop working on them?



Manaal Faruqui

@manaalfar



Follow

@yoavgo word vecs are semantic or syntactic? why dont they work well for syntactic probs? Should we continue or stop working on them?

- They are neither semantic nor syntactic.
- They are what you design them to be through context selection.
- They seem to work better for semantics than for syntax because, unlike syntax, we never quite managed to define what "semantics" really means, so everything goes.

with proper care, we can perform well on syntax, too.

- Ling, Dyer, Black and Trancoso, NAACL 2015: using positional contexts with a small window size work well for capturing parts of speech, and as features for a neural-net parser.
- In our own work, we managed to derive good features for a graph-based parser (in submission).
- also related: many parsing results at this ACL.



Manaal Faruqui

@manaalfar



Follow

@yoavgo what else is left to be done in word vectors? (apart from word senses) this might help me select my next project :) thanks!



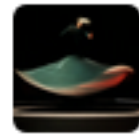
Manaal Faruqui

@manaalfar



+ Follow

@yoavgo what else is left to be done in word vectors? (apart from word senses) this might help me select my next project :) thanks!



SJ

@TheAshenLight



+ Follow

@yoavgo effective ways to integrate linguistic information with VSMs. Existing techniques or required research directions.

what's left to do?

- Pretty much nothing, and pretty much everything.
- Word embeddings are just a small step on top of distributional lexical semantics.
- All of the previous open questions remain open, including:
 - composition.
 - multiple senses.
 - multi-word units.

looking beyond words

- word2vec will easily identify that "hotfix" is similar to "hf", "hot-fix" and "patch"
- But what about "hot fix"?
- How do we know that "New York" is a single entity?
- Sure we can use a collocation-extraction method, but is it really the best we can do? can't it be integrated in the model?



Djamé
@zehavoc



 Follow

@manaalfar @yoavgo Something on the usefulness of word2vec acquired from edited text when used on really noisy data (possibly MRL) ?



Djamé
@zehavoc



+ Follow

@manaalfar @yoavgo Something on the usefulness of word2vec acquired from edited text when used on really noisy data (possibly MRL) ?

- Actually works pretty well
- But would be nice to be able to deal with typos and spelling variations without relying only on seeing them enough times in the corpus.
- I believe some people are working on that.



Djamé
@zehavoc



 Follow

@manaalfar @yoavgo Something on the usefulness of word2vec acquired from edited text when used on really noisy data (possibly MRL) ?



Djamé
@zehavoc



 Follow

@manaalfar @yoavgo Something on the usefulness of word2vec acquired from edited text when used on really noisy data (possibly MRL) ?

#####



Djamé
@zehavoc



+ Follow

@manaalfar @yoavgo Something on the usefulness of word2vec acquired from edited text when used on really noisy data (possibly MRL) ?

REDACTED

MRL: morphologically rich language

what happens when we look outside of English?

- Things don't work nearly as well.
- Known problems from English become more extreme.
- We get some new problems as well.

a quick look at Hebrew

word senses

ספר

book(N). barber(N). counted(V). tell!(V). told(V).

חומה

brown (feminine, singular)

wall (noun)

her fever (possessed noun)

multi-word units

- עורך דין
- בית ספר
- שומר ראש
- יושב ראש
- ראש עיר
- בית שימוש

words vs. tokens

וכשמהבית

and when from the house

words vs. tokens

וכשמהבית

and when from the house

בצל

in shadow

בצל

onion

and of course: inflections

- nouns, pronouns and adjectives
--> are inflected for *number* and *gender*
- verbs
--> are inflected for *number*, *gender*, *tense*, *person*
- syntax requires *agreement* between
 - nouns and adjectives
 - verbs and subjects

and of course: inflections

she **saw** a **brown** fox

he **saw** a **brown** fence

and of course: inflections

[fem] [masc]
she **saw** a **brown** fox

he **saw** a **brown** fence
[masc] [fem]

and of course: inflections

[fem] [masc]

she **saw** a **brown** fox

היא **ראתה** שועל **חום**

הוא **ראה** גדר **חומה**

he **saw** a **brown** fence
[masc] [fem]

inflections and dist-sim

- More word forms -- more sparsity
- But more importantly: *agreement patterns affect the resulting similarities.*

adjectives

green [m,sg] ירוק	green [f,sg] ירוקה	green [m,pl] ירוקים
blue [m,sg]	gray [f,sg]	gray [m,pl]
orange [m,sg]	orange [f,sg]	blue [m,pl]
yellow [m,sg]	yellow [f,sg]	black [m,pl]
red [m,sg]	magical [f,g]	heavenly [m,pl]

verbs

(he) walked הלך	(she) thought חשבה	(they) ate אכלו
(they) walked	(she) is thinking	(they) will eat
(he) is walking	(she) felt	(they) are eating
(he) turned	(she) is convinved	(he) ate
(he) came closer	(she) insisted	(they) drank

nouns

Doctor [m,sg]
רופא

Doctor [f, sg]
רופאה

psychiatrist [m,sg]

student [f, sg]

psychologist [m, sg]

nun [f, sg]

neurologist [m, sg]

waitress [f, sg]

engineer [m, sg]

photographer [f, sg]

nouns

sweater
סוודר

shirt
חולצה

jacket

suit

down

robe

overall

dress

turban

helmet

nouns

sweater
סוודר

shirt
חולצה

jacket

suit

down

robe

overall

dress

turban

helmet

masculine

feminine

nouns

sweater סוודר	shirt חולצה
jacket	suit
down	robe
overall	dress
turban	helmet
masculine	feminine

completely arbitrary

inflections and dist-sim

- Inflections and agreement really influence the results.
- We get a mix of syntax and semantics.
- Which aspect of the similarity we care about? what does it mean to be similar?
- Need better control of the different aspects.

inflections and dist-sim

- Work with lemmas instead of words!!
- Sure, but where do you get the lemmas?
- ...for unknown words?
- And what should you lemmatize? everything?
somethings? context-dependent?
- Ongoing work in my lab -- but still much to do.



Manaal Faruqui

@manaalfar



 Follow

@yoavgo what else is left to be done in word vectors? (apart from word senses) this might help me select my next project :) thanks!

looking for an
interesting project?

choose an
interesting language!

(good luck in getting it accepted to ACL, though)

to summarize

- Magic is bad. Understanding is good. Once you Understand you can control and improve.
- Word embeddings are just distributional semantics in disguise.
- Need to think of what you actually want to solve.
--> focus on a specific task!
- Inputs >> fancy math.
- Look beyond just words.
- Look beyond just English.