Aligning Packed Dependency Trees: a theory of composition for distributional semantics

JULIE WEEDS, UNIVERSITY OF SUSSEX, SPRING 2017

Based on joint work with David Weir, Jeremy Reffin and Thomas Kober

Overview

What is compositional distributional semantics?

Existing methods of composition

Elementary Anchored Packed Trees (APTs)

Composition (as Contextualisation)

Similarity

Experimental results

Conclusions, applications and further work

Compositional distributional semantics

?

Compositional distributional semantics

Semantics → the study of the meanings of words and phrases in a language

Compositional distributional semantics

Distributional → based on the position, arrangement or frequency of occurrence of members of a group throughout some space

Semantics → the study of the meanings of words and phrases in a language

Compositional distributional semantics

Compositional → based on the product of mixing or combining various elements or ingredients

Distributional → based on the position, arrangement or frequency of occurrence of members of a group throughout some space

Semantics → the study of the meanings of words and phrases in a language

Distributional Semantics 101

The distributional hypothesis:- words that occur in the same contexts tend to have similar meaning (Harris, 1954)

You shall know a word by the company it keeps (Firth, 1957)

Distributional Semantics 101

The distributional hypothesis:-
words that occur in the same contexts tend to have similar meaning (Harris, 1954)

You shall know a word by the company it keeps (Firth, 1957)

Both count-based and prediction-based methods of constructing distributional word representations probe the underlying co-occurrence statistics of the corpus (Pennington et al. 2014)

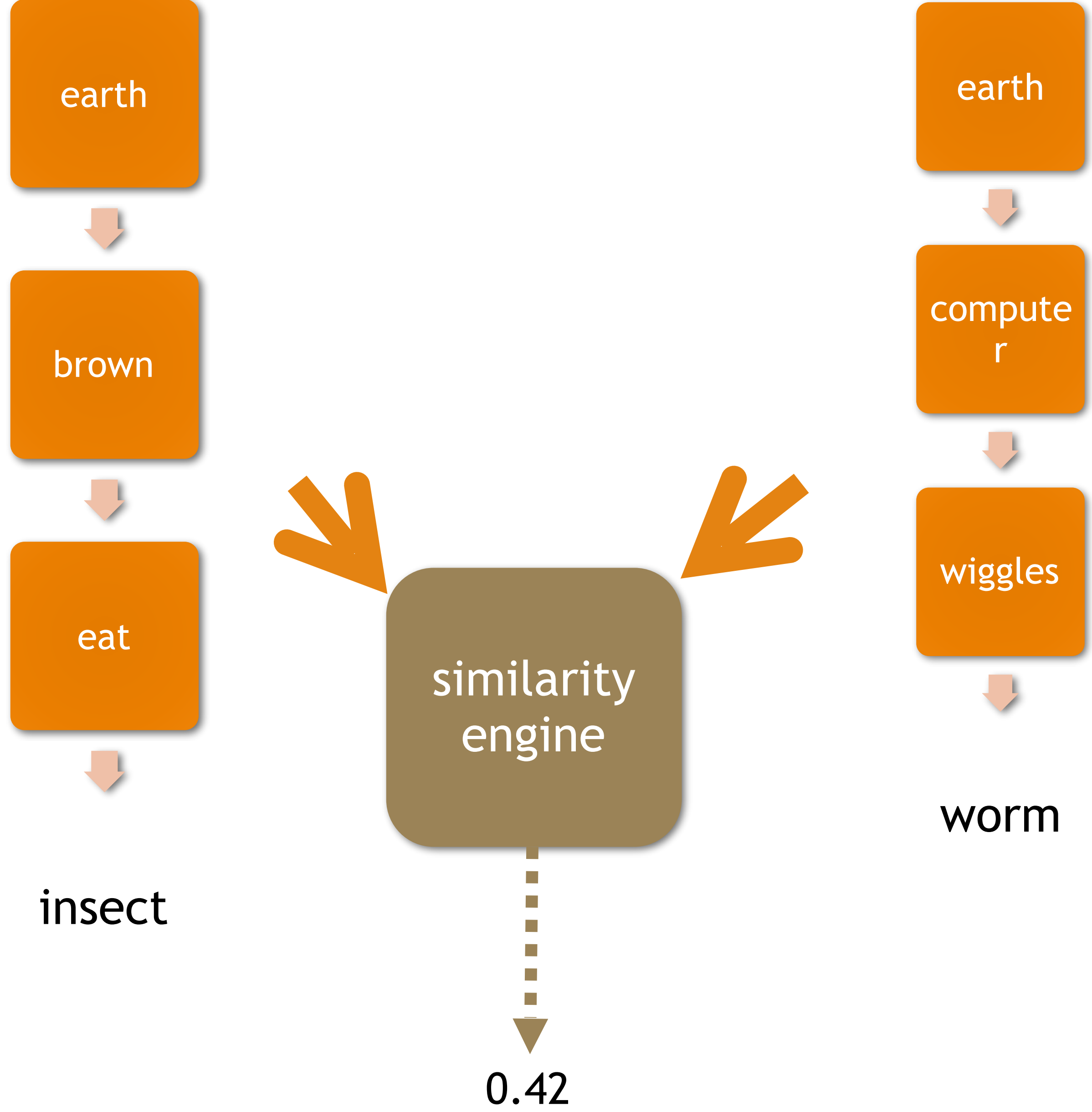
Distributional Semantics 101

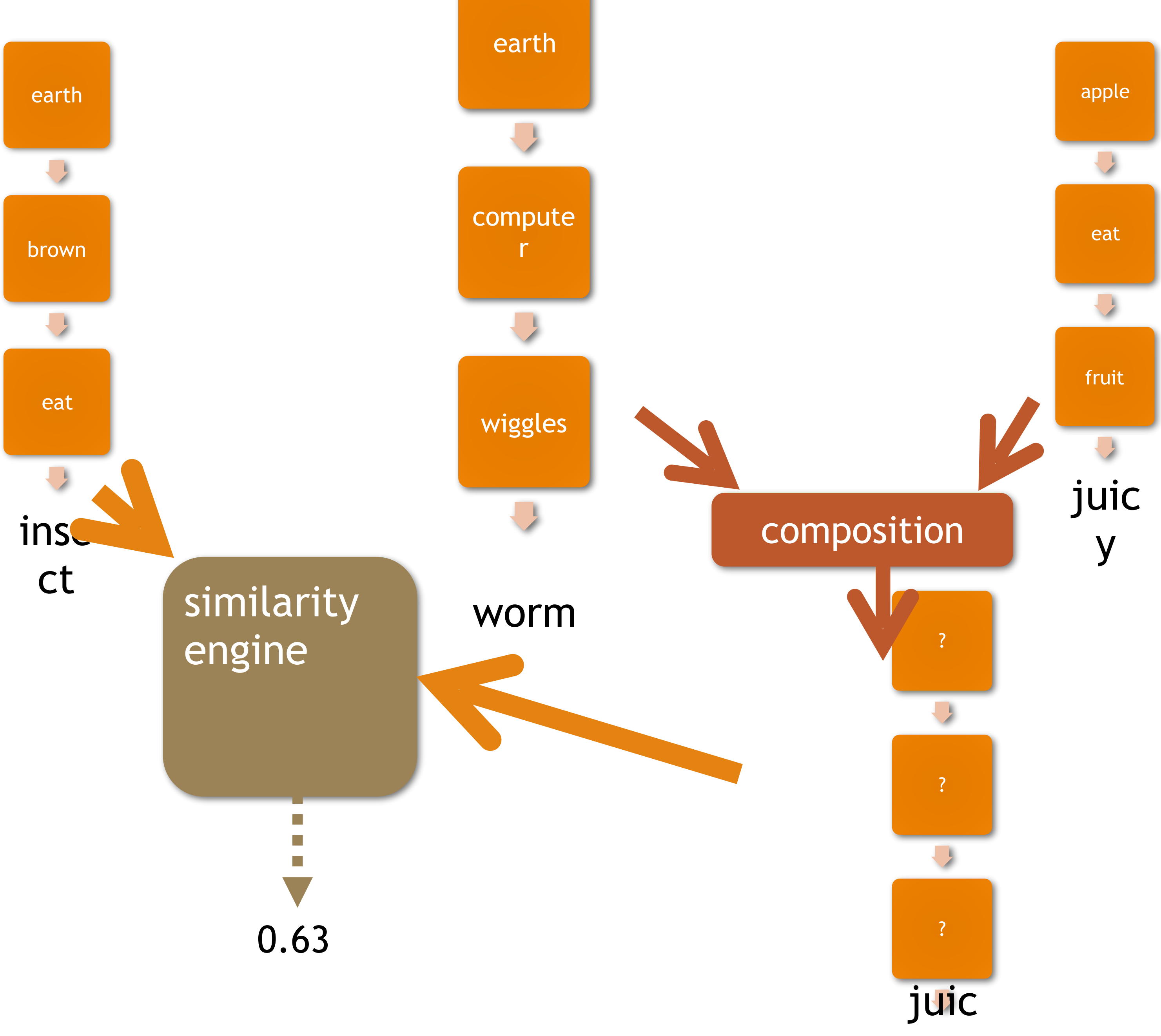
The distributional hypothesis:- words that occur in the same contexts tend to have similar meaning (Harris, 1954)

You shall know a word by the company it keeps (Firth, 1957)

Both count-based and prediction-based methods of constructing distributional word representations probe the underlying co-occurrence statistics of the corpus (Pennington et al. 2014)

The representation of a word is determined by the **contexts** in which it occurs.

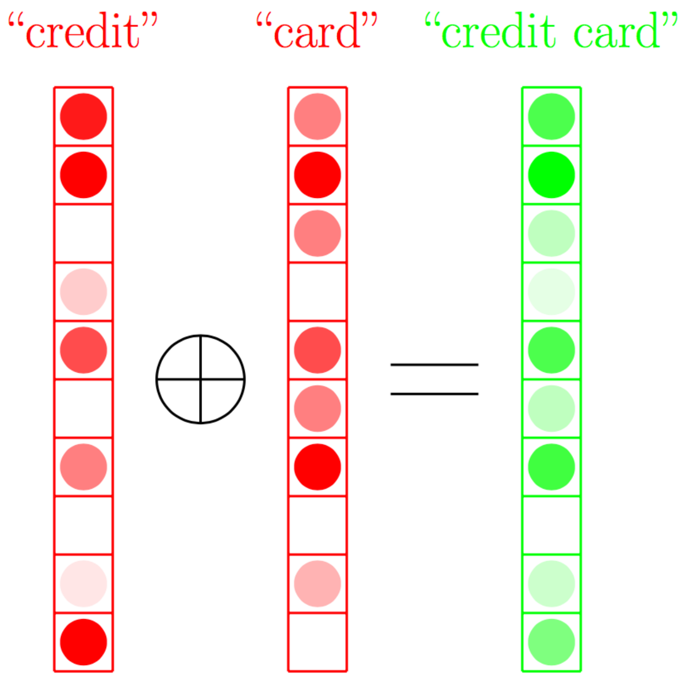

earth
brown
eat
insect
similarity engine
0.42
earth
compute r
wiggles
worm


earth
brown
eat
ins
ct
earth
compute
r
wiggles
worm
apple
eat
fruit
juic
y
composition
similarity
engine
0.63
?
?
?
juic
y
wor

Vector-based models of composition

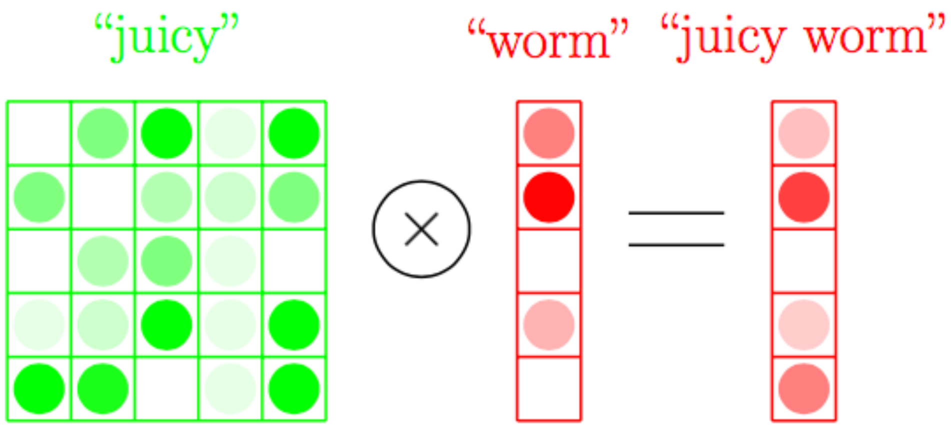
- e.g., Mitchell & Lapata (2008, 2010)
- Easy to implement.
- Word representations typically directly derived from corpus (either by counting or predicting)
- Hard to beat “naive” operations such as add (union) and multiply (intersect)
- Hard to capture interesting linguistic properties such as non-commutativity e.g.,

credit card ≠ card credit



Non vector-based models of composition

- e.g., Baroni and Zamperelli (2010) and Grefenstette et al. (2013) borrow ideas from formal semantics
- Words may be of different types
  - e.g., an adjective is a function which maps a noun to a compound noun
  - **juicy(worm) → juicy_worm**
  - adjectives modelled by matrices, nouns modelled by vectors



- At least some word representations typically learnt from observed phrasal representations
- Lots of parameters!

What should the distributional representation of a phrase or a sentence actually be?

What should the distributional representation of a phrase or a sentence actually be?

- The representation of a word is the contexts in which it occurs.

What should the distributional representation of a phrase or a sentence actually be?

- The representation of a word is the contexts in which it occurs.

➢ The representation of a phrase is the contexts in which it occurs.

What should the distributional representation of a phrase or a sentence actually be?

- The representation of a word is the contexts in which it occurs.

➢ The representation of a phrase is the contexts in which it occurs.

➢ The representation of a sentence is the contexts in which it occurs.

?

What should the distributional representation of a phrase or a sentence actually be?

- The representation of a word is the contexts in which it occurs.

➢ The representation of a phrase is the contexts in which it occurs.

➢ The representation of a sentence is the contexts in which it occurs.

?

- Should the space for sentences be the same as the space for words?

What should the distributional representation of a phrase or a sentence actually be?

- The representation of a word is the contexts in which it occurs.

➢ The representation of a phrase is the contexts in which it occurs.

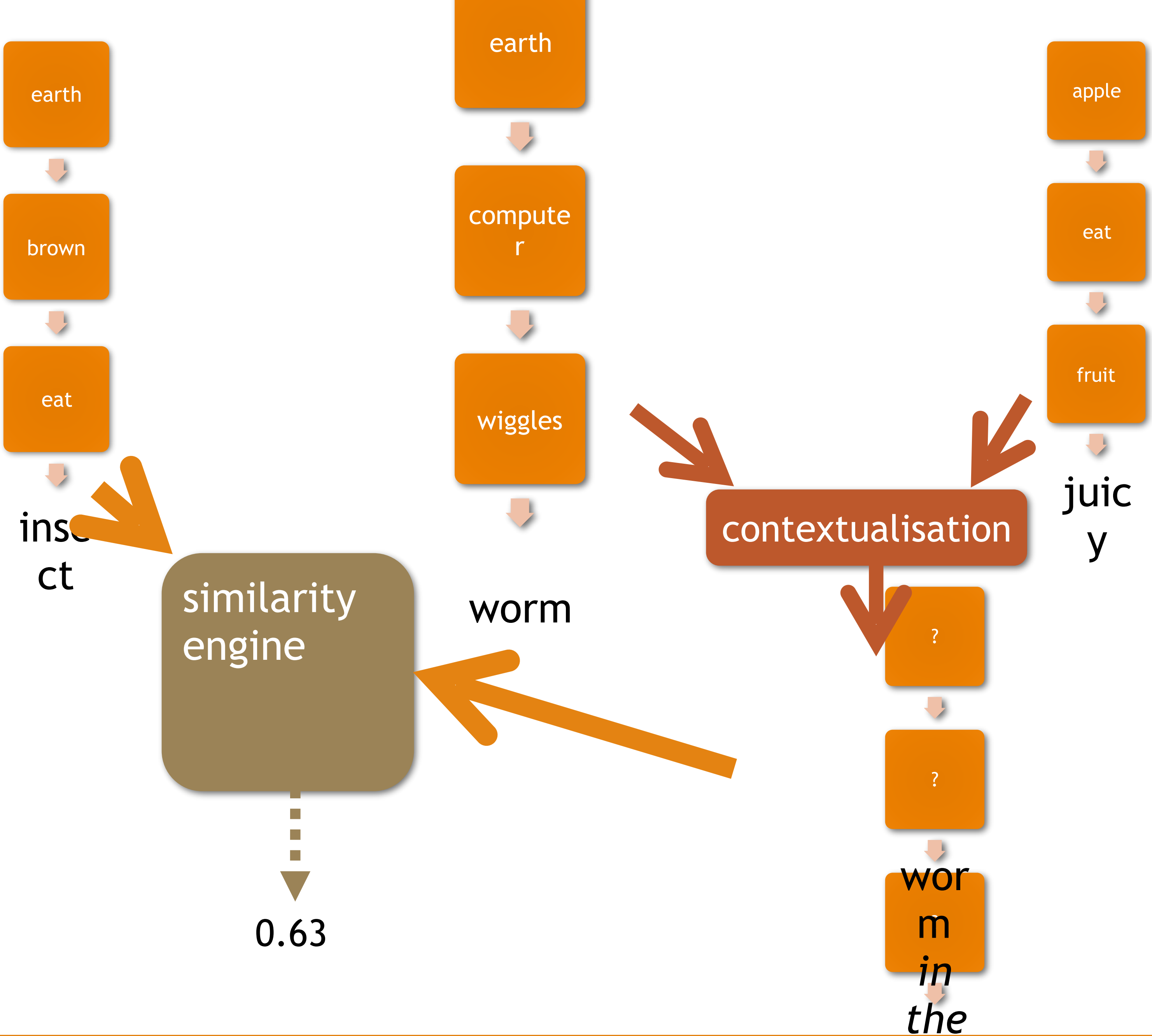
➢ The representation of a sentence is the contexts in which it occurs.

?

  - Should the space for sentences be the same as the space for words?
  - When we compose meanings, are we building a representation of something larger or smaller?

Anchored Packed Tree (APT) framework (Weir et al 2016)

- Composition is a process of mutual disambiguation or contextualisation
- Representation of a sentence is the representation of each word in the context of that sentence
- Structured, syntax-driven representations allows phrases with different structures (e.g., “credit card” vs “card credit”) to have different representations
- Uniform nature of representations for lexemes, phrases and sentences allow for direct comparison of phrases of different lengths


earth
brown
eat
insect
earth
computer
wiggles
worm
apple
eat
fruit
juicy
contextualisation
similarity engine
0.63
?
?
worm in the context

Definition of context?

What is the context of “ball” in the following sentence?

- The cricket **ball** hit the castle wall.

Definition of context?

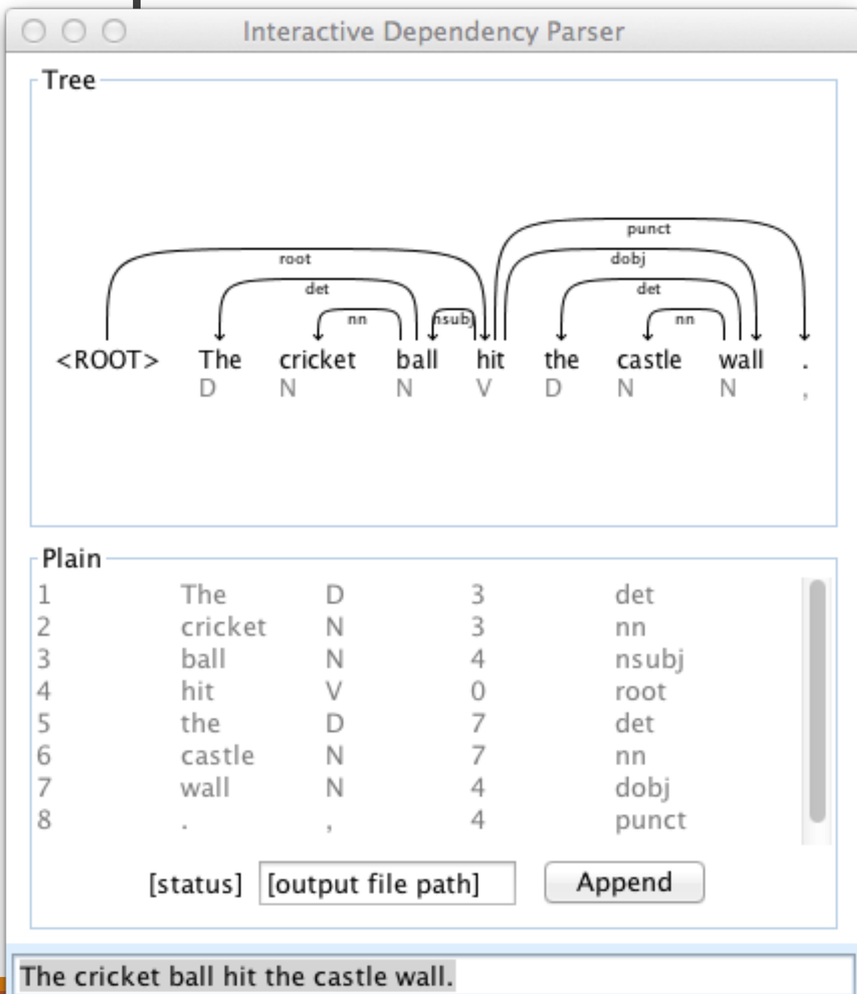
What is the context of “ball” in the following sentence?

- The cricket **ball** hit the castle wall.

- Proximity? e.g., Within n words, in the same sentence, in the same document?

	the	cricket	hit	castle	wall	....
ball	2	1	1	1	1	0

Dependency-based representations of context



	nn: cricket	_nsubj: hit	det: the	…
ball	1	1	1	0

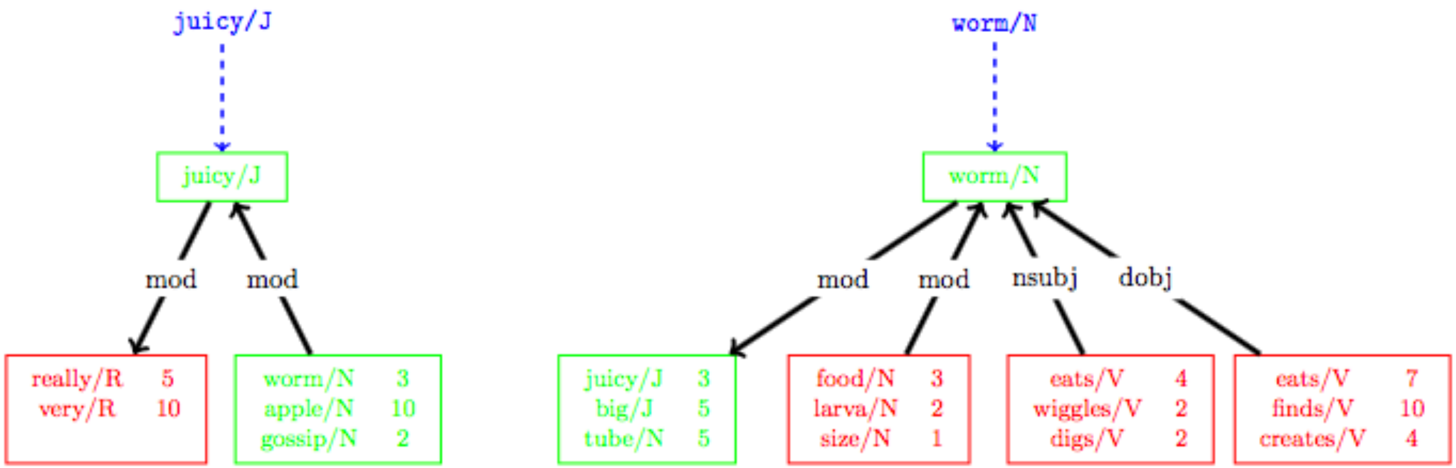
Which are the *best* nearest noun neighbours of “ball”?

neighbour	similarity
shot	0.17
stick	0.16
wheel	0.15
shell	0.15
piece	0.14
puck	0.14
bullet	0.14
barrel	0.14
ring	0.14
projectile	0.14

?

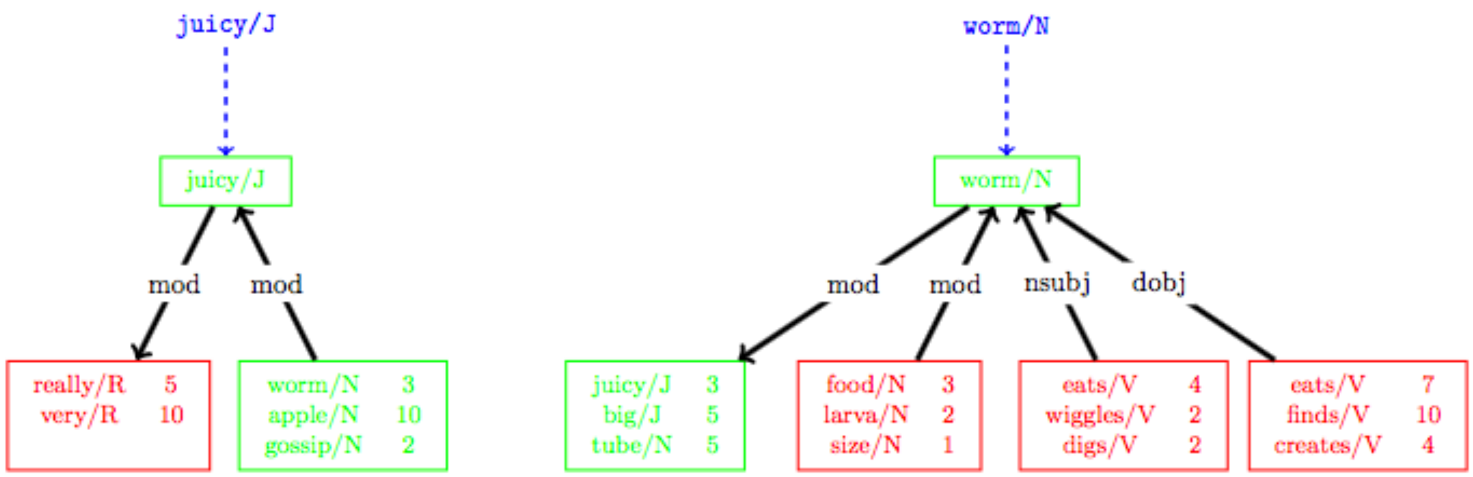
neighbour	similarity
run	0.15
game	0.15
season	0.14
play	0.14
quarter	0.13
kick	0.13
match	0.12
inning	0.12
minute	0.12
goal	0.11

Composing dependency-based representations



Nouns (N)	Verbs (V)	Adjectives (J)	Adverbs (R)
mod:_/N mod:_/J _mod:_/N _nsubj:_/V _dobj:_/V	dobj:_/N nsubj:_/N pobj:_/N mod:_/R	_mod:/_N mod:_/R	_mod:_/V _mod:_/J

Composing dependency-based representations



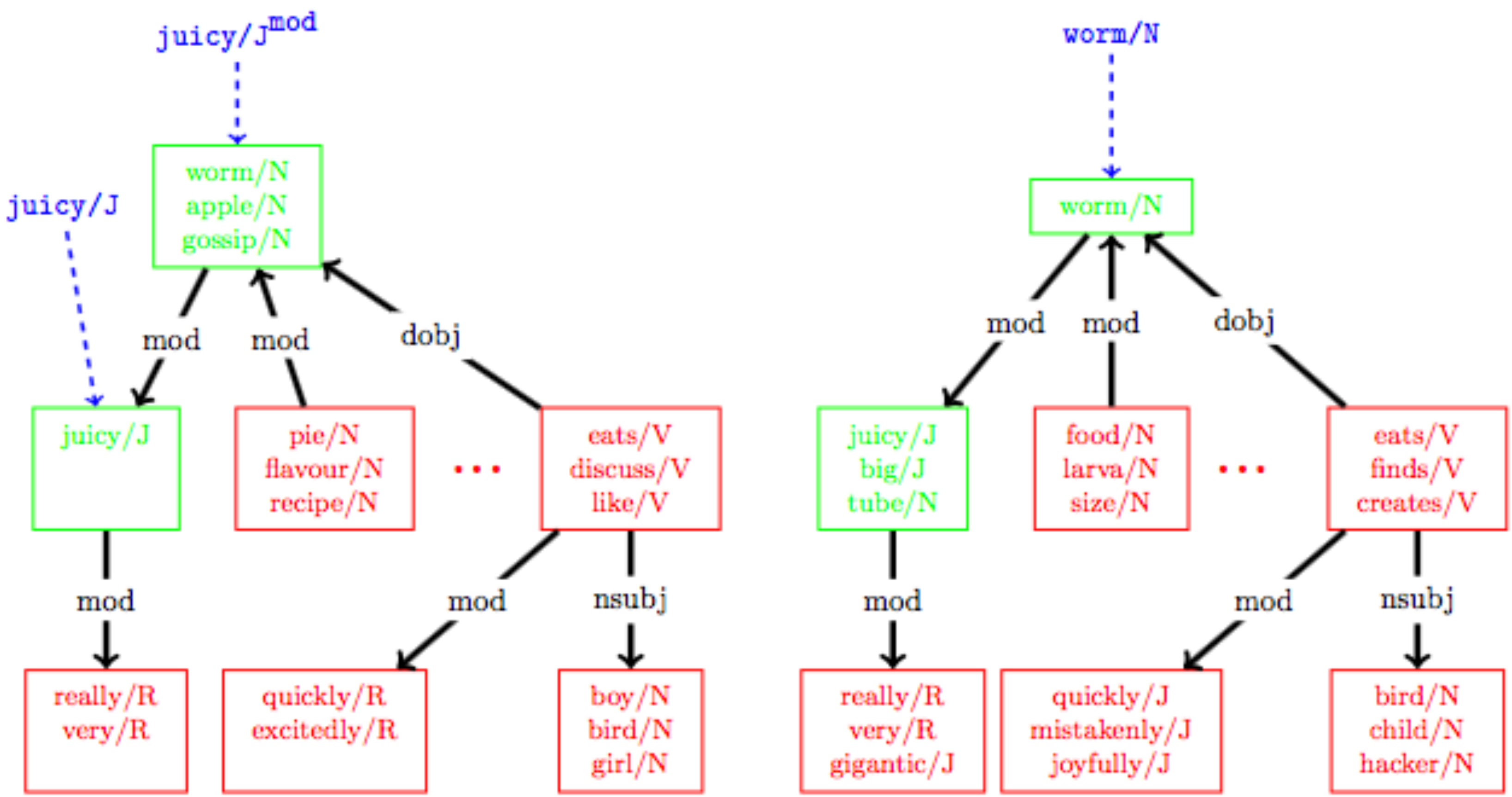
- difficult due to non-overlapping nature of feature types
- intersective methods → empty representations
- additive methods → broad, non-comparable representations

APT intuition

____ a juicy worm

- consider words which are in the _dobj relation with worm
- consider words which are in the _dobj relation with words which are in the _mod relation with juicy

Aligned anchored packed trees



What kind of worm is a *juicy worm*?

Composition requires alignment of the APTs

- *juicy* is in adjective space and *worm* is in noun space

Typically, offset non-head APTs to align with the head APT

- *worm* is the head of the phrase *juicy worm* so we offset *juicy* to make it look like a noun

If δ is the observed phrasal relation, offset all paths by prepending δ and applying reduction operation

- The relation from *worm* to *juicy* is mod so we offset all of the paths in the elementary APT by mod to produce an offset-APT which aligns with noun APTs

A[juicy]	A[juicy]^{mod}	W
<ε, juicy>	<mod,juicy>	50
<mod,really>	<mod.mod,really >	5
<_mod,apple>	<ε,apple>	7
<_mod.dobj,eats>	<dobj,eats>	10

Merging aligned APTs

- insert your favourite composition operation. This could be intersective (MIN, MULTIPLY) or more additive (ADD, MAX)
- Due to the asymmetric nature of the alignment, composition is not commutative even if a symmetric composition operation is used.

Similarity of APTs

- map APTs to vectors. Features/dimensions are the typed co-occurrences
- apply favourite measure of feature association e.g., PPMI
- optionally carry out feature selection or dimensionality reduction
- apply favourite similarity measure e.g. cosine

Disambiguation Examples: ADD

	Aligned: add		Unaligned: add	
shoot	green shoot	six-week shoot	green shoot	six-week shoot
shot leaf shooting fight scene video tour footage interview flower	shoot leaf flower fruit orange tree color shot colour cover	shoot tour shot break session show shooting concert interview leaf	shoot shot leaf shooting fight scene video tour flower footage	shoot shot shooting leaf scene video fight footage photo interview

Disambiguation Examples: MIN

	Aligned: min		Unaligned: min	
shoot	green shoot	six-week shoot	green shoot	six-week shoot
shot	shoot	shoot	shoot	e/f
leaf	leaf	photoshoot	pyrite	uemtsu
shooting	fruit	taping	plosive	confederations
fight	stalk	tour	handlebars	shortlist
scene	flower	airing	annual	all-ireland
video	twig	rehearsal	roundel	dern
tour	sprout	broadcast	affricate	gerwen
footage	bud	session	phosphor	tactics
interview	shrub	q&a	connections	backstroke
flower	inflorescence	post-production	reduplication	gabler

Disambiguation Examples: ADD vs MIN

	Aligned: add		Aligned: min	
group	musical group	ethnic group	musical group	ethnic group
organization organisation company community corporation unit movement association society entity	group company band music movement community society corporation category association	group organization organisation community company movement society minority unit entity	group band troupe ensemble artist trio genre music duo supergroup	group community organization grouping sub-group faction ethnicity minority organisation tribe

Compositionality Detection (Weeds et al. 2017)

set of 90 compound nouns rated for compositionality / literality by humans (Reddy et al., 2011) on a scale of 0 to 5, e.g.,

- climate change: 5
- gravy train: 0.3
- cocktail dress: 3

ASSUMPTION: using an effective method of composition, similarity of composed representations with observed phrasal representations will correlate with compositionality of compound

Experimental methodology

1. Build elementary representations for phrases and for lexemes
2. Compose lexemes to infer compositional representation of phrase
3. Compute similarity of observed and compositional representations
4. Compute correlation between computed similarity judgements and human judgements of compositionality

Results -APTs

Alignment	Comp. Op	Spearman's ρ
Aligned	MIN	0.70
Aligned	SUM	0.72*
Unaligned	MIN	0.72*
Unaligned	SUM	0.75*
Hybrid	MIN	0.73*
Hybrid	SUM	**0.78***

These results obtained by composing representations where feature weights are standard PPMI scores.

Differences > 0.005 are significant at the 95% level

* significantly higher than Reddy et al 2011 result (0.714) at 95% level (using ukWaC corpus, proximity vectors and likelihood ratio as feature association)

Results – Word2Vec

Subsampling dilutes words which occur with a frequency greater than the threshold t

Embedding method	$t=10^{-3}$	$t=10^{-4}$	$t=10^{-5}$
cbow, 50d	0.73	0.65	0.62
cbow, 100d	**0.74**	0.65	0.64
cbow, 300d	0.70	0.70	0.67
skip-gram, 50d	0.59	0.64	0.62
skip-gram, 100d	0.62	0.64	0.64
skip-gram, 300d	0.63	0.64	0.68

Comments

- Unaligned APTs (analogous to conventional dependency reps) do very well at this task – likely due to large proportion of NN relations
- Best performance using combination of aligned and unaligned APTs
- ADD generally better than MIN (particularly for smaller corpus)
- Experiments with other corpora (e.g., wikipedia) show similar pattern of results and that the most notable factor in performance is size of corpus

Overcoming sparsity

Overcoming sparsity

size of vocabulary

Number of grammatical relations

APT depth

Overcoming sparsity

size of vocabulary

Number of grammatical relations

APT depth

Very sparse model can be a problem ….

- ➢ computationally
- ➢ theoretically.

Overcoming sparsity

size of vocabulary

Number of grammatical relations

APT depth

Very sparse model can be a problem ….
- ➢ computationally
- ➢ theoretically.

What does 0 mean?

Overcoming sparsity

Number of dimensions is $O(r^kV)$

size of vocabulary

Number of grammatical relations

APT depth

Very sparse model can be a problem ….

- ➢ computationally
- ➢ theoretically.

What does 0 mean?

Standard strategies to dealing with sparsity include

- ➢ dimensionality reduction
- ➢ smoothing

Distributional Inference (Kober et al. 2016)

Improve elementary representations using ***distributional inference*** (Dagan et al. 1994)

Add in to each elementary representation, co-occurrences which were observed with word's neighbours

1. Build sparse APT representations, M
2. For all w in M do:
    3. w' ← w ✖ α
    4. for all n in neighbours(M,w) do
        5. w' ← w' + n

Hyper-parameters include:

➢ neighbourhood retrieval function, weighting of neighbours and original distrib

➢ similarity measure used, feature weighting used in similarity calculations

Word similarity experiments

	Without DI	With DI
MEN	0.63	0.68
SimLex-999	0.30	0.32
WordSim-353 (rel)	0.55	0.61
WordSim-353 (sub)	0.75	0.76

Benchmark word similarity tasks which compare distributional similarity with human judgements of similarity

Neighbourhood retrieval function: static top 30
α: 30
Similarity measure: cosine
Feature weighting: Shifted PPMI (k=40) with context distribution smoothing (α=0.75)
(Levy et al. 2015)

Phrase Similarity Benchmarks

		AN	NN	VO	Avg
No DI	Union	0.45	0.43	0.37	0.42
	Intersect	0.38	0.44	0.36	0.39
With DI	Union	0.45	0.45	0.38	0.43
	Intersect	**0.50**	**0.49**	**0.43**	**0.47**
B&L 2012 *		0.48	0.50	0.35	0.44
Hashimoto 2014 **		0.52	0.46	0.45	0.48

* Blacoe and Lapata (2012): untyped VSM using multiplication as composition

** Hashimoto et al. (2014) : neural network based model (PAS-CLBLM with add)

Results on the phrase similarity benchmark task from Mitchell and Lapata (2010)

Conclusions

- the APT is a single structure which represent distributional semantics of lexemes, phrases and sentences
- definition of context is crucial: retention of higher-order grammatical structure enables syntax-sensitive composition
- APT composition captures mutual disambiguation (and generalisation)

Applications

- word sense induction
- semantic relation discovery
- sentence completion, parse reranking, language modelling
- paraphrase recognition, question answering

Further Work

- consider limitations in underlying grammar formalism
  - surface disparities in syntactic structure e.g., active vs passive
  - modifier scope (happiest blonde person = blonde happiest person)
- explore other grammar formalisms e.g., CCG
- investigate how to handle function words:- the dog vs a dog vs all dogs
- develop continuous model of syntax?
- combine with predictive approaches to learning word embeddings /dimensionality reduction

References 1

Baroni, Marco and Zamperelli, Roberto. 2010. Nouns are vectors, adjectives are matrices: representing adjective-noun constructions in semantic space. In *Proceedings of EMNLP*

Blaco, William and Lapata, Mirella. 2012. A comparison of vector-based representations for semantic composition. In *Proceedings of EMNLP*

Ferraresi, Adriano, Eros Zanchetta, Marco Baroni and Silvia Bernardini. 2008. Introducing and evaluating ukWaC, a very large web-derived corpus of English. In Proceedings of the WAC4 workshop at LREC.

Firth, J.R. 1957. A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*

Grefenstette, Ed, Dinum Georgiana, Zhang, Y-Z, Sadrzadeh, Mehrnoosh and Baroni, Marco. 2013. Multi-step regression learning for compositional distributional semantics. In *Proceedings of ICWS*.

Harris, Z. 1954. Distributional structure. In *Word*

Hashimoto, Kazuma, Pontus Stenetorp, Makoto Miwa and Yoshimasa Tsuruoka. 2014. Jointly learning word representations and composition functions using predicate-argument structures. In *Proceedings of EMNLP*

Kober, Thomas, Julie Weeds, Jeremy Reffin and David Weir. 2016. Improving sparse word representations with distributional inference for semantic composition. In *Proceedings of EMNLP*

References 2

Levy, Omar, Yoav Goldberg and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. In *Transactions of the Association for Computational Linguistics (TACL)*

Mitchell, Jeff and Lapata, Mirella. 2008. Vector-based Models of Semantic Composition. In *Proceedings of ACL*.

Mitchell, Jeff and Lapata, Mirella. 2010. Composition in Distributional Models of Semantics. In *Cognitive Science*.

Pennington, Jeffrey, Richard Socher and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representations. In *Proceedings of EMNLP*

Reddy, Siva, Diana McCarthy and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of IJCNLP*

Weir, David, Julie Weeds, Jeremy Reffin and Thomas Kober. 2016. Aligning packed dependency trees: a theory of compositional for distributional semantics. In *Compositional Linguistics*

Weeds, Julie, Thomas Kober, Jeremy Reffin and David Weir. 2017. When a red herring is not a red herring: using compositional methods to detect non-compositional phrases. In *Proceedings of EACL*