

Outline

Introduction

- Syntax-based distributional similarity
- Transitivity of similarity

Method

- Introducing transitivity

Experiments

- Data
- Evaluation
- Results

Conclusions

Distributional similarity (syntax-based)

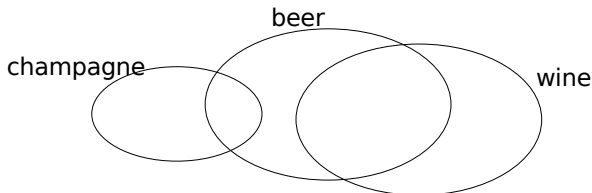
- ▶ Extract syntactic co-occurrences from parsed Dutch newspaper text

	drink-obj 'drink-obj'	koud-adj 'cold-adj'	praat-subj 'talk-subj'
bier 'beer'	89	79	1
wijn 'wine'	76	14	0
premier 'prime-minister'	1	2	240

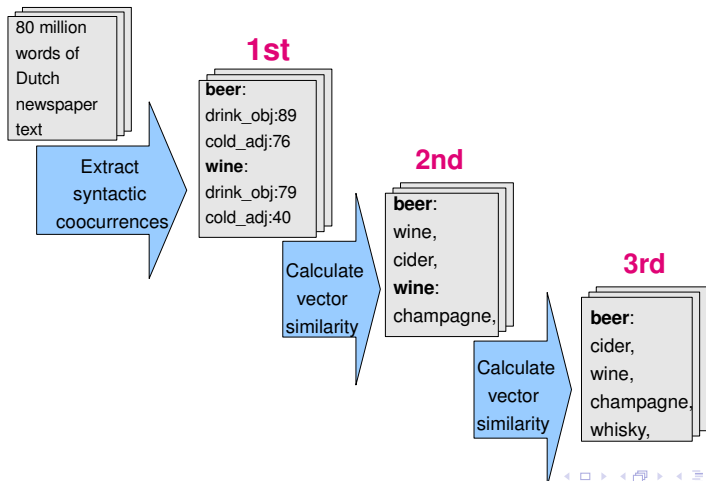
- ▶ Compare the vectors using Dice[†] (Curran and Moens, 2002) as measure and pointwise mutual information as weight

Transitivity of similarity

- ▶ If A is similar to B, and B is similar to C \rightarrow A is similar to C
- ▶ We can infer things we have not seen from evidence
- ▶ Particularly handy when there is data sparseness



First, second, third



Second-order similarity score

The second-order similarity score (SOSS) for a given headword (h) and a given nearest neighbour (nn) is defined as follows:

$$SOSS(h,nn) = \frac{\text{max.freq.of.first.order}(h)}{\text{rank}(nn)}$$

The second-order affinity score does not dominate.

1st-order gracht 'canal'			2nd-order gracht 'canal'		
97	Amsterdams_ADJ	'Amsterdam_ADJ'	97	gracht	'canal'
26	is_SUBJ	'is_SUBJ'	48	laan	'avenue'
12	wordt_SUBJ	'becomes_SUBJ'	32	sloot	'ditch'

Outline

Introduction

- Syntax-based distributional similarity
- Transitivity of similarity

Method

- Introducing transitivity

Experiments

- Data
- Evaluation
- Results

Conclusions

Test sets

Four test sets of each 1000 words:

1. **High-frequency** test set:
scène 'scene' (2,278) up to
jaar 'year' (258,253)
2. **Middle-frequency** test set:
vredesverdrag 'peace treaty' (364) up to
celstraf 'jail sentence' (541)
3. **Low-frequency** test set:
vriendenprijs 'special price' (23) up to
röntgenonderzoek 'x-ray research' (28)
4. **Very-low-frequency** test set:
cederhout 'cedar wood' (8) up to
slaginstrument 'percussion instrument' (9)

Wu and Palmer EWN similarity

		EWN similarity		
		top-1	top-5	top-10
HF	Orig	0.72	0.65	0.61
	Combi	0.72	0.65	0.61
	Nn	0.72	0.64	0.62
MF	Orig	0.64	0.59	0.56
	Combi	0.65	0.60	0.56
	Nn	0.64	0.60	0.58
LF	Orig	0.43	0.39	0.37
	Combi	0.43	0.40	0.38
	Nn	0.44	0.43	0.43
VLF	Orig	0.39	0.36	0.35
	Combi	0.40	0.38	0.36
	Nn	0.41	0.41	0.41

Synonyms

		top-1		top-10	
		#	%	#	%
HF	Orig	143	14.4	461	4.6
	Combi	148	15.0	465	4.7
	Nn	154	15.5	382	5.3
MF	Orig	105	10.6	312	3.1
	Combi	109	11.0	318	3.2
	Nn	107	11.4	214	4.0
LF	Orig	33	3.8	108	1.3
	Combi	34	3.9	113	1.3
	Nn	25	4.0	54	3.2
VLF	Orig	2	0.5	10	0.3
	Combi	2	0.5	10	0.3
	Nn	2	0.9	2	0.4

Example Output

Nearest neighbours for
videoband 'video tape',
cassette 'cassette'
bandje 'tape' and
CDi 'CDi'

	Orig			Combi
cassette	videoband	bandje	CDi	cassette
cassette	videoband	bandje	CDi	cassette
videoband	cassette	cassette	DCC	videoband
CDi	videofilm	videoband	CD	bandje

Outline

Introduction

- Syntax-based distributional similarity
- Transitivity of similarity

Method

- Introducing transitivity

Experiments

- Data
- Evaluation
- Results

Conclusions

