Syntactic subtree embedding

Martin Kroon | LUCL—LUCDH—LCDS

23-3-2018

Universiteit Leiden
The Netherlands

Syntactic subtree embedding

1. The idea

2. Application
    1. Syntactic equivalents
    2. Syntactic subtree alignment

3. Issues

The idea

Goldberg and Levy (2014) define their context based on dependency parses

1.	children play with LEGO
2.	LEGO is a line of construction toys

window=1	play/obj^{-1}	with/case	line/nsubj^{-1}	play/nsubj^{-1}	…
LEGO	1	1	1	0	…
children	0	0	0	1	…
…	…	…	…	…	…

The idea

Goldberg and Levy (2014) define their context based on dependency parses

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

Result: embedding on syntactic context

Very similar contexts mean they appear in the same syntactic contexts

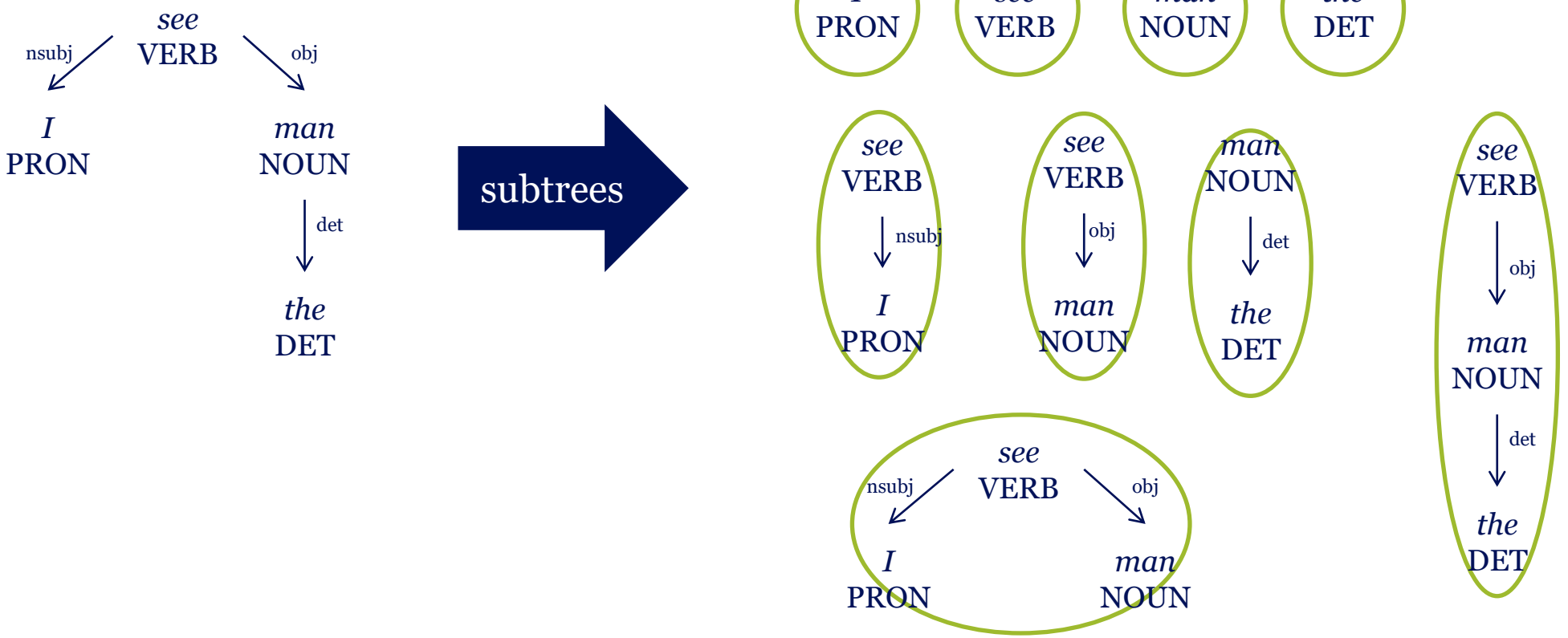
The idea

Now: embed **syntactic subtrees** based on their syntactic contexts

The idea

Now: embed **syntactic subtrees** based on their syntactic contexts

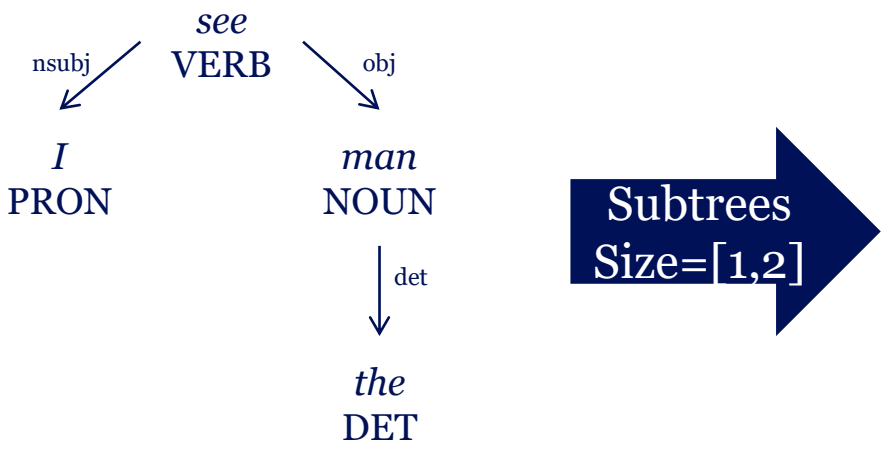
- **Subtree:** All possible combinations of nodes in a tree that are **connected**



The idea

Now: embed **syntactic subtrees** based on their syntactic contexts

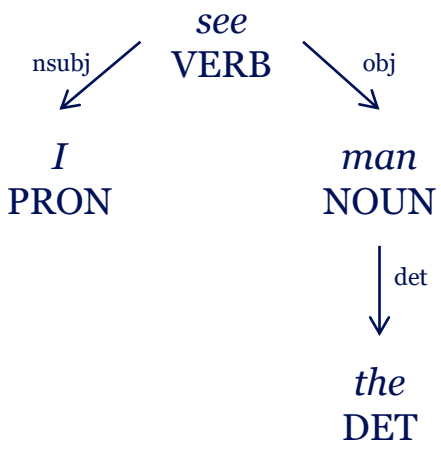
- **Subtree:** All possible combinations of nodes in a tree that are **connected**



- **Parameters:** subtree size (phrase length) and distance range (window size)

The idea

Now: embed **syntactic subtrees** based on their syntactic contexts



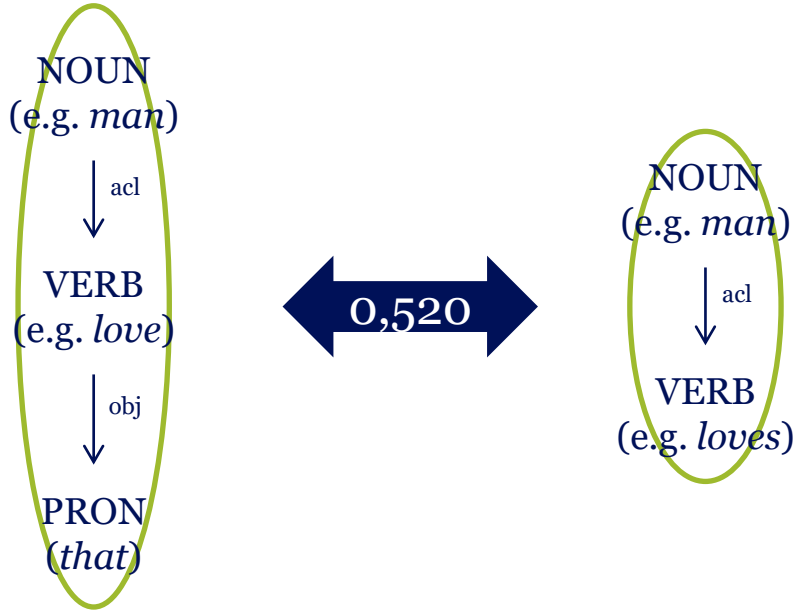
vectorize

Size=2, dist=[0,1]	[PRON]/nsubj	[NOUN]/obj	[NOUN det:[DET]]/obj	[VERB]	[VERB]/nsubj^{-1}	…
[VERB]	1	1	1	1	0	…
[PRON]	0	0	0	0	1	…
…	…	…	…	…	…	…
[VERB nsubj:[PRON]]	0	1	1	0	0	…
…	…	…	…	…	…	…

The idea

- **Normally:** vector space with **semantically related** words closely embedded together

- **Now: syntactic subtrees** with **similar syntactic behaviour** closely emdedded together
  - English:
    - 100 sentences, size=[1,2,3], distance=[1]
    - 1341 subtrees, 9773 context dimensions

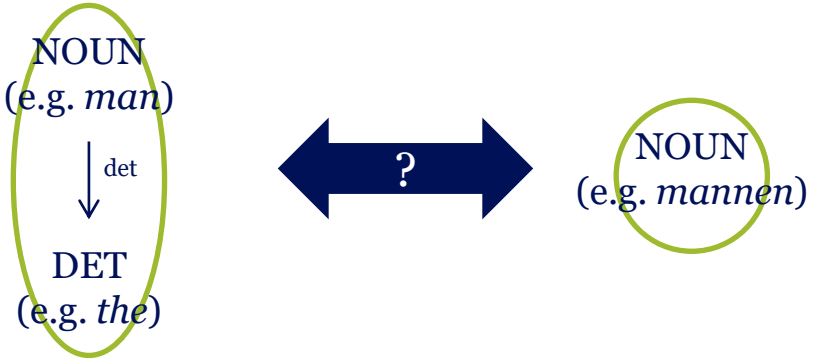


Application – syntactic equivalents

- Embed **two languages** into the **same vector space**
  - English-Swedish:
    - 100 sentences each, size=[1,2,3], distance=[0,1]
    - 2623 subtrees, 17877 context dimensions

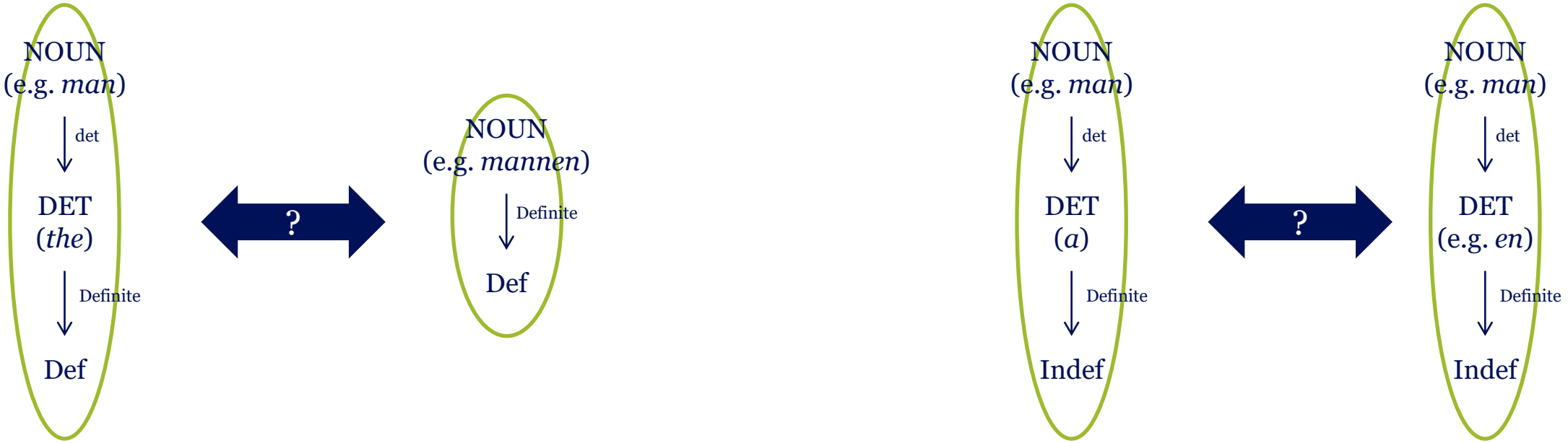
Application – syntactic equivalents

- Embed **two languages** into the **same vector space**
  - English-Swedish:
    - Morphology: more detailed results



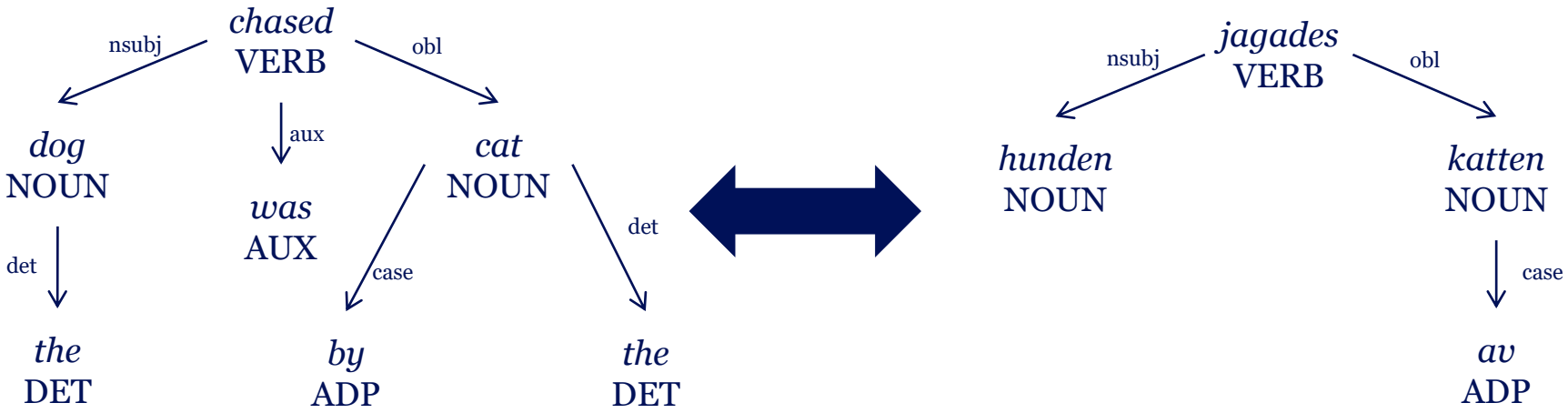
Application – syntactic equivalents

- Embed **two languages** into the **same vector space**
  - English-Swedish:
    - Morphology: more detailed results



Application – subtree alignment

- The dog was chased by the cat. ↔ Hunden jagades av katten.



Application – subtree alignment

- The dog was chased by the cat. ↔ Hunden jagades av katten.

English	Swedish	Cosine sim.
the dog	hunden	0.612
was chased	jagades	0.667
by	av	0.548
the cat	katten	0.603

Application – subtree alignment

- How to find the **best alignment**?
  - All nodes/words must receive exactly one alignment:

    by ↔ av | by the cat ↔ av katten

    cat ↔ katten; the ↔ Ø

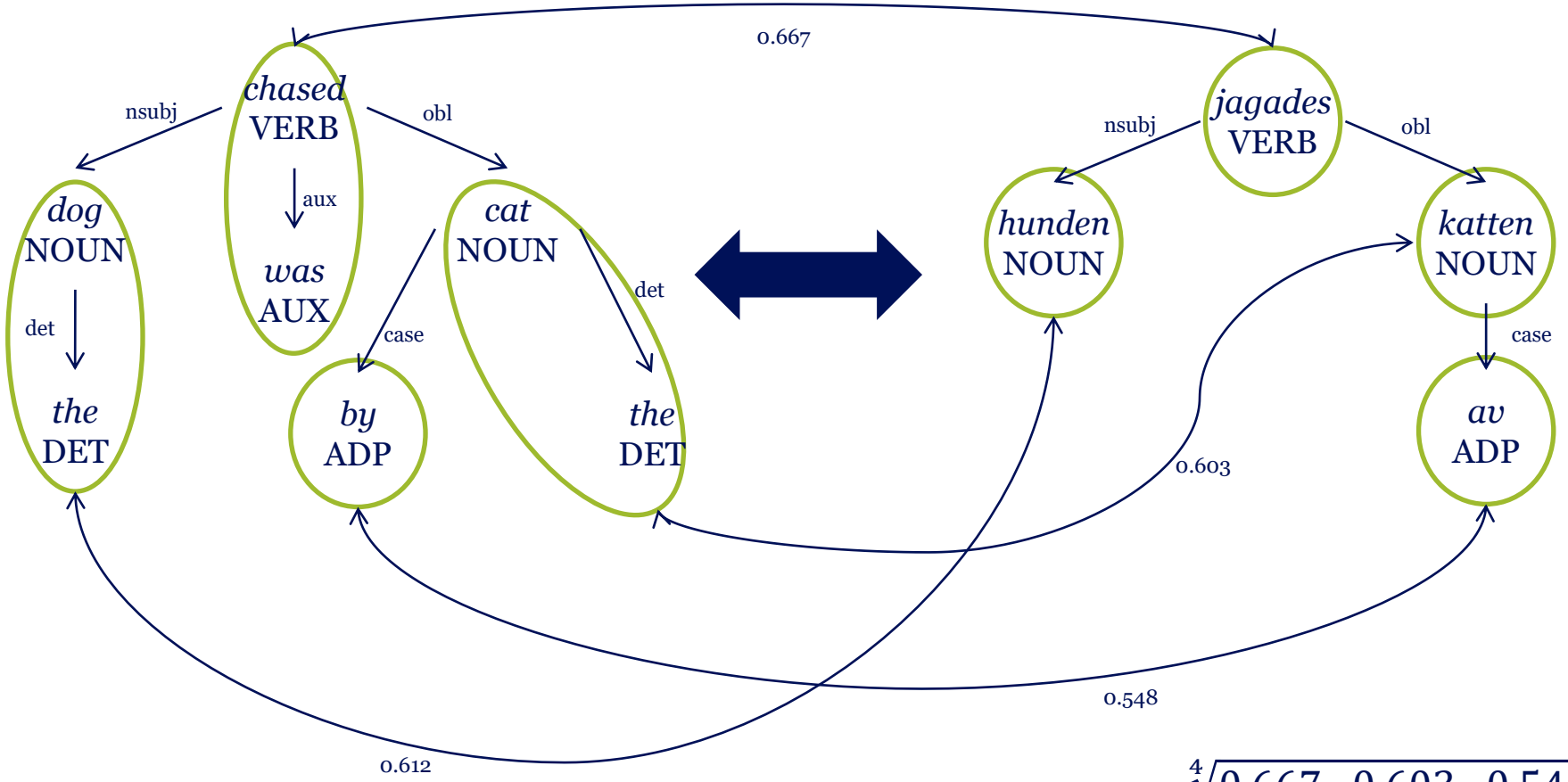
  - For all possible alignments, calculate alignment score

  - Alignment with highest score is best

  - But how to efficiently find all possible alignments? What should be the score?

Application – subtree alignment

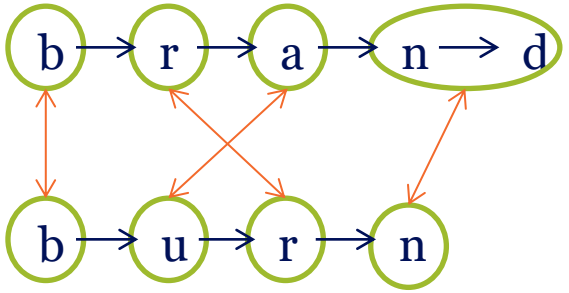
- Geometric average of sub-alignment scores?



$$\sqrt[4]{0.667 \cdot 0.603 \cdot 0.548 \cdot 0.612} = 0.606$$

Application – subgraph alignment

- Substring alignment:
  - Dutch *brand* ↔ English *burn*



- Chemistry…

- Multi-output classification?

Issues

- Slow.
  - The longer the sentences, the larger the amount of subtrees (combinatorics)
  - For alignment, the number of possible alignments also grows

Issues

- Bidirectional probability distribution?
  - If in 95% of the cases Dutch *de* wants to be aligned to English *the*, but English *the* only want to be aligned to *de* in 75% of the cases (20% being *het*), what is the probability of *de* aligning to *the* **and** *the* aligning to *de*?

	the	that
de	1425	75
het	380	20
dat	95	405

  - $\boldsymbol{P_n(x \leftrightarrow y) = \frac{P_{n-1}(x \leftrightarrow y)}{\sum_y P_{n-1}(x \leftrightarrow y)} \times \frac{P_{n-1}(y \leftrightarrow x)}{\sum_x P_{n-1}(y \leftrightarrow x)}}$ **where** $\boldsymbol{P_0(x \leftrightarrow y) = \frac{N(x \leftrightarrow y)}{\sum_y N(x \leftrightarrow y)} \times \frac{N(y \leftrightarrow x)}{\sum_x N(y \leftrightarrow x)}}$

Issues

- Bidirectional probability distribution?
  - If in 95% of the cases Dutch *de* wants to be aligned to English *the*, but English *the* only want to be aligned to *de* in 75% of the cases (20% being *het*), what is the probability of *de* aligning to *the* **and** *the* aligning to *de*?

	the	that
de	0.789	0.000
het	0.211	0.000
dat	0.000	1.000

  - $\boldsymbol{P_n(x \leftrightarrow y) = \frac{P_{n-1}(x \leftrightarrow y)}{\sum_y P_{n-1}(x \leftrightarrow y)} \times \frac{P_{n-1}(y \leftrightarrow x)}{\sum_x P_{n-1}(y \leftrightarrow x)}}$ **where** $\boldsymbol{P_0(x \leftrightarrow y) = \frac{N(x \leftrightarrow y)}{\sum_y N(x \leftrightarrow y)} \times \frac{N(y \leftrightarrow x)}{\sum_x N(y \leftrightarrow x)}}$

  - Very sensitive to rounding errors! But reduces runtime significantly.

Issues

- Slow.
  - The longer the sentences, the larger the amount of subtrees (combinatorics)
  - For alignment, the number of possible alignments also grows

That’s all Folks

Universiteit
Leiden
The Netherlands

Discover the world at Leiden University

Application – multi-output classification?

- Latin nouns
  - Training set: 2050 words with UD morphology tags

Size=1,2,3	Acc	Sg	Masc	Acc, Sg	Acc, Masc	…
m$	1	1	1	1	1	…
um$	1	1	1	1	1	…
…	…	…	…	…	…	…
am$	1	1	0	1	0	…
…	…	…	…	…	…	…

	1	**2**	**3**
Acc,Masc,Pl	os$	os	los
Acc,Fem,Sg	am$	am	em$
…	…	…	…
Abl,Plur	bus	bu	ibu
…	…	…	…

Interesting…