Inferring unobserved co-occurrence events in Anchored Packed Trees

Thomas Kober
TAG Lab, University of Sussex
t.kober@sussex.ac.uk

15th June 2017 (1497535200)

Outline

- Distributional Semantics & Distributional Composition
- Anchored Packed Trees (APTs)
- The issue of sparsity
- Distributional Inference & Offset Inference

Outline

- **Distributional Semantics & Distributional Composition**
- Anchored Packed Trees (APTs)
- The issue of sparsity
- Distributional Inference & Offset Inference

Distributional Semantics

Distributional Semantics

- Distributional Hypothesis

Distributional Semantics

- Distributional Hypothesis
- Similar words occur in similar contexts ("*You shall know...*")

Distributional Semantics

- Distributional Hypothesis
- Similar words occur in similar contexts ("*You shall know...*")
- Underlying idea can be traced back starting from Harris (1954) and Firth (1962), via Firth (1935), to Saussure (1916). Also Wittgenstein (1953) had thoughts along similar lines.

Distributional Semantics

- Distributional Hypothesis
- Similar words occur in similar contexts ("*You shall know...*")
- Underlying idea can be traced back starting from Harris (1954) and Firth (1962), via Firth (1935), to Saussure (1916). Also Wittgenstein (1953) had thoughts along similar lines.
- First to be interested in comparing words distributionally were probably Church and Hanks (1989) and Hindle (1990)

Distributional Composition

Distributional Composition

- Capturing complex semantic phenomena in word space

Distributional Composition

- Capturing complex semantic phenomena in word space
- Would be nice if there was a general mechanism to combine elementary representations into longer phrases

Distributional Composition

- Capturing complex semantic phenomena in word space
- Would be nice if there was a general mechanism to combine elementary representations into longer phrases
- Different ideas proposed:

Distributional Composition

- Capturing complex semantic phenomena in word space
- Would be nice if there was a general mechanism to combine elementary representations into longer phrases
- Different ideas proposed:
    - adding/multiplying (Mitchell and Lapata, 2008; 2010; Zanzotto et al., 2010; Guevara 2010; 2011)

Distributional Composition

- Capturing complex semantic phenomena in word space
- Would be nice if there was a general mechanism to combine elementary representations into longer phrases
- Different ideas proposed:
    - adding/multiplying (Mitchell and Lapata, 2008; 2010; Zanzotto et al., 2010; Guevara 2010; 2011)
    - Formal Semantics (Baroni and Zamparelli, 2010; Coecke et al., 2011)

Distributional Composition

- Capturing complex semantic phenomena in word space
- Would be nice if there was a general mechanism to combine elementary representations into longer phrases
- Different ideas proposed:
    - adding/multiplying (Mitchell and Lapata, 2008; 2010; Zanzotto et al., 2010; Guevara 2010; 2011)
    - Formal Semantics (Baroni and Zamparelli, 2010; Coecke et al., 2011)
    - Neural Networks (Socher et al., 2012; 2014; Kalchbrenner et al., 2014; Tai et al., 2015; 10000s more)

Distributional Composition

- Capturing complex semantic phenomena in word space
- Would be nice if there was a general mechanism to combine elementary representations into longer phrases
- Different ideas proposed:
    - adding/multiplying (Mitchell and Lapata, 2008; 2010; Zanzotto et al., 2010; Guevara 2010; 2011)
    - Formal Semantics (Baroni and Zamparelli, 2010; Coecke et al., 2011)
    - Neural Networks (Socher et al., 2012; 2014; Kalchbrenner et al., 2014; Tai et al., 2015; 10000s more)
    - Anchored Packed Trees (Weir et al., 2016)

Distributional Composition

Distributional Composition

- But what does it actually mean?

Distributional Composition

- But what does it actually mean?
- Distributional composition as contextualisation

Distributional Composition

- But what does it actually mean?
- Distributional composition as contextualisation
    - The meaning of a lexeme in a particular context

Distributional Composition

- But what does it actually mean?
- Distributional composition as contextualisation
    - The meaning of a lexeme in a particular context
    - Composition can recover sense specific information (Kober et al., 2017a): **bank** account *vs.* river **bank**

Outline

- Distributional Semantics & Distributional Composition
- Anchored Packed Trees (APTs)
- The issue of sparsity
- Distributional Inference & Offset Inference

Outline

- Distributional Semantics & Distributional Composition
- **Anchored Packed Trees (APTs)**
- The issue of sparsity
- Distributional Inference & Offset Inference

Anchored Packed Trees (APTs)

Anchored Packed Trees (APTs)

- They are basically just vectors

Anchored Packed Trees (APTs)

Anchored Packed Trees (APTs)

- *~~They are basically just vectors~~*

Anchored Packed Trees (APTs)

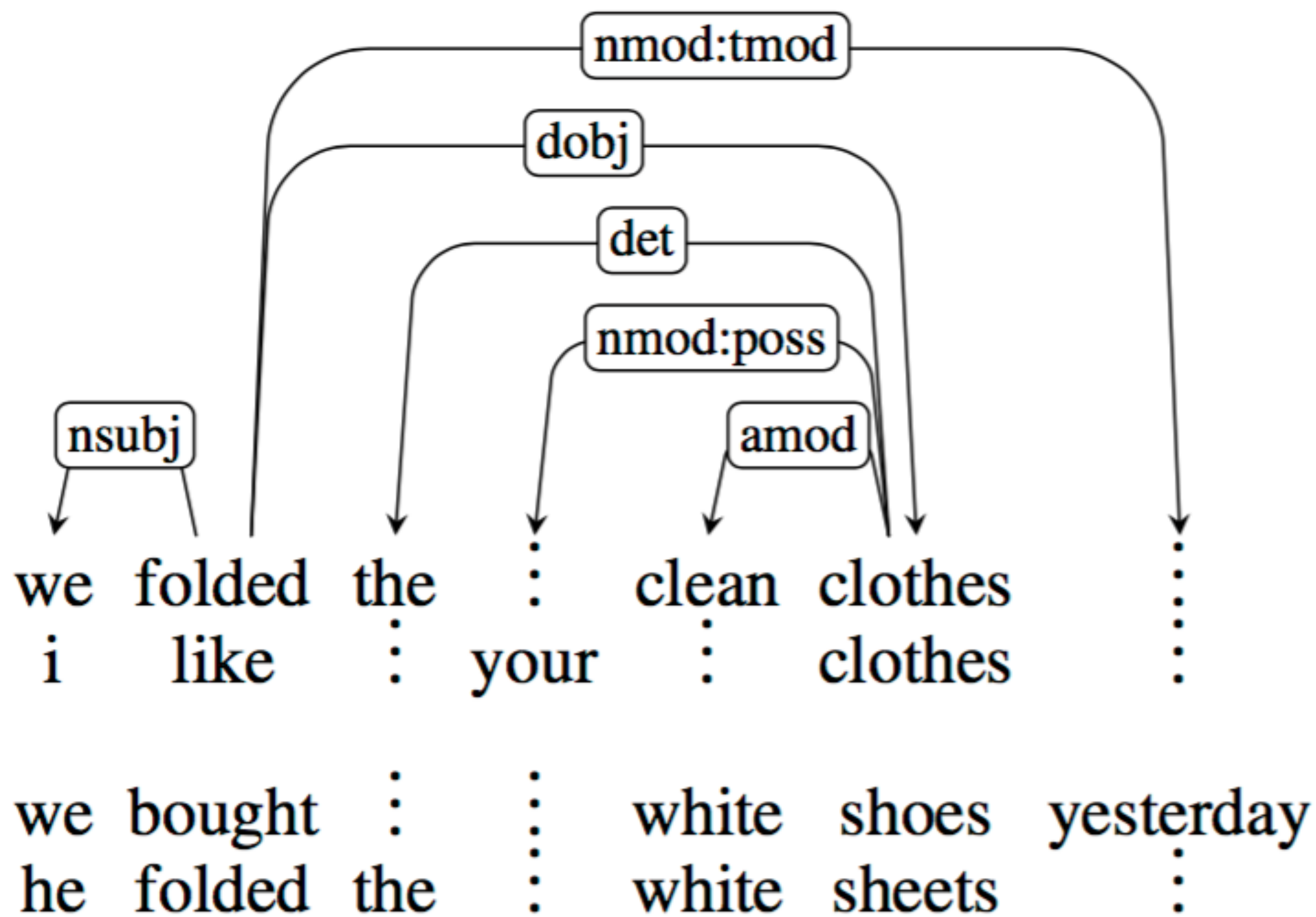
- ~~*They are basically just vectors*~~
- They are *****not***** vectors

Anchored Packed Trees (APTs)

- ~~*They are basically just vectors*~~
- They are *****not***** vectors
- (But sometimes it can be useful to vectorise them or think of them as vectors)

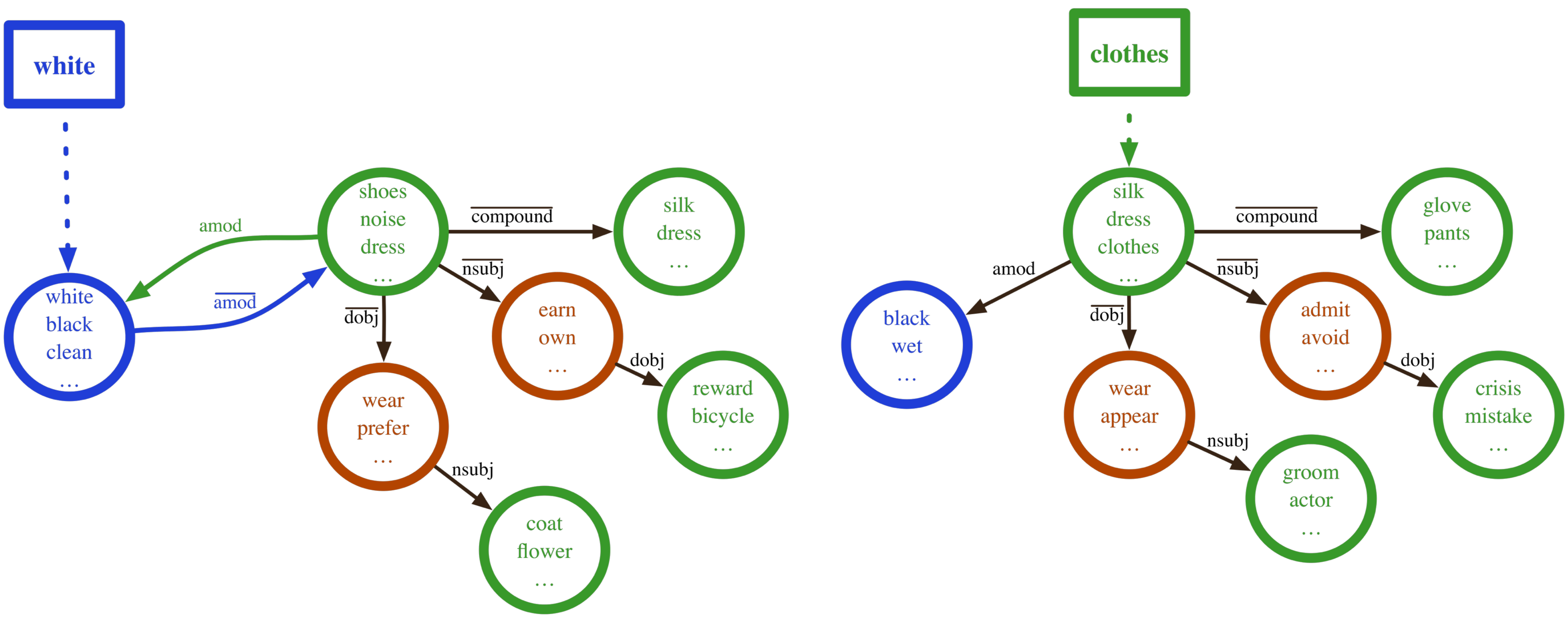
Anchored Packed Trees (APTs)

Anchored Packed Trees (APTs)

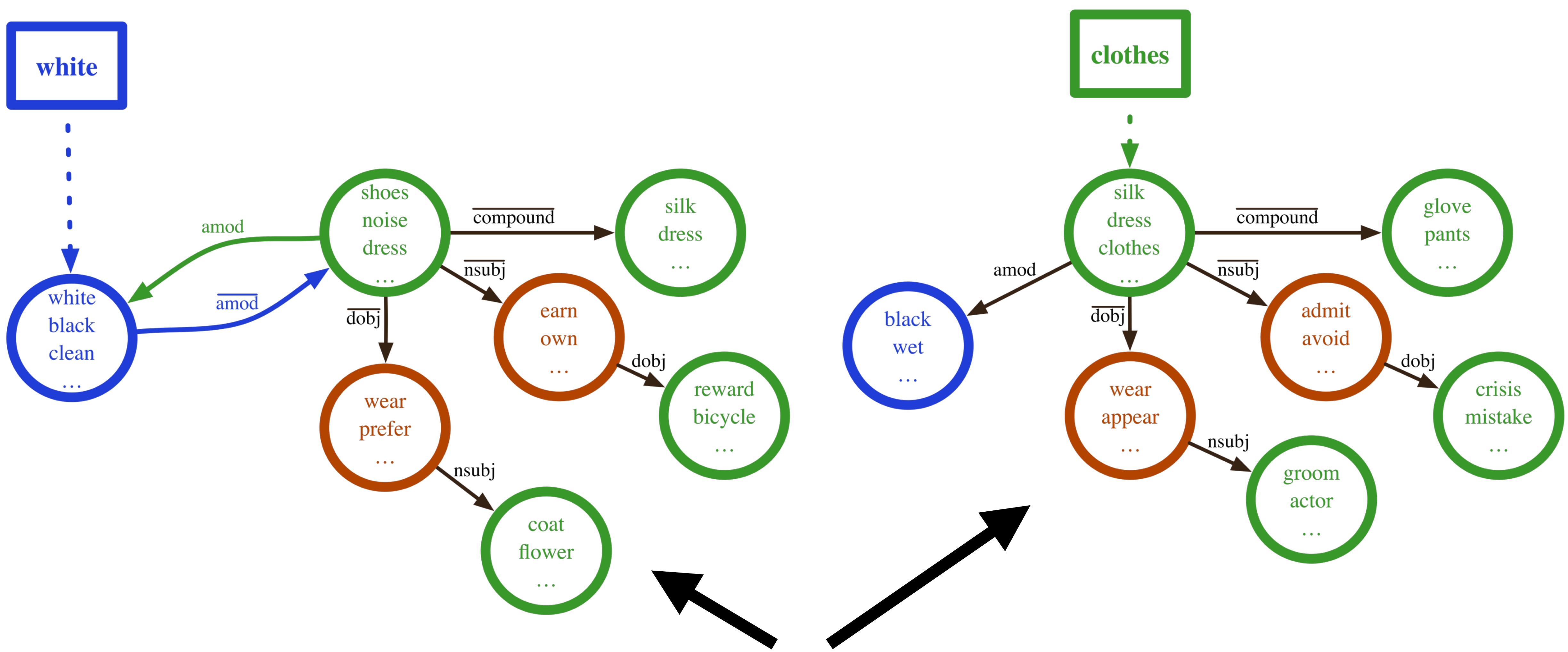


Anchored Packed Trees (APTs)

Anchored Packed Trees (APTs)



Anchored Packed Trees (APTs)



These two don't look like vectors, right?!

Characterising the APT distributional space

Characterising the APT distributional space

- Typed DSMs (such as APTs) give rise to a neighbourhood governed by co-hyponymy (e.g. dog - cat) and hypernymy (e.g. animal - dog)

Characterising the APT distributional space

- Typed DSMs (such as APTs) give rise to a neighbourhood governed by co-hyponymy (e.g. dog - cat) and hypernymy (e.g. animal - dog)

- Untyped DSMs (such as word2vec) give rise to a neighbourhood governed by relatedness (e.g. bee - honey; dog - kennel) or meronymy (e.g. dog - tail)

Characterising the APT distributional space

- Typed DSMs (such as APTs) give rise to a neighbourhood governed by co-hyponymy (e.g. dog - cat) and hypernymy (e.g. animal - dog)

- Untyped DSMs (such as word2vec) give rise to a neighbourhood governed by relatedness (e.g. bee - honey; dog - kennel) or meronymy (e.g. dog - tail)

- See e.g. Peirsman (2008), Baroni and Lenci (2011), Levy and Goldberg (2014) for work on typed vs. untyped DSMs

Characterising the APT distributional space

Characterising the APT distributional space

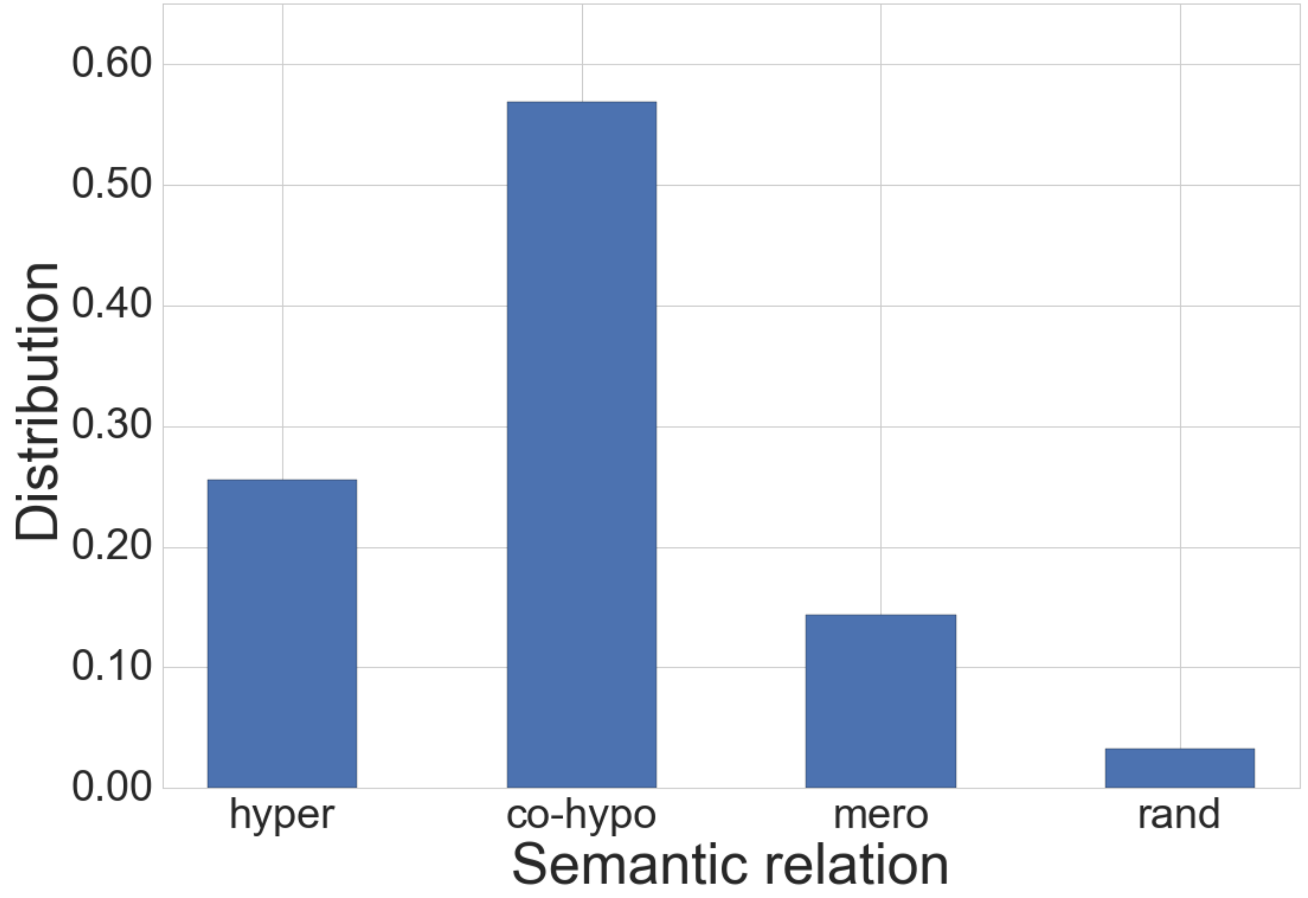
- Preliminary experiment using the BLESS dataset (Baroni and Lenci, 2011)

Characterising the APT distributional space

- Preliminary experiment using the BLESS dataset (Baroni and Lenci, 2011)
- Compare a target word to a hypernym, co-hyponym, meronym and a random word by cosine similarity, and tally up which relation is closest (e.g. given "dog", which of "animal", "cat", "tail" or "vector" is closest?)

Characterising the APT distributional space

- Preliminary experiment using the BLESS dataset (Baroni and Lenci, 2011)
- Compare a target word to a hypernym, co-hyponym, meronym and a random word by cosine similarity, and tally up which relation is closest (e.g. given "dog", which of "animal", "cat", "tail" or "vector" is closest?)



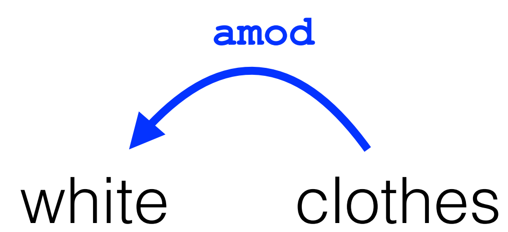
APT composition

APT composition

- Want to compose

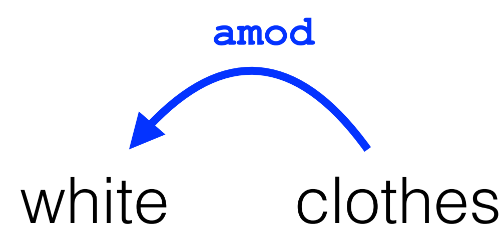
APT composition

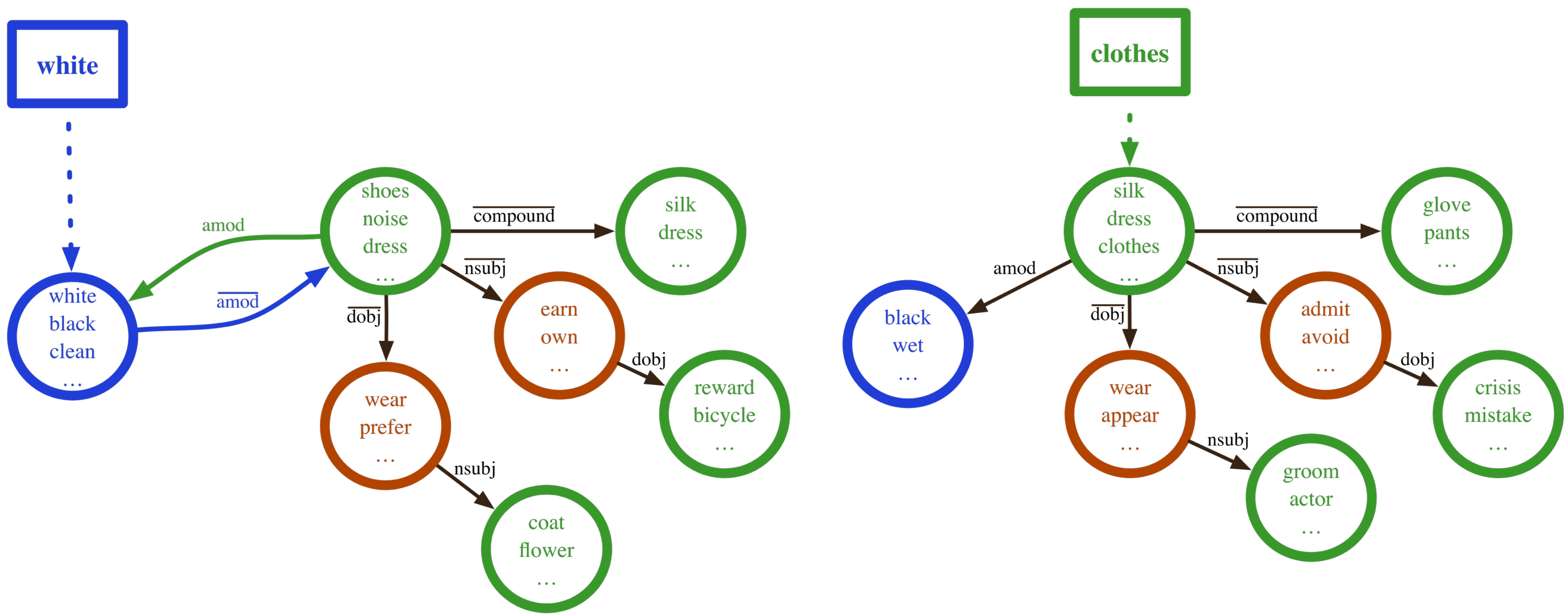
- Want to compose



APT composition

- Want to compose





APT composition

APT composition

- Composition in APTs is structured and driven by the given dependency tree

APT composition

- Composition in APTs is structured and driven by the given dependency tree
- Due to the structure, the feature spaces of adjectives and nouns are incompatible

APT composition

- Composition in APTs is structured and driven by the given dependency tree

- Due to the structure, the feature spaces of adjectives and nouns are incompatible

    - E.g. many paths for nouns start with **`amod`**, but this doesn't happen for verbs or adjectives

APT composition

- Composition in APTs is structured and driven by the given dependency tree
- Due to the structure, the feature spaces of adjectives and nouns are incompatible
    - E.g. many paths for nouns start with **`amod`**, but this doesn't happen for verbs or adjectives
- Need to align the representations first

APT composition

- Composition in APTs is structured and driven by the given dependency tree
- Due to the structure, the feature spaces of adjectives and nouns are incompatible
    - E.g. many paths for nouns start with **`amod`**, but this doesn't happen for verbs or adjectives
- Need to align the representations first
- Lets ***vectorise*** the feature space to make it more obvious!

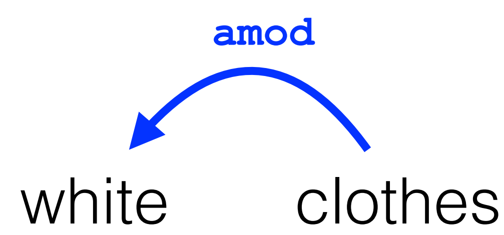
APT composition

APT composition

- Want to compose

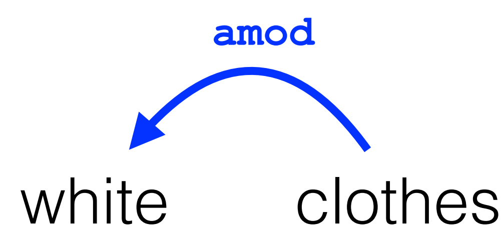
APT composition

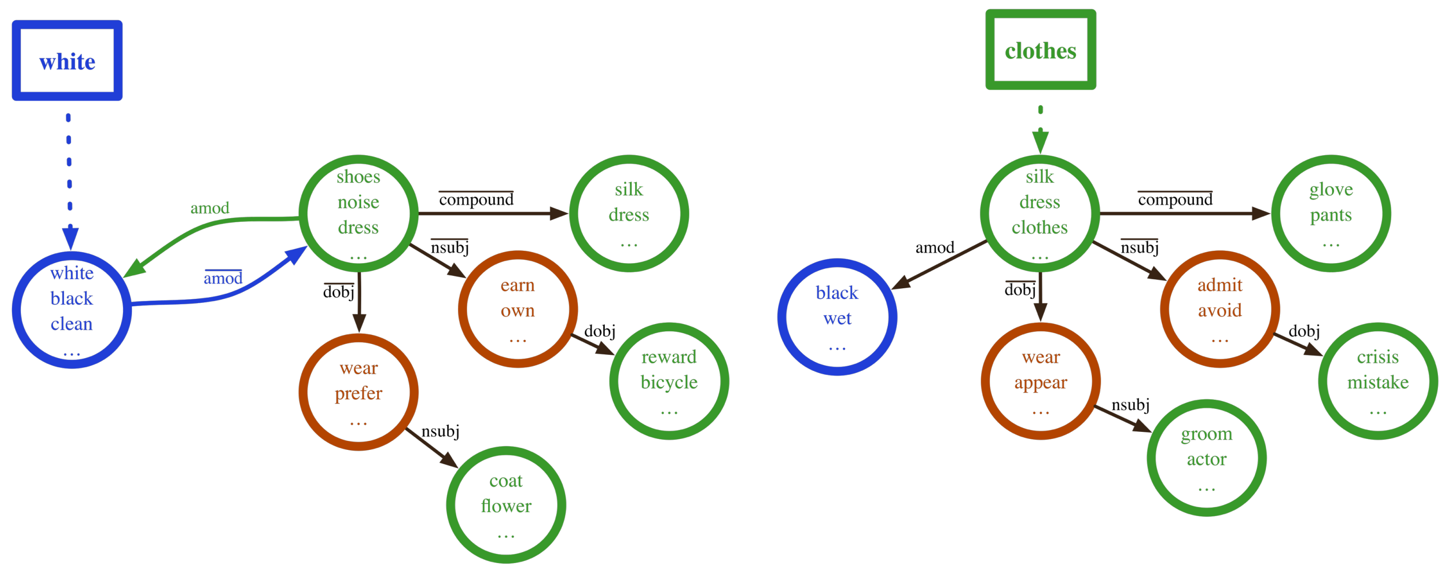
- Want to compose



APT composition

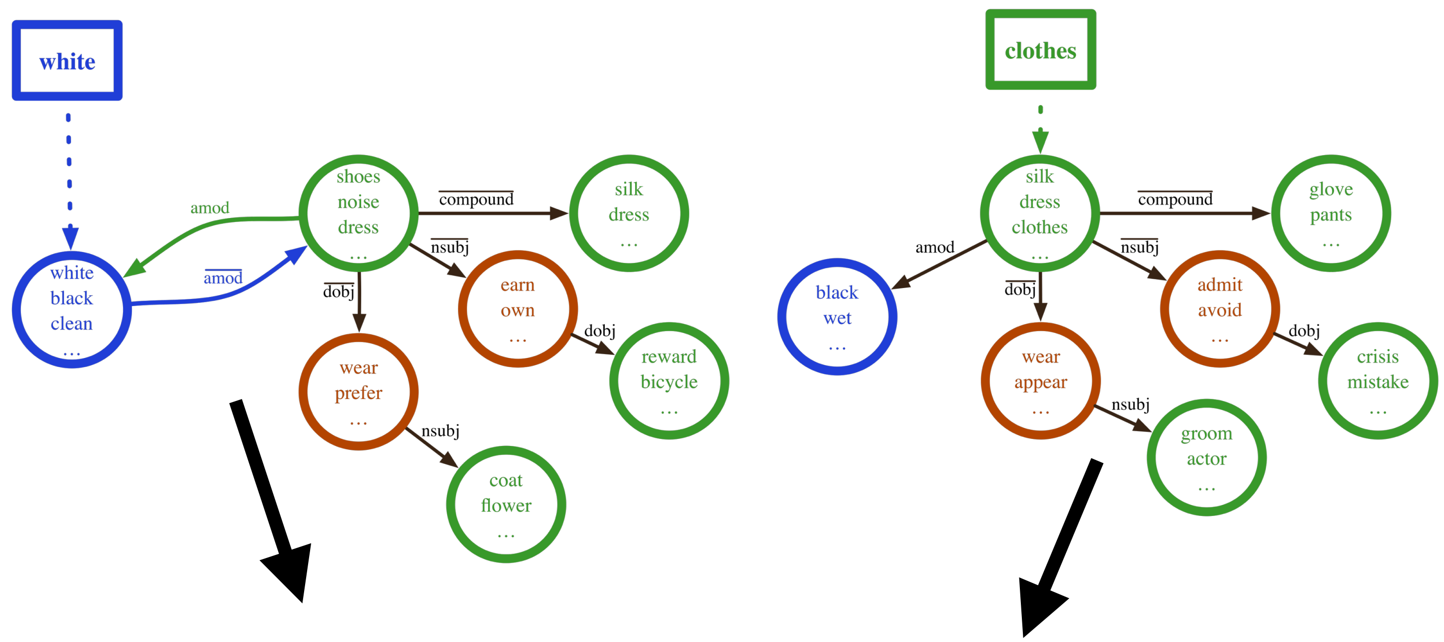
- Want to compose





APT composition

- Want to compose white $\xleftarrow{\textbf{amod}}$ clothes

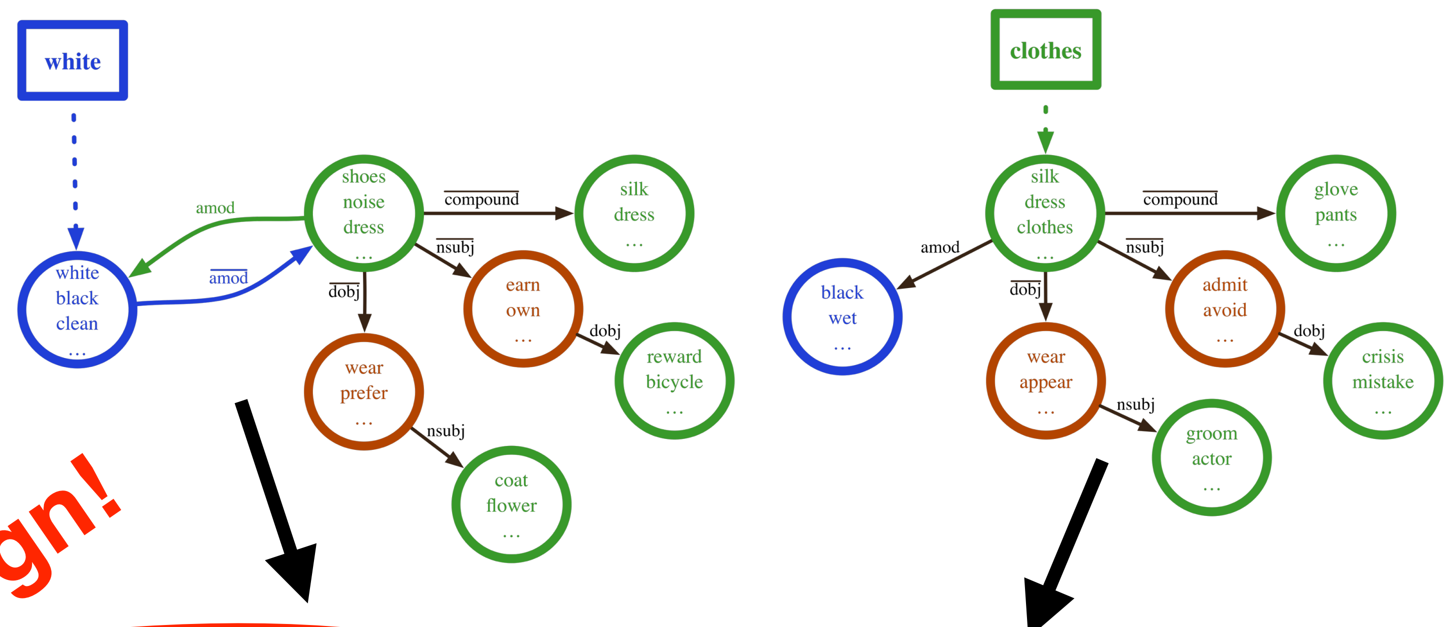


white	clothes
:clean	$\textbf{amod}$:wet
$\overline{\textbf{amod}}$:shoes	:dress
$\overline{\textbf{amod}}.\overline{\textbf{dobj}}$:wear	$\overline{\textbf{dobj}}$:wear
$\overline{\textbf{amod}}.\overline{\textbf{nsubj}}$:earn	$\overline{\textbf{nsubj}}$:admit

APT composition

- Want to compose

$$\text{white} \xleftarrow{\texttt{amod}} \text{clothes}$$



Paths don't align!

white	clothes
:clean	**amod**:wet
$\overline{\texttt{amod}}$:shoes	:dress
$\overline{\texttt{amod}}$ $\overline{\texttt{dobj}}$:wear	$\overline{\texttt{dobj}}$:wear
$\overline{\texttt{amod}}$ $\overline{\texttt{nsubj}}$:earn	$\overline{\texttt{nsubj}}$:admit

APT composition

APT composition

- Feature spaces between different parts of speech do not align

APT composition

- Feature spaces between different parts of speech do not align
- Need to align two representations in order to leverage their distributional commonalities

APT composition

- Feature spaces between different parts of speech do not align
- Need to align two representations in order to leverage their distributional commonalities
- Alignment can be achieved by "offsetting" one of the constituents

APT composition

- Feature spaces between different parts of speech do not align
- Need to align two representations in order to leverage their distributional commonalities
- Alignment can be achieved by "offsetting" one of the constituents
- Either offset *white* to make it a noun or offset *clothes* to make it an adjective

APT composition

- Feature spaces between different parts of speech do not align
- Need to align two representations in order to leverage their distributional commonalities
- Alignment can be achieved by "offsetting" one of the constituents
- Either offset *white* to make it a noun or offset *clothes* to make it an adjective
- We offset the dependent (so *white*) in a given dependency relation

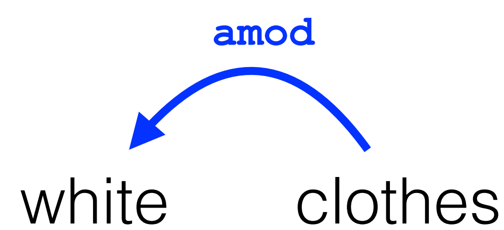
APT composition

APT composition

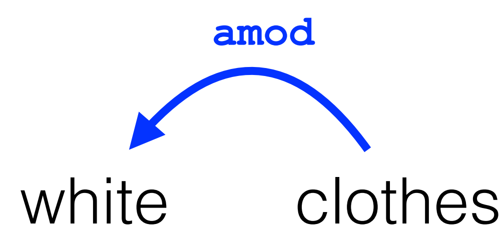
- Want to compose

APT composition

- Want to compose



APT composition

- Want to compose 

- Offset white by **`amod`** to make it a noun: white$^{\texttt{amod}}$

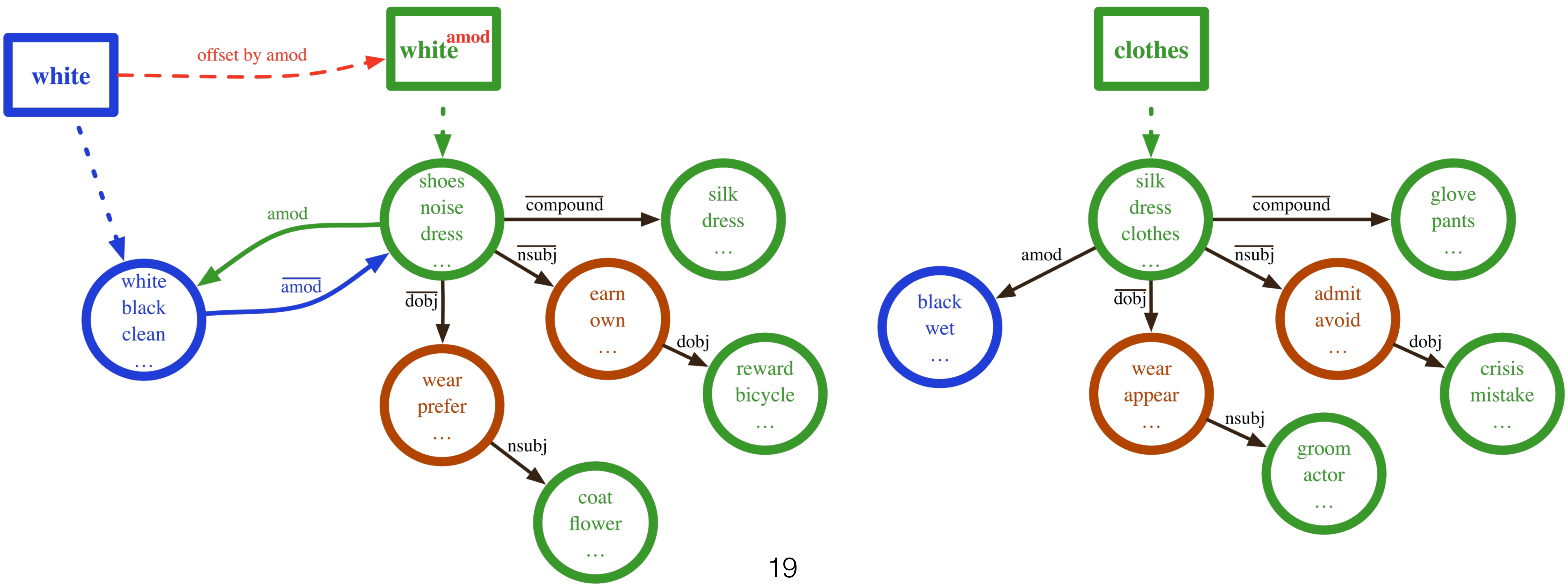
APT composition

- Want to compose

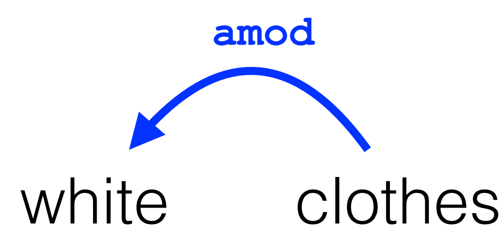
  amod

  white clothes

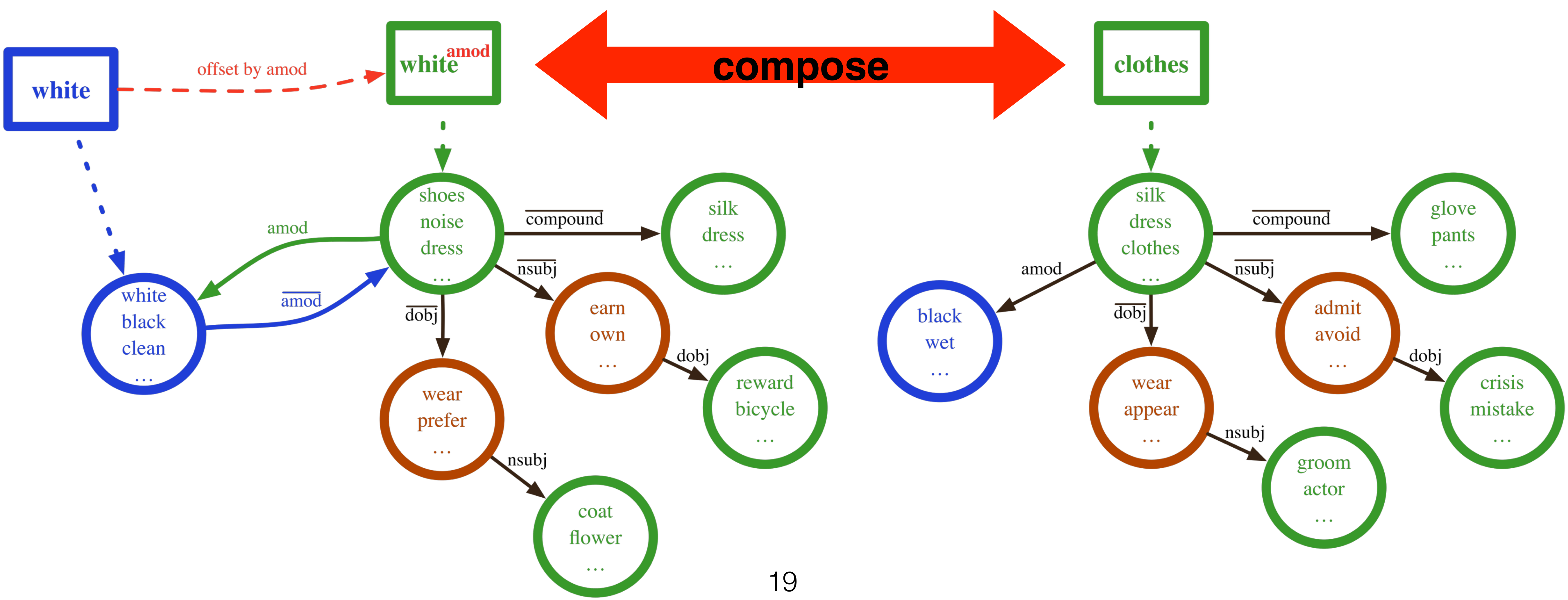
- Offset white by **amod** to make it a noun: $\text{white}^{\textbf{amod}}$



APT composition

- Want to compose 

- Offset white by **amod** to make it a noun: white$^{\mathbf{amod}}$

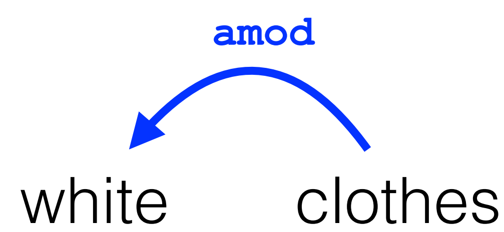


APT composition

APT composition

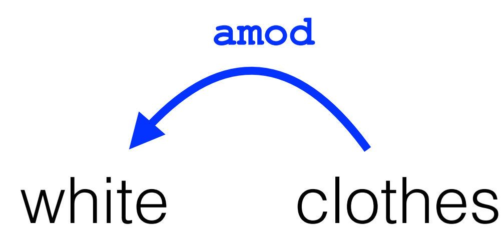
- Want to compose

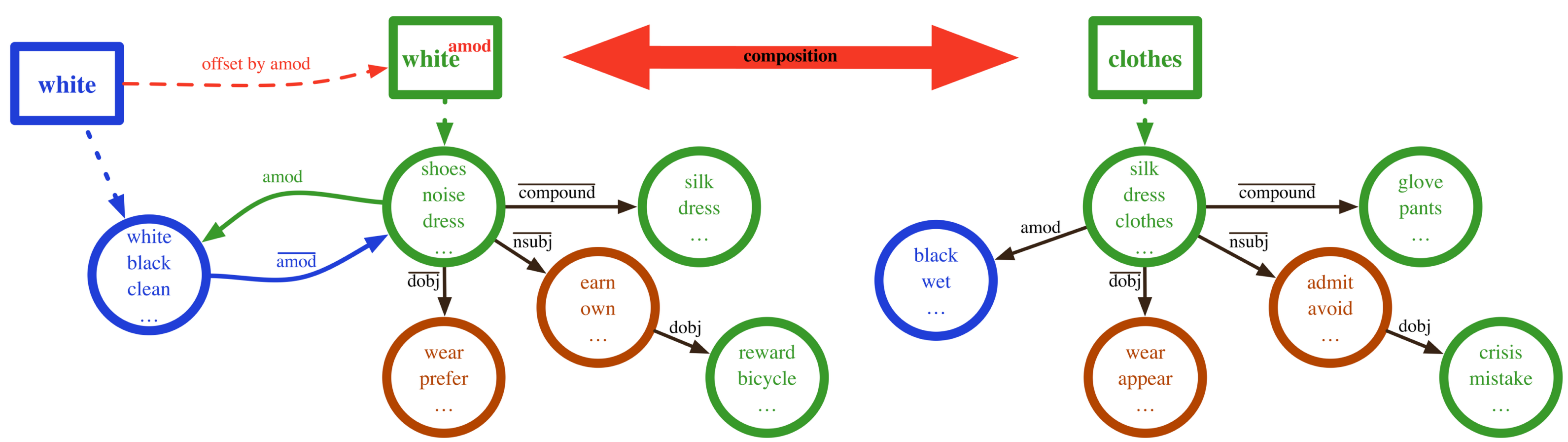
APT composition

- Want to compose 

APT composition

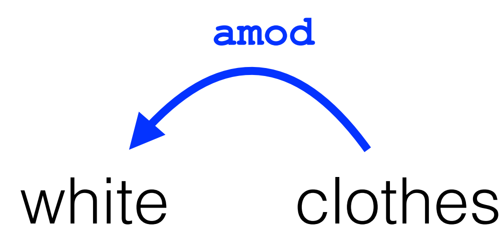
- Want to compose

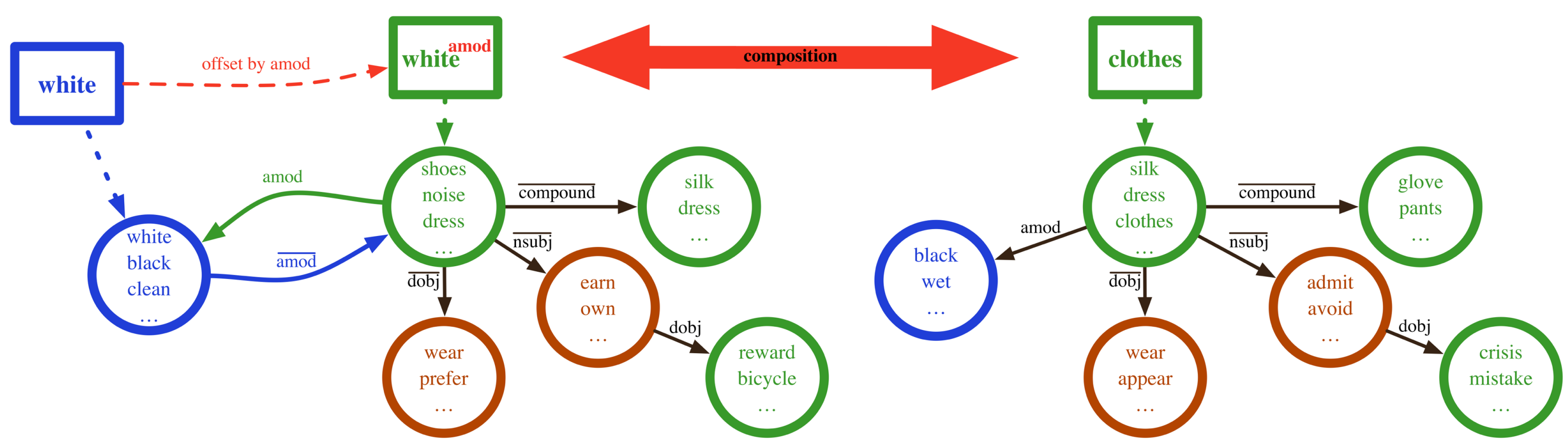




APT composition

- Want to compose





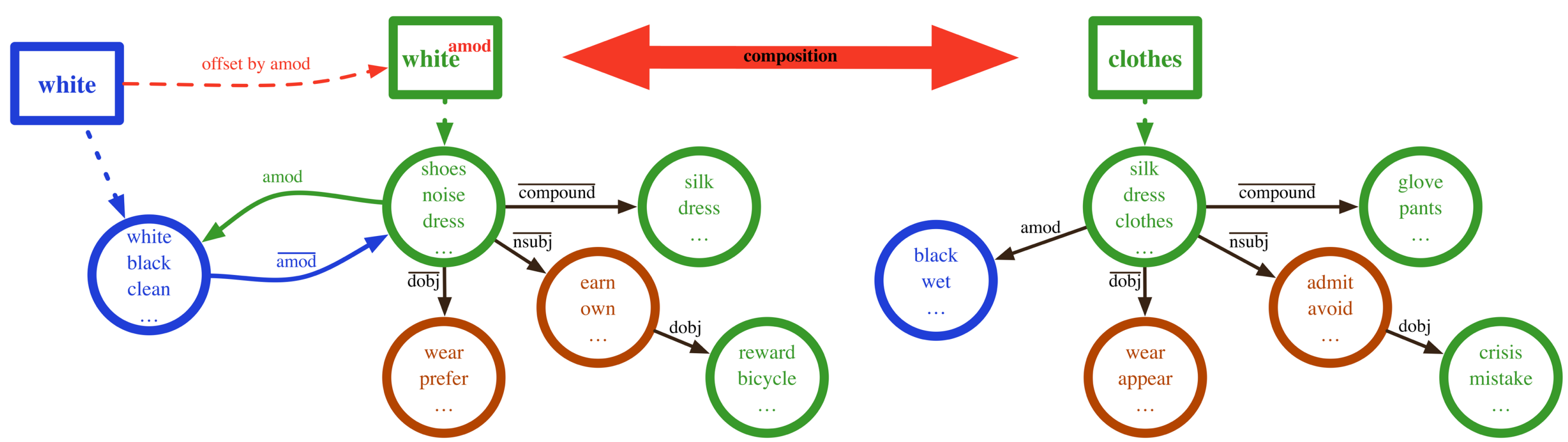
white	**white**$^{\mathtt{amod}}$	**clothes**
:clean	**amod**:clean	**amod**:wet
$\overline{\mathtt{amod}}$:shoes	:shoes	:dress
$\overline{\mathtt{amod}}.\overline{\mathtt{dobj}}$:wear	$\overline{\mathtt{dobj}}$:wear	$\overline{\mathtt{dobj}}$:wear
$\overline{\mathtt{amod}}.\overline{\mathtt{nsubj}}$:earn	$\overline{\mathtt{nsubj}}$:earn	$\overline{\mathtt{nsubj}}$:admit

APT composition

- Want to compose

amod

white clothes



Paths align! \o/

white	white$^{\textbf{amod}}$	clothes
:clean	**amod**:clean	**amod**:wet
$\overline{\textbf{amod}}$:shoes	:shoes	:dress
$\overline{\textbf{amod}}.\overline{\textbf{dobj}}$:wear	$\overline{\textbf{dobj}}$:wear	$\overline{\textbf{dobj}}$:wear
$\overline{\textbf{amod}}.\overline{\textbf{nsubj}}$:earn	$\overline{\textbf{nsubj}}$:earn	$\overline{\textbf{nsubj}}$:admit

Outline

- Distributional Semantics & Distributional Composition
- Anchored Packed Trees (APTs)
- The issue of sparsity
- Distributional Inference & Offset Inference

Outline

- Distributional Semantics & Distributional Composition
- Anchored Packed Trees (APTs)
- **The issue of sparsity**
- Distributional Inference & Offset Inference

The issue of sparsity

The issue of sparsity

- Model sparsity vs. Data sparsity

The issue of sparsity

- Model sparsity vs. Data sparsity
  - Model sparsity: discriminative and explicit (and therefore interpretable) representation

The issue of sparsity

- Model sparsity vs. Data sparsity
    - Model sparsity: discriminative and explicit (and therefore interpretable) representation
    - Data sparsity: Not observing all plausible co-occurrences in the given corpus

The issue of sparsity

- Model sparsity vs. Data sparsity
    - Model sparsity: discriminative and explicit (and therefore interpretable) representation
    - Data sparsity: Not observing all plausible co-occurrences in the given corpus
- APTs are a sparse model by design

The issue of sparsity

- Model sparsity vs. Data sparsity
  - Model sparsity: discriminative and explicit (and therefore interpretable) representation
  - Data sparsity: Not observing all plausible co-occurrences in the given corpus
- APTs are a sparse model by design
- This work addresses the data sparsity problem

The issue of ***data*** sparsity

The issue of ***data*** sparsity

- Caused by incomplete data in any collection

The issue of ***data*** sparsity

- Caused by incomplete data in any collection
- Especially problematic for intersective composition functions

The issue of ***data*** sparsity

- Caused by incomplete data in any collection
- Especially problematic for intersective composition functions
- If we compose intersectively a few times, we might end up with nothing left in the intersection

The issue of ***data*** sparsity

- Caused by incomplete data in any collection
- Especially problematic for intersective composition functions
- If we compose intersectively a few times, we might end up with nothing left in the intersection
- Composition by union avoids that, but lacks the discriminative power of composition by intersection

The issue of ***data*** sparsity

- Caused by incomplete data in any collection
- Especially problematic for intersective composition functions
- If we compose intersectively a few times, we might end up with nothing left in the intersection
- Composition by union avoids that, but lacks the discriminative power of composition by intersection
- If we apply composition by union a few times, we might end up with everything (or at least too much)

Why not just use dimensionality reduction?

Why not just use dimensionality reduction?

- Could do but...

Why not just use dimensionality reduction?

- Could do but...
- APT composition relies on offsetting to align two lexemes in a given dependency relation

Why not just use dimensionality reduction?

- Could do but...
- APT composition relies on offsetting to align two lexemes in a given dependency relation
- [As of now] There is no low-dimensional counterpart to achieve this precise operation

Why not just use dimensionality reduction?

- Could do but...
- APT composition relies on offsetting to align two lexemes in a given dependency relation
- [As of now] There is no low-dimensional counterpart to achieve this precise operation
- If the individual dimensions are not explicit, APTs degrade to just adding vectors

Outline

- Distributional Semantics & Distributional Composition
- Anchored Packed Trees (APTs)
- The issue of sparsity
- Distributional Inference & Offset Inference

Outline

- Distributional Semantics & Distributional Composition
- Anchored Packed Trees (APTs)
- The issue of sparsity
- **Distributional Inference & Offset Inference**

Enter Distributional Inference

Enter Distributional Inference

- Basic idea originates from language modelling in the speech processing community (Essen and Steinbiss, 1992)

Enter Distributional Inference

- Basic idea originates from language modelling in the speech processing community (Essen and Steinbiss, 1992)
- Smoothing bigrams with unseen words

Enter Distributional Inference

- Basic idea originates from language modelling in the speech processing community (Essen and Steinbiss, 1992)
- Smoothing bigrams with unseen words
- Picked up by Dagan et al. (1993) for WSD and Dagan et al. (1994) for LM

Enter Distributional Inference

- Basic idea originates from language modelling in the speech processing community (Essen and Steinbiss, 1992)
- Smoothing bigrams with unseen words
- Picked up by Dagan et al. (1993) for WSD and Dagan et al. (1994) for LM
- We in turn picked it up for distributional composition (Kober et al., 2016)

Enter Distributional Inference

- Basic idea originates from language modelling in the speech processing community (Essen and Steinbiss, 1992)
- Smoothing bigrams with unseen words
- Picked up by Dagan et al. (1993) for WSD and Dagan et al. (1994) for LM
- We in turn picked it up for distributional composition (Kober et al., 2016)
- Though there are traces of it in earlier work on composition (Kintsch, 2001) as well as modelling semantic relations (Turney, 2006) and modelling word meaning in context (Erk and Pado, 2010)

Distributional Inference

Distributional Inference

- For any word ***w*** in a distributional model ***M***

Distributional Inference

- For any word ***w*** in a distributional model ***M***
- Find the ***n*** nearest neighbours, ***w'***, of ***w***

Distributional Inference

- For any word ***w*** in a distributional model ***M***
- Find the ***n*** nearest neighbours, ***w'***, of ***w***
- Combine ***w*** and ***w'*** (and re-scale appropriately)

Distributional Inference

- For any word ***w*** in a distributional model ***M***
- Find the ***n*** nearest neighbours, ***w'***, of ***w***
- Combine ***w*** and ***w'*** (and re-scale appropriately)
- Profit!

Distributional Inference

- For any word ***w*** in a distributional model ***M***
- Find the ***n*** nearest neighbours, ***w'***, of ***w***
- Combine ***w*** and ***w'*** (and re-scale appropriately)
- Profit!
- ...or at least improved performance on some task

Distributional Inference

Distributional Inference

Dataset	APTs	APTs + DI	VSM	VSM + DI

Distributional Inference

Dataset	APTs	APTs + DI	VSM	VSM + DI
MEN	0,63	0.67 **(+0.04)**	0,71	0.71 **(+0.00)**

Distributional Inference

Dataset	**APTs**	**APTs + DI**	**VSM**	**VSM + DI**
MEN	0,63	0.67 **(+0.04)**	0,71	0.71 **(+0.00)**
SimLex	0,30	0.32 **(+0.02)**	0,30	0.29 **(-0.01)**

Distributional Inference

Dataset	APTs	APTs + DI	VSM	VSM + DI
MEN	0,63	0.67 **(+0.04)**	0,71	0.71 **(+0.00)**
SimLex	0,30	0.32 **(+0.02)**	0,30	0.29 **(-0.01)**
WS353 (rel)	0,55	0.62 **(+0.07)**	0,60	0.64 **(+0.04)**

Distributional Inference

Dataset	APTs	APTs + DI	VSM	VSM + DI
MEN	0,63	0.67 **(+0.04)**	0,71	0.71 **(+0.00)**
SimLex	0,30	0.32 **(+0.02)**	0,30	0.29 **(-0.01)**
WS353 (rel)	0,55	0.62 **(+0.07)**	0,60	0.64 **(+0.04)**
WS353 (sub)	0,75	0.78 **(+0.03)**	0,70	0.73 **(+0.03)**

Distributional Inference

Distributional Inference

ML10 Task	APTs	APTs + DI	VSM	VSM + DI

Distributional Inference

ML10 Task	APTs	APTs + DI	VSM	VSM + DI
AN	0,38	0.50 **(+0.12)**	0,42	0.46 **(+0.04)**

Distributional Inference

ML10 Task	APTs	APTs + DI	VSM	VSM + DI
AN	0,38	0.50 **(+0.12)**	0,42	0.46 **(+0.04)**
NN	0,44	0.49 **(+0.05)**	0,45	0.48 **(+0.03)**

Distributional Inference

ML10 Task	APTs	APTs + DI	VSM	VSM + DI
AN	0,38	0.50 **(+0.12)**	0,42	0.46 **(+0.04)**
NN	0,44	0.49 **(+0.05)**	0,45	0.48 **(+0.03)**
VO	0,36	0.43 **(+0.07)**	0,39	0.40 **(+0.01)**

Distributional Inference

ML10 Task	APTs	APTs + DI	VSM	VSM + DI
AN	0,38	0.50 **(+0.12)**	0,42	0.46 **(+0.04)**
NN	0,44	0.49 **(+0.05)**	0,45	0.48 **(+0.03)**
VO	0,36	0.43 **(+0.07)**	0,39	0.40 **(+0.01)**
Average	0,39	0.47 **(+0.08)**	0,42	0.45 **(+0.03)**

Distributional Inference

Distributional Inference

- Really important to get composition by intersection working

Distributional Inference

- Really important to get composition by intersection working
- Effect on composition by union was mildly positive for APTs (but not to such an extend as for composition by intersection)

Distributional Inference

- Really important to get composition by intersection working
- Effect on composition by union was mildly positive for APTs (but not to such an extend as for composition by intersection)
- Interesting relation between composition and inference - for window based VSMs, its the same operation

Distributional Inference

- Really important to get composition by intersection working

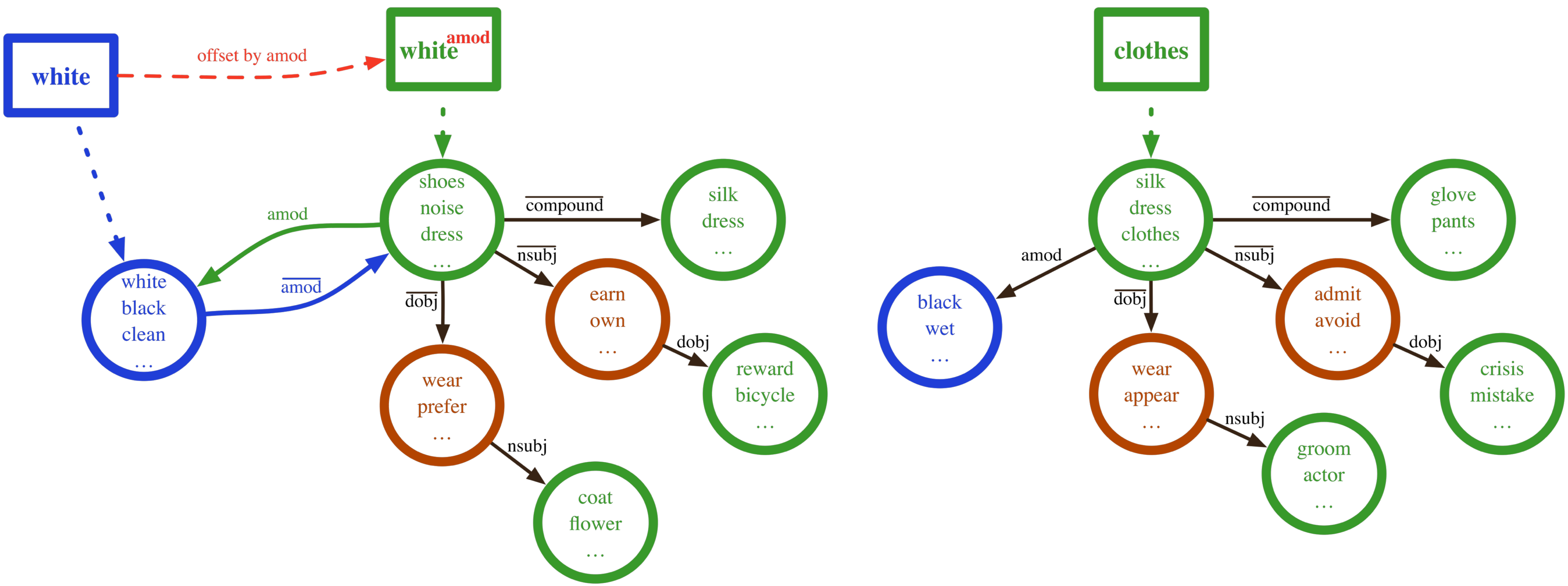
- Effect on composition by union was mildly positive for APTs (but not to such an extend as for composition by intersection)

- Interesting relation between composition and inference - for window based VSMs, its the same operation

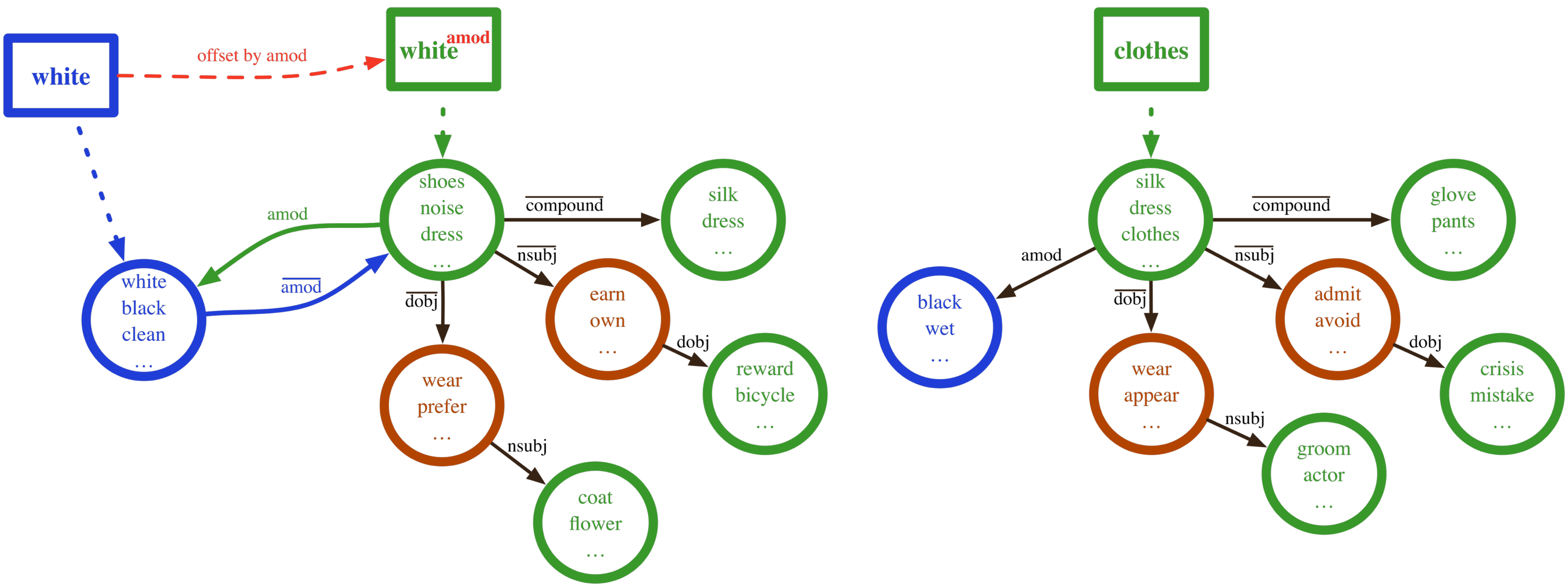
- Not yet for APTs, but if we leverage offsets...

Offset APTs

Offset APTs


white
offset by amod
white amod
clothes
shoes
noise
dress
…
compound
silk
dress
…
amod
amod
white
black
clean
…
dobj
nsubj
earn
own
…
dobj
reward
bicycle
…
wear
prefer
…
nsubj
coat
flower
…
silk
dress
clothes
…
compound
glove
pants
…
amod
black
wet
…
dobj
nsubj
admit
avoid
…
dobj
crisis
mistake
…
wear
appear
…
nsubj
groom
actor
…

Offset APTs



- Infer distributional information from *"other things that can be white"* (Kober et al., 2017b)

Offset APTs

Offset APTs

Offset Representation	Nearest Neighbours
$\text{ancient}^{\texttt{amod}}$	civilisation, mythology, tradition, ruin, monument
$\text{red}^{\texttt{amod}}$	$\text{blue}^{\texttt{amod}}$, $\text{black}^{\texttt{amod}}$, $\text{green}^{\texttt{amod}}$, $\text{dark}^{\texttt{amod}}$, onion
$\text{economic}^{\texttt{amod}}$	$\text{political}^{\texttt{amod}}$, $\text{societal}^{\texttt{amod}}$, cohabiting, economy, growth

Offset APTs

Offset Representation	Nearest Neighbours
ancient$^{\mathtt{amod}}$	civilisation, mythology, tradition, ruin, monument
red$^{\mathtt{amod}}$	blue$^{\mathtt{amod}}$, black$^{\mathtt{amod}}$, green$^{\mathtt{amod}}$, dark$^{\mathtt{amod}}$, onion
economic$^{\mathtt{amod}}$	political$^{\mathtt{amod}}$, societal$^{\mathtt{amod}}$, cohabiting, economy, growth
government$^{\overline{\mathtt{dobj}}}$	overthrow, party$^{\overline{\mathtt{dobj}}}$, authority$^{\overline{\mathtt{dobj}}}$, leader$^{\overline{\mathtt{dobj}}}$
problem$^{\overline{\mathtt{dobj}}}$	difficulty$^{\overline{\mathtt{dobj}}}$, solve, coded, issue$^{\overline{\mathtt{dobj}}}$, injury$^{\overline{\mathtt{dobj}}}$
law$^{\overline{\mathtt{dobj}}}$	violate, rule$^{\overline{\mathtt{dobj}}}$, enact, repeal, principle$^{\overline{\mathtt{dobj}}}$

Offset APTs

Offset Representation	Nearest Neighbours
ancient$^{\texttt{amod}}$	civilisation, mythology, tradition, ruin, monument
red$^{\texttt{amod}}$	blue$^{\texttt{amod}}$, black$^{\texttt{amod}}$, green$^{\texttt{amod}}$, dark$^{\texttt{amod}}$, onion
economic$^{\texttt{amod}}$	political$^{\texttt{amod}}$, societal$^{\texttt{amod}}$, cohabiting, economy, growth
government$^{\overline{\texttt{dobj}}}$	overthrow, party$^{\overline{\texttt{dobj}}}$, authority$^{\overline{\texttt{dobj}}}$, leader$^{\overline{\texttt{dobj}}}$
problem$^{\overline{\texttt{dobj}}}$	difficulty$^{\overline{\texttt{dobj}}}$, solve, coded, issue$^{\overline{\texttt{dobj}}}$, injury$^{\overline{\texttt{dobj}}}$
law$^{\overline{\texttt{dobj}}}$	violate, rule$^{\overline{\texttt{dobj}}}$, enact, repeal, principle$^{\overline{\texttt{dobj}}}$
researcher$^{\overline{\texttt{nsubj}}}$	physician$^{\texttt{nsubj}}$, writer$^{\texttt{nsubj}}$, theorize, thwart, theorise
mother$^{\overline{\texttt{nsubj}}}$	wife$^{\overline{\texttt{nsubj}}}$, father$^{\overline{\texttt{nsubj}}}$, parent$^{\overline{\texttt{nsubj}}}$, woman$^{\overline{\texttt{nsubj}}}$
law$^{\overline{\texttt{nsubj}}}$	rule$^{\overline{\texttt{nsubj}}}$, principle$^{\overline{\texttt{nsubj}}}$, policy$^{\overline{\texttt{nsubj}}}$, criminalize

Offset Inference

Offset Inference

- Generalises the Distributional Inference algorithm, which falls out as a special case

Offset Inference

- Generalises the Distributional Inference algorithm, which falls out as a special case
- For any word ***w*** in the APT space ***M***

Offset Inference

- Generalises the Distributional Inference algorithm, which falls out as a special case
- For any word ***w*** in the APT space ***M***
- Offset ***w*** by some dependency path ***p*** to get ***w'***

Offset Inference

- Generalises the Distributional Inference algorithm, which falls out as a special case
- For any word ***w*** in the APT space ***M***
- Offset ***w*** by some dependency path ***p*** to get ***w'***
- Find the ***n*** nearest neighbours, ***w''***, of ***w'***

Offset Inference

- Generalises the Distributional Inference algorithm, which falls out as a special case
- For any word ***w*** in the APT space ***M***
- Offset ***w*** by some dependency path ***p*** to get ***w'***
- Find the ***n*** nearest neighbours, ***w''***, of ***w'***
- Combine ***w'*** and ***w''*** (and re-scale appropriately)

Offset Inference

- Generalises the Distributional Inference algorithm, which falls out as a special case
- For any word ***w*** in the APT space ***M***
- Offset ***w*** by some dependency path ***p*** to get ***w'***
- Find the ***n*** nearest neighbours, ***w''***, of ***w'***
- Combine ***w'*** and ***w''*** (and re-scale appropriately)
- If ***p*** == ***ε***, the original DI algorithm is recovered (Kober et al., 2017b)

Offset Inference

Offset Inference

Dataset	APTs	APTs + DI	APTs + OI

Offset Inference

Dataset	**APTs**	**APTs + DI**	**APTs + OI**
ML10 - AN	0,35	0,48	0.49 (**+0.01**)

Offset Inference

Dataset	APTs	APTs + DI	APTs + OI
ML10 - AN	0,35	0,48	0.49 (**+0.01**)
ML10 - NN	0,50	0,51	0.52 (**+0.01**)

Offset Inference

Dataset	APTs	APTs + DI	APTs + OI
ML10 - AN	0,35	0,48	0.49 (**+0.01**)
ML10 - NN	0,50	0,51	0.52 (**+0.01**)
ML10 - VO	0,39	0,43	0.44 (**+0.01**)

Offset Inference

Dataset	APTs	APTs + DI	APTs + OI
ML10 - AN	0,35	0,48	0.49 (**+0.01**)
ML10 - NN	0,50	0,51	0.52 (**+0.01**)
ML10 - VO	0,39	0,43	0.44 (**+0.01**)
ML10 - Average	0,41	0,47	0.48* (**+0.01**)

Offset Inference

Dataset	APTs	APTs + DI	APTs + OI
ML10 - AN	0,35	0,48	0.49 (**+0.01**)
ML10 - NN	0,50	0,51	0.52 (**+0.01**)
ML10 - VO	0,39	0,43	0.44 (**+0.01**)
ML10 - Average	0,41	0,47	0.48* (**+0.01**)
ML08	0,22	0,29	0.31* (**+0.02**)

Offset Inference

Offset Inference

- Improvements are not striking in magnitude, but...

Offset Inference

- Improvements are not striking in magnitude, but...
- Consistent and statistically significant improvements

Offset Inference

- Improvements are not striking in magnitude, but...
- Consistent and statistically significant improvements
- Powerful concept, can travel the APT structure inferring unobserved co-occurrences at different nodes

Offset Inference

- Improvements are not striking in magnitude, but...
- Consistent and statistically significant improvements
- Powerful concept, can travel the APT structure inferring unobserved co-occurrences at different nodes
- Could also be done offline, but so far this didn't yield any improvements

Offset Inference

- Improvements are not striking in magnitude, but...
- Consistent and statistically significant improvements
- Powerful concept, can travel the APT structure inferring unobserved co-occurrences at different nodes
- Could also be done offline, but so far this didn't yield any improvements
- Realised by the same mechanism as composition

Offset Inference

- Improvements are not striking in magnitude, but...
- Consistent and statistically significant improvements
- Powerful concept, can travel the APT structure inferring unobserved co-occurrences at different nodes
- Could also be done offline, but so far this didn't yield any improvements
- Realised by the same mechanism as composition
    - An offset followed by a merge

Relation to Distributional Composition

Relation to Distributional Composition

- Works complementary with an intersective composition function

Relation to Distributional Composition

- Works complementary with an intersective composition function
- Co-occurrence embellishment and filtering

Relation to Distributional Composition

- Works complementary with an intersective composition function
- Co-occurrence embellishment and filtering
    - Distributional Inference embellishes an APT representation, at the cost of introducing some noise

Relation to Distributional Composition

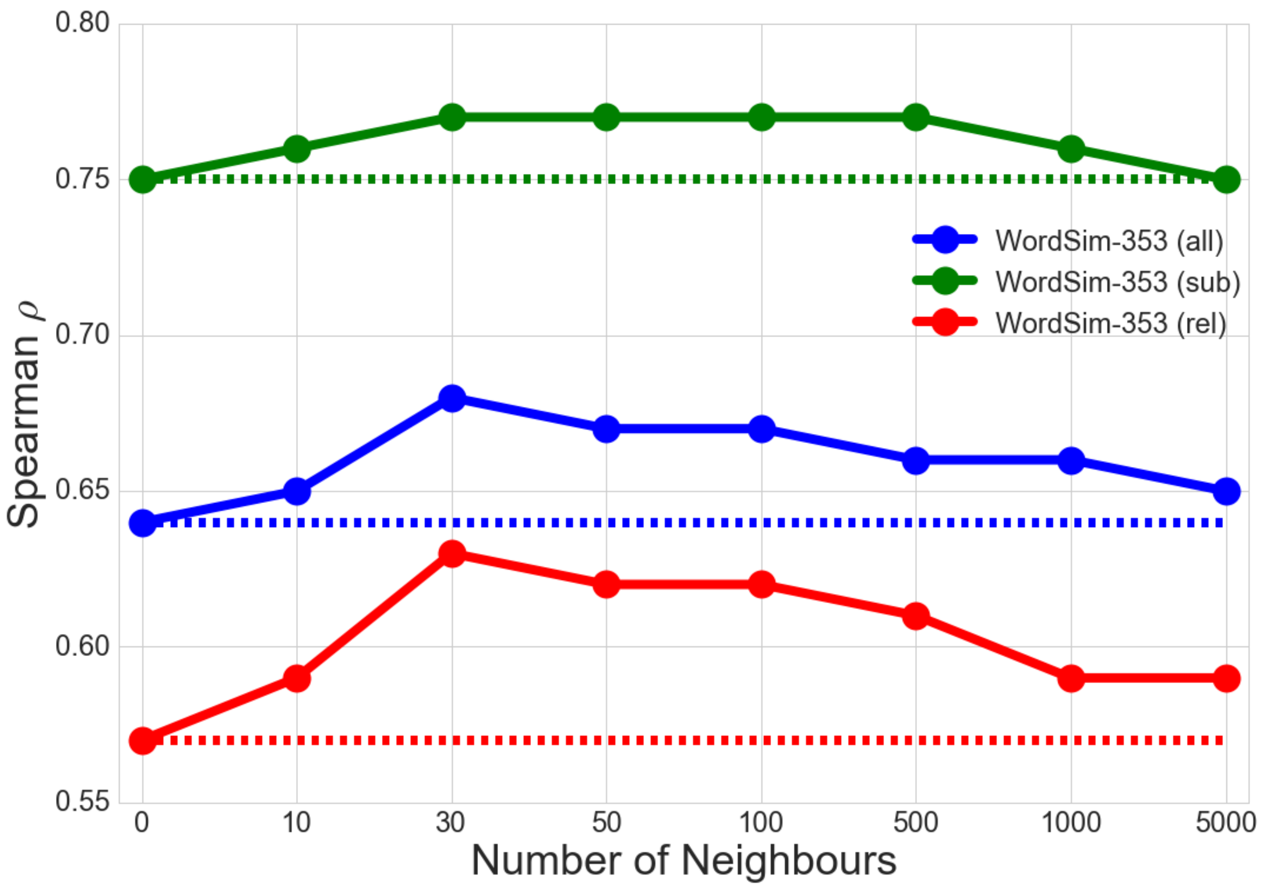
- Works complementary with an intersective composition function
- Co-occurrence embellishment and filtering
  - Distributional Inference embellishes an APT representation, at the cost of introducing some noise
  - Distributional Composition filters an APT representation, at the cost of removing some plausible information (data sparsity!)

Relation to Distributional Composition

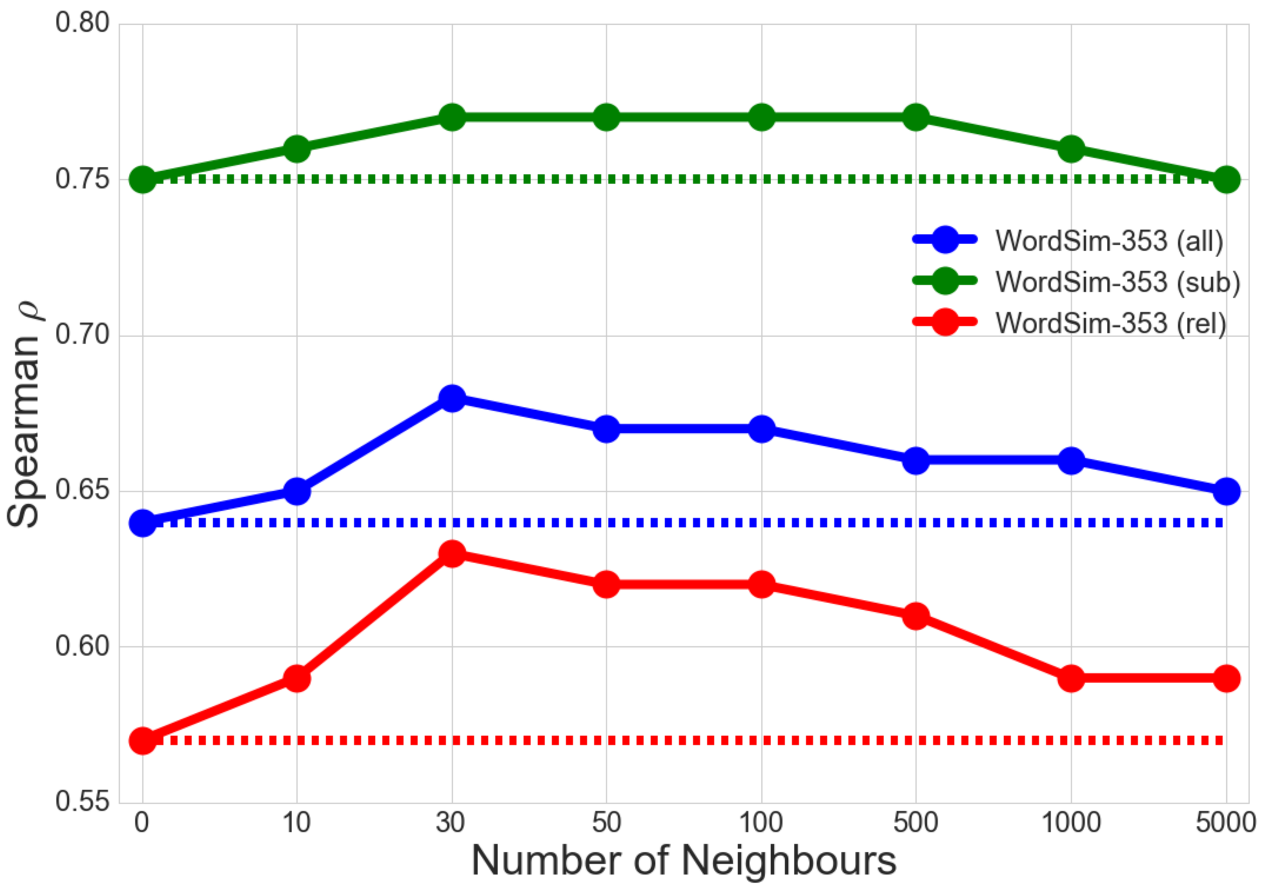
- Works complementary with an intersective composition function
- Co-occurrence embellishment and filtering
    - Distributional Inference embellishes an APT representation, at the cost of introducing some noise
    - Distributional Composition filters an APT representation, at the cost of removing some plausible information (data sparsity!)
- Potential to scale to longer phrases with an intersective composition function before running out of features

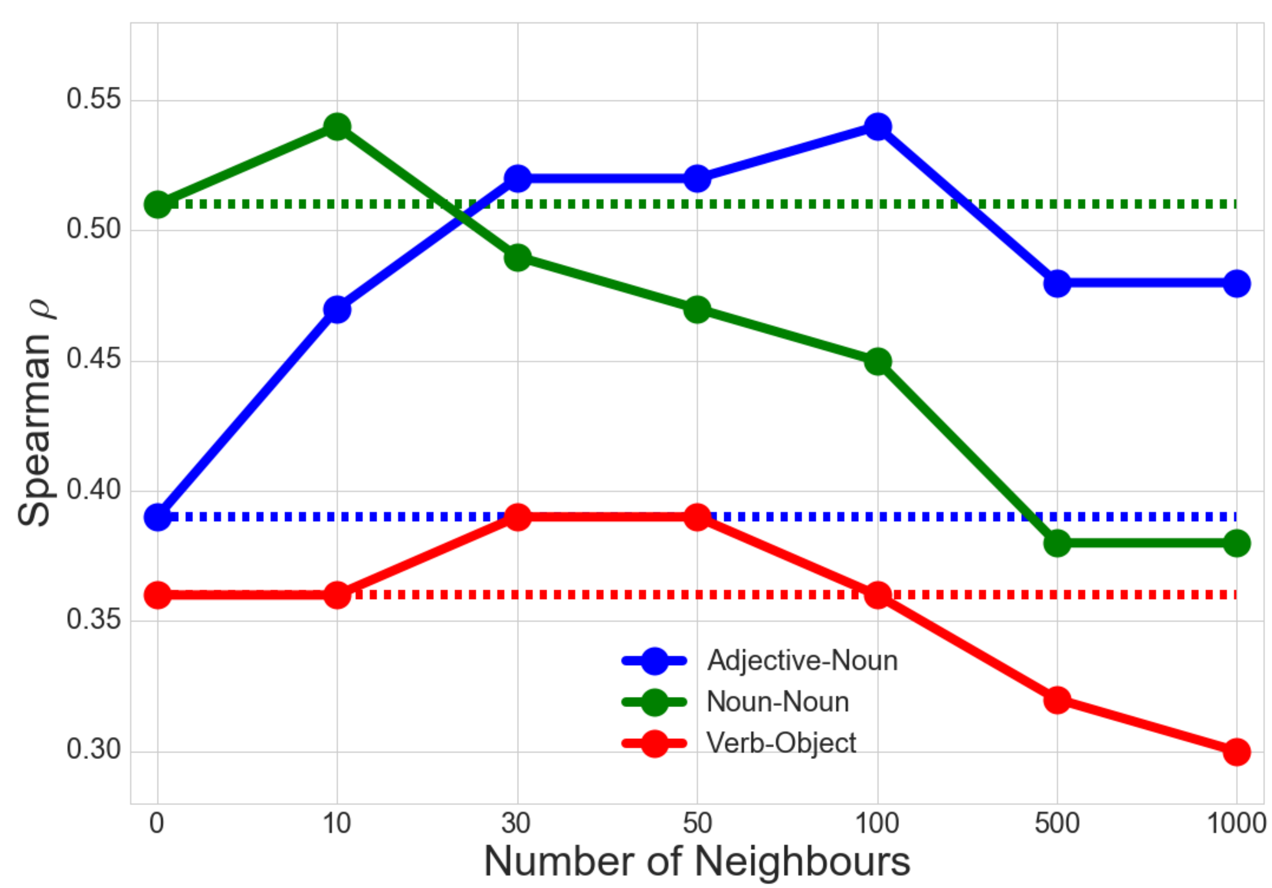
Its more than just smoothing

Its more than just smoothing

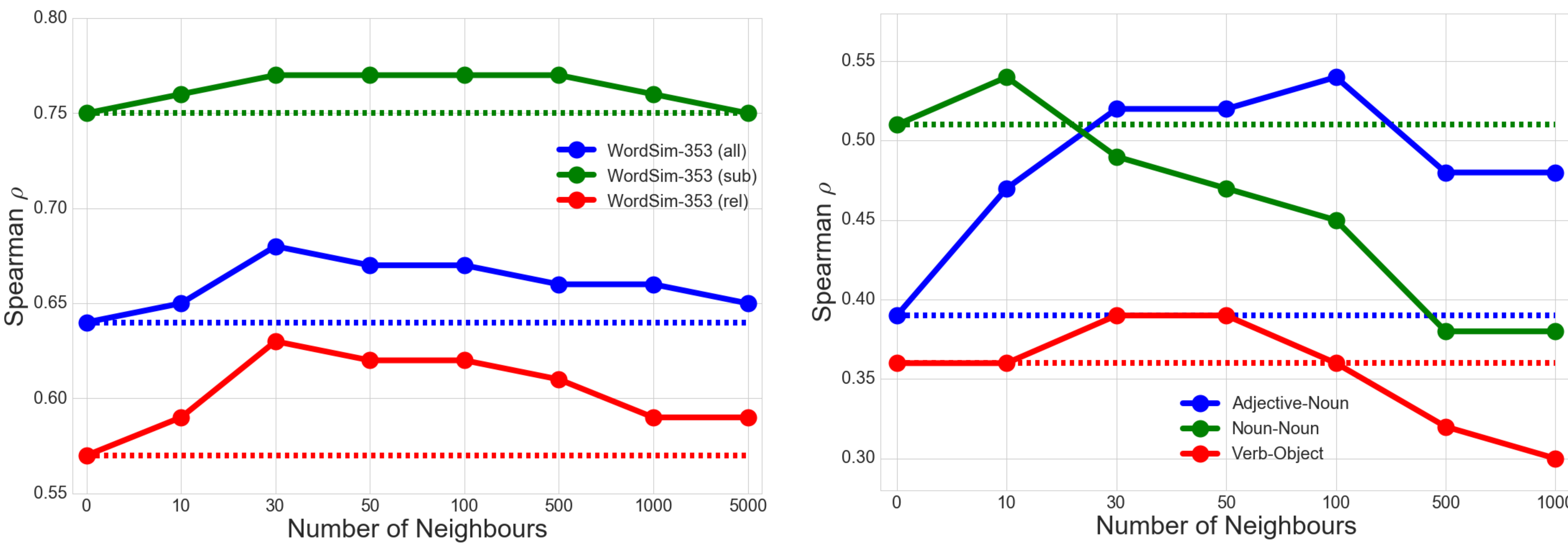


Its more than just smoothing





Its more than just smoothing



- The number of neighbours is the only hyperparameter that needs tuning

Limitations of Distributional Inference

Limitations of Distributional Inference

- But getting the number of neighbours right can be crucial

Limitations of Distributional Inference

- But getting the number of neighbours right can be crucial
- Too many neighbours leads to an overflow of the representations with noise

Limitations of Distributional Inference

- But getting the number of neighbours right can be crucial
- Too many neighbours leads to an overflow of the representations with noise
- Without a "post-processing step" (such as composition) to clean up the representations, this could lead to a mess

Limitations of Distributional Inference

- But getting the number of neighbours right can be crucial
- Too many neighbours leads to an overflow of the representations with noise
- Without a "post-processing step" (such as composition) to clean up the representations, this could lead to a mess
- Still difficult to scale beyond short sentences with an intersective composition function

Thats it!

Thats it!

Thats it!

Or ask some ques…cake…did somebody mention cake?!

Thats it!

Or ask some ques…cake…did somebody mention cake?!

(You can also email me - t.kober@sussex.ac.uk - and I might even reply!)

References (1)

- Marco Baroni and Alessandro Lenci. 2011. How we BLESSed Semantic Evaluation. In Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics, 1-10
- Marco Baroni and Roberto Zamparelli. 2010. Nouns are Vectors, Adjectives are Matrices: Representing Adjective-Noun Constructions in Semantic Space. In Proceedings of EMNLP, 1183-1193
- Kenneth Ward Church and Patrick Hanks. 1989. Word Association, Mutual Information, and Lexicography. In Proceedings of ACL, 76-83
- Bob Coecke, Mehrnoosh Sadrzadeh and Stephen Clark. 2011. Mathematical Foundations for a Compositional Distributed Model of Meaning. Linguistic Analysis, 36(1-4): 345-384
- Ido Dagan, Shaul Marcus and Shaul Markovitch. 1993. Contextual Word Similarity and Estimation from Sparse Data. In Proceedings of ACL, 164-171
- Ido Dagan, Fernando Pereira and Lillian Lee. 1994. Similarity-Based Estimation of Word Cooccurrence Probabilities. In Proceedings of ACL, 272-278
- Katrin Erk and Sebastian Pado. 2010. Exemplar-Based Models for Word Meaning in Context. In Proceedings of ACL, 92-97
- Ute Essen and Volker Steinbiss. 1992. Co-occurrence Smoothing for Stochastic Language Modeling. In Proceedings of ICASSP, 161-164
- John Rupert Firth. 1935. The Technique of Semantics. Transactions of the Philological Society 34(1):36-73
- John Rupert Firth. 1962. A Synopsis of Linguistic Theory. Selected Papers of JR Firth 1952-1959, 168-205

References (2)

- Emiliano Guevara. 2010. A Regression Model of Adjective-Noun Compositionality in Distributional Semantics. In Proceedings of the 2010 GEMS Workshop on GEometrical Models of Natural Language Semantics, 33-37
- Emiliano Guevara. 2011. Computing Semantic Compositionality in Distributional Semantics. In Proceedings of IWCS, 135-144
- Zellig Harris. 1954. Distributional Structure. Word 10:146-162
- Donald Hindle. 1990. Noun Classification from Predicate-Argument Structures. In Proceedings of ACL, 268-275
- Nal Kalchbrenner, Edward Grefenstette and Phil Blunsom. 2014. A Convolutional Neural Network for Modelling Sentences. In Proceedings of ACL, 655-665
- Walter Kintsch. 2001. Predication. Cognitive Science 25
- Thomas Kober, Julie Weeds, Jeremy Reffin and David Weir. 2016. Improving Sparse Word Representations with Distributional Inference for Semantic Composition. In Proceedings of EMNLP, 1691-1702
- Thomas Kober, Julie Weeds, John Wilkie, Jeremy Reffin and David Weir. 2017. One Representation per Word - Does it make Sense for Composition? In Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications, 79-90
- Thomas Kober, Julie Weeds, Jeremy Reffin and David Weir. 2017. Improving Semantic Composition with Offset Inference. In Proceedings of ACL
- Omer Levy and Yoav Goldberg. 2014. Dependency-Based Word Embeddings. In Proceedings of ACL, 302-308
- Jeff Mitchell and Mirella Lapata. 2008. Vector-based Models of Semantic Composition. In Proceedings of ACL, 236-244
- Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. Cognitive Science, 34(8):1388–1429

References (3)

- Yves Peirsman. 2008. Word space models of semantic similarity and relatedness. In Proceedings of ESSLLI, 143-152

- Ferdinand de Saussure. 1916. Cours de linguistique générale

- Richard Socher, Brody Huval, Christopher Manning and Andrew Ng. 2012. Semantic Compositionality through Recursive Matrix-Vector Spaces. In Proceedings of EMNLP, 1201-1211

- Kai Sheng Tai, Richard Socher and Christopher Manning. 2015. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. In Proceedings of ACL, 1556-1566

- Peter Turney. 2006. Similarity of Semantic Relations. Computational Linguistics 32(3):379-416

- David Weir, Julie Weeds, Jeremy Reffin and Thomas Kober. 2016. Aligning Packed Dependency Trees: A theory of composition for distributional semantics. Computational Linguistics 42(4):727-761

- Ludwig Wittgenstein. 1953. Philosophical Investigations

- Fabio Massimo Zanzotto, Ioannis Korkontzelos, Francesca Fallucchi and Suresh Manandhar. 2010. Estimating Linear Models for Compositional Distributional Semantics. In Proceedings of COLING, 1263-1271