



Subtrees as a new type of context in Word Space Models

Margaux Smets, Dirk Speelman and Dirk Geeraerts



KULeuven
Quantitative Lexicology and Variational Linguistics

Gent, February 11th, 2011

Overview

Introduction

Subtrees as contexts

Let statistics decide

Extension to phrases

Conclusion



Two types of context

1. bag-of-words models
 - how large a window?



Two types of context

1. bag-of-words models
 - how large a window?
2. syntactic models
 - perform better in general
 - subject/verb and verb/object, other (in)direct dependencies?



Problems

- choice of contexts
⇒ DOP-philosophy:

“take all and let statistics decide”



Problems

- choice of contexts
⇒ DOP-philosophy:

“take all and let statistics decide”

- representation of individual **words**
⇒ uniform representation of **phrases**
(e.g. 'the young boys', 'kids')



Overview

Introduction

Subtrees as contexts

Let statistics decide

Extension to phrases

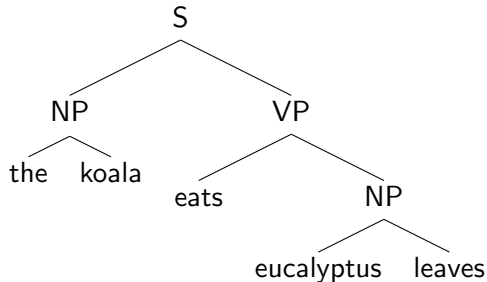
Conclusion



New type of context: subtrees

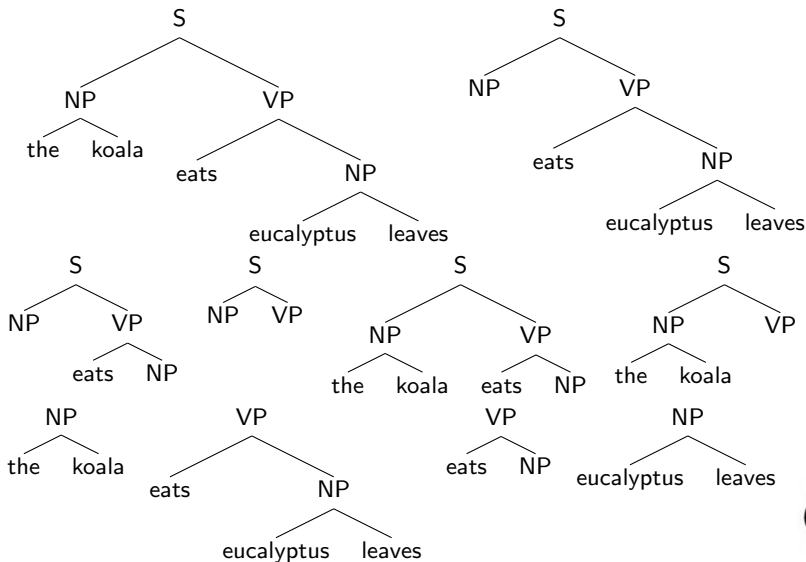
Training corpus:

1. the koala eats eucalyptus leaves
2. the wombat eats eucalyptus leaves
3. the lion eats a lamb
4. the tree crushes a car



New type of context: **subtrees**

Extracted subtrees from 'the koala eats eucalyptus leaves':

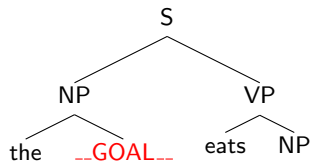


New type of context: subtrees

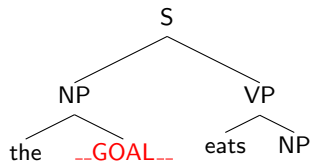
'Two words are similar
if they can occur at the same place in the same subtrees'



New type of context: subtrees



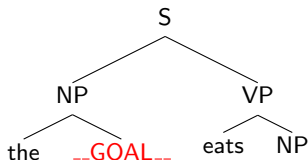
New type of context: subtrees



- 'koala': value 1
- 'wombat': value 1
- 'lion': value 1
- 'tree': value 0



New type of context: subtrees



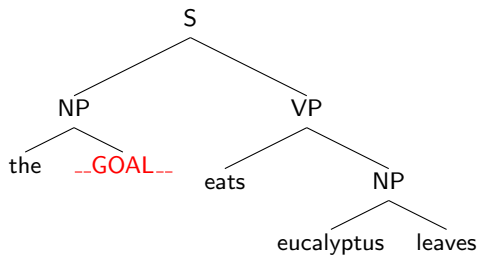
- 'koala': value 1
- 'wombat': value 1
- 'lion': value 1
- 'tree': value 0

⇒ 'koala', 'wombat' and 'lion' are similar, because they can occur as the subject of 'eat'

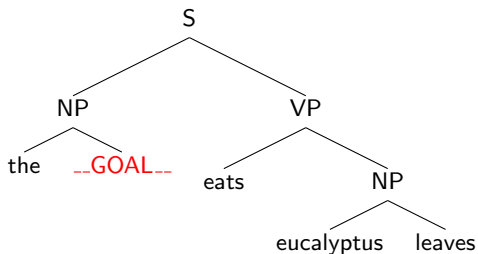
⇒ 'tree' is different, because it doesn't occur as the subject of 'eat'



New type of context: subtrees



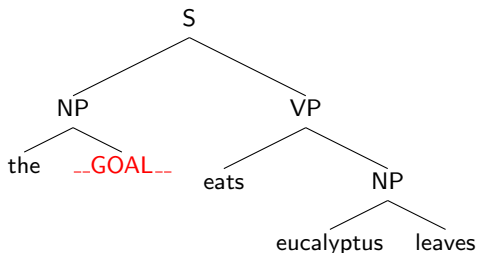
New type of context: subtrees



- 'koala': value 1
- 'wombat': value 1
- 'lion': value 0
- 'tree': value 0



New type of context: subtrees



- 'koala': value 1
- 'wombat': value 1
- 'lion': value 0
- 'tree': value 0

⇒ 'koala' and 'wombat' are similar, because they both co-occur with 'eucalyptus leaves'

⇒ 'lion' is different, because it doesn't occur with 'eucalyptus leaves'



New type of context: subtrees

- 'koala' and 'wombat' most similar
- 'lion' more similar to 'koala' and 'wombat' than to 'tree'
- 'tree' not similar to others



New type of context: subtrees

- subtrees as contexts:
 - capture co-occurrences, all window lengths
 - capture all syntactic dependencies, direct and indirect
- 'take all' \approx 'extracting all subtrees'
 - efficient (chart representation, Goodman-reduction, . . .)



Overview

Introduction

Subtrees as contexts

Let statistics decide

Extension to phrases

Conclusion



Let statistics decide

Various selection mechanisms:

- **subtree depth**: only consider subtrees with a depth of maximum `subtreeDepth` as contexts
≈ reducing context window
- **subtree frequency**: only consider subtrees with a frequency of minimum `subtreeFrequency` as contexts
- **subtree variance**: only consider the top `subtreeVariance` subtrees with the highest variance
(variance only calculated on target elements themselves)



Overview

Introduction

Subtrees as contexts

Let statistics decide

Extension to phrases

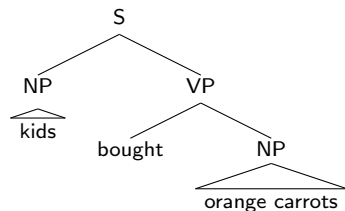
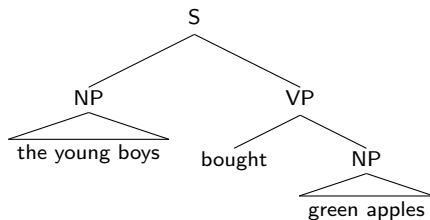
Conclusion



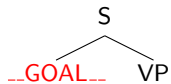
Extension to phrases

Training corpus:

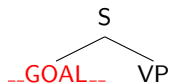
1. the young boy bought green apples
2. kids bought orange carrots
3. the giant dog ate fresh food



Extension to phrases



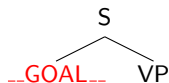
Extension to phrases



- 'the young boys': value 1
- 'kids': value 1
- 'the giant dog': value 1
- 'fresh food': value 0



Extension to phrases



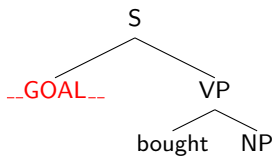
- 'the young boys': value 1
- 'kids': value 1
- 'the giant dog': value 1
- 'fresh food': value 0

⇒ 'the young boys', 'kids' and 'the giant dog' are similar, because they can occur as a subject

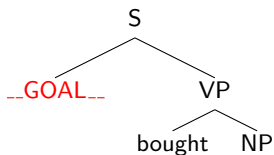
⇒ 'fresh food' is different, because it doesn't occur as a subject



Extension to phrases



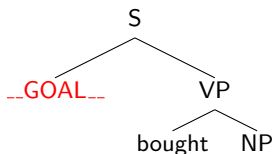
Extension to phrases



- 'the young boys': value 1
- 'kids': value 1
- 'the giant dog': value 0
- 'fresh food': value 0



Extension to phrases



- 'the young boys': value 1
- 'kids': value 1
- 'the giant dog': value 0
- 'fresh food': value 0

⇒ 'the young boys' and 'kids' are similar, because they can occur as the subject of 'bought'

⇒ 'the giant dog' is different, because it doesn't occur as the subject of 'bought'



Extension to phrases

- 'the young boys' and 'kids' most similar
- 'the giant dog' more similar to 'the young boys' and 'kids' than to 'fresh food'
- 'fresh food' not similar to others



Overview

Introduction

Subtrees as contexts

Let statistics decide

Extension to phrases

Conclusion



Conclusion and further work

- subtrees as contexts
 - capture all kinds of contexts
 - 'easy' to take all
 - easy to extend to phrases



Conclusion and further work

- subtrees as contexts
 - capture all kinds of contexts
 - 'easy' to take all
 - easy to extend to phrases
- to do
 - experiments, compare with other approaches
 - make representation of phrases fully compositional



Thank you!

