

Language in 3D

Multi-way factorization algorithms to tackle semantics

Tim Van de Cruys

University of Cambridge

CCLS, Columbia University
Thursday 30 June, 2011

Distributional similarity

Distributional similarity models are able to infer (lexical) semantics from text:

- semantically similar words (syntactic context, small window)
 - **train**: *bus, ferry, boat, coach, car, plane, vehicle, taxi, ship, truck, ...*
 - **doctor**: *nurse, GP, physician, practitioner, midwife, dentist, surgeon, ...*

Distributional similarity

Distributional similarity models are able to infer (lexical) semantics from text:

- topically related words (large window)
 - **train**: *bus, journey, railway, station, passenger, ride, stop, taxi, fare, ...*
 - **doctor**: *medication, GP, surgery, hospital, sufferer, clinic, nurse, treatment, illness, ...*

Two-way vs. three-way

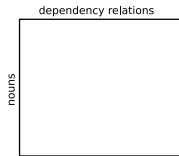
- all methods use two way co-occurrence frequencies \longrightarrow matrix
- suitable for two-way problems
 - words \times documents
 - nouns \times dependency relations
- not suitable for n -way problems
 - words \times documents \times authors
 - verbs \times subjects \times direct objects

Two-way vs. three-way

- all methods use two way co-occurrence frequencies \longrightarrow matrix
- suitable for two-way problems
 - words \times documents
 - nouns \times dependency relations
- not suitable for n -way problems \longrightarrow ?
 - words \times documents \times authors
 - verbs \times subjects \times direct objects

Two-way vs. three-way

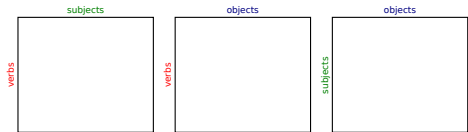
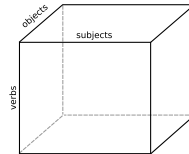
two-way



matrix

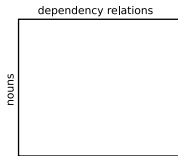


three-way



Two-way vs. three-way

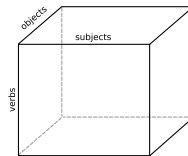
two-way



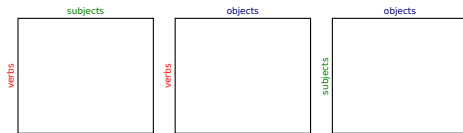
matrix



three-way

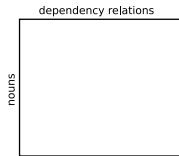


tensor



Two-way vs. three-way

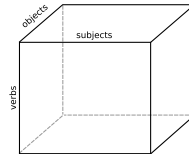
two-way



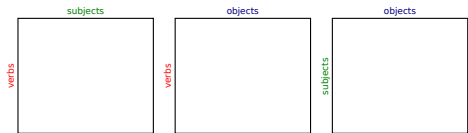
matrix



three-way



tensor



pairwise matrices

Factorization

Two reasons for performing dimensionality reduction:

- Intractable computations
 - When number of elements and number of features is too large, similarity computations may become intractable
 - reduction of the number of features makes computation tractable again
- Generalization capacity
 - the dimensionality reduction is able to describe the data better, or is able to capture intrinsic semantic features
 - dimensionality reduction is able to improve the results (counter data sparseness and noise)

Different flavours

- Principal component analysis
- Latent semantic analysis (singular value decomposition)
- Probabilistic latent semantic analysis
- Topic models – latent dirichlet allocation
- Non-negative matrix factorization

Non-negative matrix factorization

- Given a non-negative matrix V , find non-negative matrix factors W and H such that:

$$\mathbf{V}_{n \times m} \approx \mathbf{W}_{n \times r} \mathbf{H}_{r \times m} \quad (1)$$

- Choosing $r \ll n, m$ reduces data
- Constraint on factorization: all values in three matrices need to be *non-negative values* (≥ 0)
- Constraint brings about a *parts-based* representation: only additive, no subtractive relations are allowed

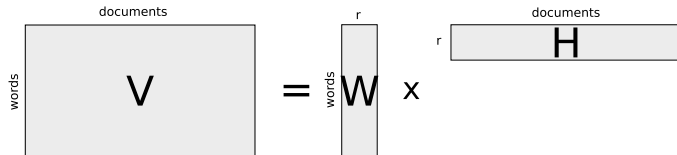
Non-negative matrix factorization

- Different kinds of NMF's that minimize different cost functions:
 - Square of Euclidean distance (L1-norm)
 - Kullback-Leibler Divergence (L2-norm)
⇒ better suited for language phenomena
- To find NMF is to minimize $D(V \| WH)$ with respect to W and H , subject to the constraints $W, H \geq 0$
- This can be done with *update rules*

$$\mathbf{H}_{a\mu} \leftarrow \mathbf{H}_{a\mu} \frac{\sum_i \mathbf{W}_{ia} \frac{\mathbf{v}_{i\mu}}{(\mathbf{WH})_{i\mu}}}{\sum_k \mathbf{W}_{ka}} \quad \mathbf{W}_{ia} \leftarrow \mathbf{W}_{ia} \frac{\sum_{\mu} \mathbf{H}_{a\mu} \frac{\mathbf{v}_{i\mu}}{(\mathbf{WH})_{i\mu}}}{\sum_v \mathbf{H}_{av}} \quad (2)$$

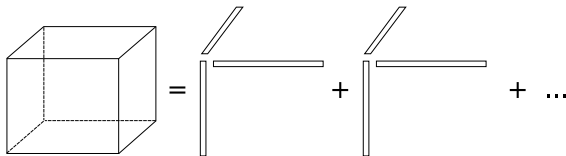
- these update rules converge to a *local optimum* in the minimization of KL divergence

Graphical Representation

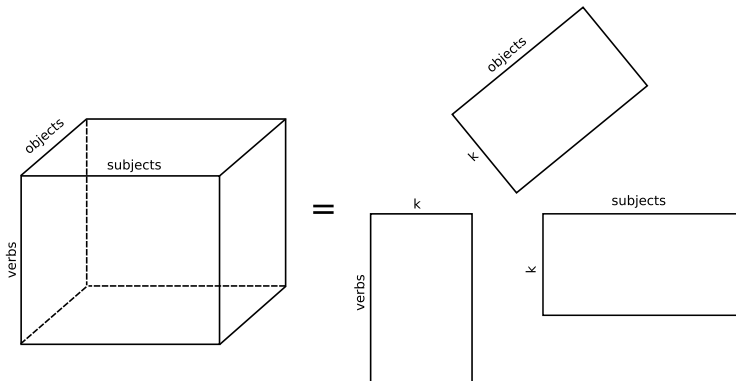


Technique

- Idea similar to non-negative matrix factorization
- Calculations are different
- $\min_{x_i \in \mathbb{R}_{\geq 0}^{D_1}, y_i \in \mathbb{R}_{\geq 0}^{D_2}, z_i \in \mathbb{R}_{\geq 0}^{D_3}} \| T - \sum_{i=1}^k x_i \circ y_i \circ z_i \|_F^2$



Graphical representation



Introduction 1/2

- Standard selectional preference models: two-way co-occurrences
- Keeping track of single relationships
- But: two-way selectional preference models are not sufficiently rich
- Compare:
 - *The skyscraper is playing coffee.*
 - *The turntable is playing the piano.*

Introduction 2/2

- *The skyscraper is playing coffee.*
 - $(play, su, scyscraper) \downarrow$
 - $(play, obj, coffee) \downarrow$
- *The turntable is playing the piano.*
 - $(play, su, turntable) \uparrow$
 - $(play, obj, piano) \uparrow$
 - $(play, su, turntable, obj, piano) \downarrow$

Methodology

- Three-way extraction of selectional preferences
- Approach applied to Dutch, using TWENTE NIEUWS CORPUS (500M words of newspaper texts)
- parsed with Dutch dependency parser ALPINO
- three-way co-occurrence of verbs with subjects and direct objects extracted
- adapted with extension of pointwise mutual information
- Resulting tensor 1K verbs \times 10K subjects \times 10K direct objects
- reduction to k dimensions ($k = 50, 100, 300$)

Examples: police action

subjects	su_s	verbs	v_s	objects	obj_s
<i>politie</i> 'police'	.99	<i>houd_aan</i> 'arrest'	.64	<i>verdachte</i> 'suspect'	.16
<i>agent</i> 'policeman'	.07	<i>arresteer</i> 'arrest'	.63	<i>man</i> 'man'	.16
<i>autoriteit</i> 'authority'	.05	<i>pak_op</i> 'run in'	.41	<i>betoger</i> 'demonstrator'	.14
<i>Justitie</i> 'Justice'	.05	<i>schiet_dood</i> 'shoot'	.08	<i>relschopper</i> 'rioter'	.13
<i>recherche</i> 'detective force'	.04	<i>verdenk</i> 'suspect'	.07	<i>raddraaier</i> 'instigator'	.13
<i>marechaussee</i> 'military police'	.04	<i>tref_aan</i> 'find'	.06	<i>overvaller</i> 'raider'	.13
<i>justitie</i> 'justice'	.04	<i>achterhaal</i> 'overtake'	.05	<i>Roemeen</i> 'Romanian'	.13
<i>arrestatieteam</i> 'special squad'	.03	<i>verwijder</i> 'remove'	.05	<i>actievoerder</i> 'campaigner'	.13
<i>leger</i> 'army'	.03	<i>zoek</i> 'search'	.04	<i>hooligan</i> 'hooligan'	.13
<i>douane</i> 'customs'	.02	<i>spoor_op</i> 'track'	.03	<i>Algerijn</i> 'Algerian'	.13

Examples: legislation

subjects	su_s	verbs	v_s	objects	obj_s
<i>meerderheid</i> 'majority'	.33	<i>steun</i> 'support'	.83	<i>motie</i> 'motion'	.63
<i>VVD</i>	.28	<i>dien_in</i> 'submit'	.44	<i>voorstel</i> 'proposal'	.53
<i>D66</i>	.25	<i>neem_aan</i> 'pass'	.23	<i>plan</i> 'plan'	.28
<i>Kamermeerderheid</i>	.25	<i>wijs_af</i> 'reject'	.17	<i>wetsvoorstel</i> 'bill'	.19
<i>fractie</i> 'party'	.24	<i>verwerp</i> 'reject'	.14	<i>hem</i> 'him'	.18
<i>PvdA</i>	.23	<i>vind</i> 'think'	.08	<i>kabinet</i> 'cabinet'	.16
<i>CDA</i>	.23	<i>aanvaard</i> 'accepts'	.05	<i>minister</i> 'minister'	.16
<i>Tweede Kamer</i>	.21	<i>behandel</i> 'treat'	.05	<i>beleid</i> 'policy'	.13
<i>partij</i> 'party'	.20	<i>doe</i> 'do'	.04	<i>kandidatuur</i> 'candidature'	.11
<i>Kamer</i> 'Chamber'	.20	<i>keur_goed</i> 'pass'	.03	<i>amendement</i> 'amendment'	.09

Examples: exhibition

subjects	su_s	verbs	v_s	objects	obj_s
<i>tentoonstelling</i> 'exhibition'	.50	<i>toon</i> 'display'	.72	<i>schilderij</i> 'painting'	.47
<i>expositie</i> 'exposition'	.49	<i>omvat</i> 'cover'	.63	<i>werk</i> 'work'	.46
<i>galerie</i> 'gallery'	.36	<i>bevat</i> 'contain'	.18	<i>tekening</i> 'drawing'	.36
<i>collectie</i> 'collection'	.29	<i>presenteer</i> 'present'	.17	<i>foto</i> 'picture'	.33
<i>museum</i> 'museum'	.27	<i>laat</i> 'let'	.07	<i>sculptuur</i> 'sculpture'	.25
<i>oeuvre</i> 'oeuvre'	.22	<i>koop</i> 'buy'	.07	<i>aquarel</i> 'aquarelle'	.20
<i>Kunsthall</i>	.19	<i>bezit</i> 'own'	.06	<i>object</i> 'object'	.19
<i>kunstenaar</i> 'artist'	.15	<i>zie</i> 'see'	.05	<i>beeld</i> 'statue'	.12
<i>dat</i> 'that'	.12	<i>koop_aan</i> 'acquire'	.05	<i>overzicht</i> 'overview'	.12
<i>hij</i> 'he'	.10	<i>in huis heb</i> 'own'	.04	<i>portret</i> 'portrait'	.11

Examples: quality count

- 44 dimensions contain similar, framelike semantics
- 43 dimensions contain less clear-cut semantics
 - single verbs account for one dimension
 - verb senses are mixed up
- 13 dimensions based on syntax rather than semantics
 - fixed expressions
 - pronomina

Evaluation: methodology

- pseudo-disambiguation task to test generalization capacity (standard automatic evaluation for selectional preferences)

<i>s</i>	<i>v</i>	<i>o</i>	<i>s'</i>	<i>o'</i>
<i>jongere</i>	<i>drink</i>	<i>bier</i>	<i>coalitie</i>	<i>aandeel</i>
'youngster'	'drink'	'beer'	'coalition'	'share'
<i>werkgever</i>	<i>riskeer</i>	<i>boete</i>	<i>doel</i>	<i>kopzorg</i>
'employer'	'risk'	'fine'	'goal'	'worry'
<i>directeur</i>	<i>zwaai</i>	<i>scepter</i>	<i>informatieur</i>	<i>vodka</i>
'manager'	'sway'	'sceptre'	'informer'	'wodka'

- 10-fold cross validation ($\pm 300,000$ co-occurrences)

Evaluation: models

- Evaluation of 4 different models
- 2 matrix models
 - $1\text{K verbs} \times (10\text{K subjects} + 10\text{K direct objects})$
 - singular value decomposition (\mathbb{R})
 - non-negative matrix factorization ($\mathbb{R}_{\geq 0}$)
- 2 tensor models
 - $1\text{K verbs} \times 10\text{K subjects} \times 10\text{K direct objects}$
 - parallel factor analysis (\mathbb{R})
 - non-negative tensor factorization ($\mathbb{R}_{\geq 0}$)

Evaluation: results

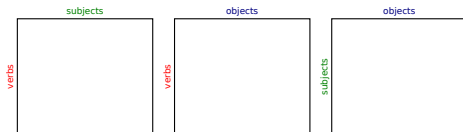
	dimensions		
	50 (%)	100 (%)	300 (%)
SVD	69.60 \pm 0.41	62.84 \pm 1.30	45.22 \pm 1.01
NMF	81.79 \pm 0.15	78.83 \pm 0.40	75.74 \pm 0.63
PARAFAC	85.57 \pm 0.25	83.58 \pm 0.59	80.12 \pm 0.76
NTF	89.52 \pm 0.18	90.43 \pm 0.14	90.89 \pm 0.16

What if tensor factorization is infeasible?

- Significant space requirements, not feasible for large dataset

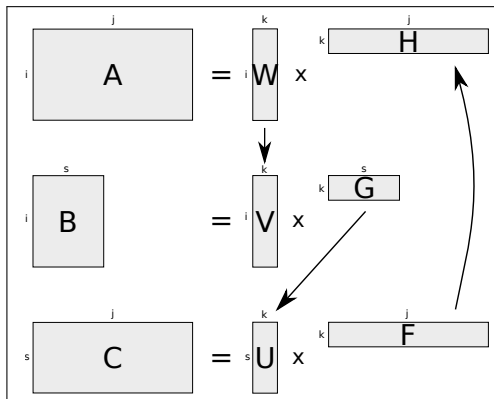
What if tensor factorization is infeasible?

- Significant space requirements, not feasible for large dataset
- Solution: use pairwise co-occurrences and combine matrices in factorization



- \Rightarrow Apply NMF to matrices, but interleave the process
- Result of former update step is used to initialize the next one

Graphical Representation



Ambiguity

- **Problem:** ambiguity
- BAR

Ambiguity

- **Problem:** ambiguity
- BAR



Ambiguity

- **Problem:** ambiguity
- BAR



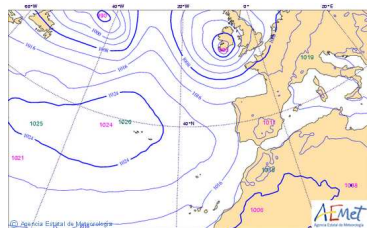
Ambiguity

- **Problem:** ambiguity
- BAR



Ambiguity

- **Problem:** ambiguity
- BAR



Ambiguity

- **Problem:** ambiguity
- BAR



Ambiguity

- **Problem:** ambiguity
- BAR

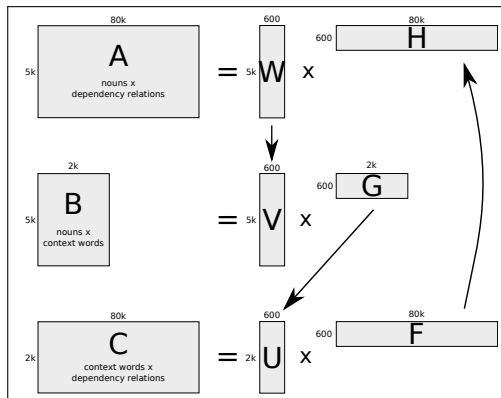


- Main research question: can 'topical' similarity and tight, synonym-like similarity be combined to compute meaning of word in a particular context?

Methodology

- Goal: classification of nouns according to both window-based context (with large window) and syntactic context
- \Rightarrow Construct three matrices capturing co-occurrence frequencies for each mode
 - nouns cross-classified by dependency relations
 - nouns cross-classified by (bag of words) context words
 - dependency relations cross-classified by context words
- \Rightarrow Apply NMF to matrices, but interleave the process
- Result of former factorization is used to initialize factorization of the next one

Graphical Representation



Word meaning in context

- NMF can be interpreted probabilistically
 - Matrix \mathbf{W} $\rightarrow p(w_i|\mathbf{z})$
 - Matrix \mathbf{H} $\rightarrow p(\mathbf{z}|d_j), p(\mathbf{d}|\mathbf{z})$
 - Matrix \mathbf{G} $\rightarrow p(\mathbf{z}|c_i)$
- $p(\mathbf{z}|C) = \frac{\sum_{c_i \in C} p(\mathbf{z}|c_i)}{|C|}$ – the probability distribution over latent factors given the context ('semantic fingerprint')
- $p(\mathbf{d}|C) = p(\mathbf{z}|C)p(\mathbf{d}|\mathbf{z})$ – probability distribution over dependency features given the context
- $p(\mathbf{d}|w_i, C) = p(\mathbf{d}|w_i) \cdot p(\mathbf{d}|C)$ – weight each dependency feature according to the importance given the context

Example

- ① Jack is listening to a **record**. $\rightarrow C_1 = \{listen_{prep(to)}^{-1}\}$
 - $p(\mathbf{z}|C_1) \rightarrow p(\mathbf{d}|C_1) \rightarrow p(\mathbf{d}|w_i, C_1)$
 - **record**_N, C_1 : *album, song, recording, track, cd*

- ② Jill updated the **record**. $\rightarrow C_2 = \{update_{obj}^{-1}\}$
 - $p(\mathbf{z}|C_2) \rightarrow p(\mathbf{d}|C_2) \rightarrow p(\mathbf{d}|w_i, C_2)$
 - **record**_N, C_2 : *file, datum, document, database, list*

Implementational details

- method applied to English and French
 - UKWaC corpus, parsed with MaltParser
 - French Wikipedia, parsed with FRMG
- one model per pos (noun, adjective, verb, adverb)
- NMF model: $K = 600$, 100 iterations
- interleaved NMF algorithm implemented in Matlab, preprocessing and vector computation in Python.

Example dimension 44

nouns	context words	dependency relations
building/NN	building/NN	dobj-1#redevelop/VB
factory/NN	construction/NN	conj_and/cc#warehouse/NN
center/NN	build/VB	prep_of/in-1#redevelopment/NN
refurbishment/NN	station/NN	prep_of/in-1#refurbishment/NN
warehouse/NN	store/NN	conj_and/cc#dock/NN
store/NN	open/VB	prep_by/in-1#open/VB
station/NN	center/NN	nn#refurbishment/NN
construction/NN	industrial/JJ	prep_of/in-1#ft/NN
complex/NN	Street/NNP	amod#multi-storey/JJ
headquarters/NN	close/VB	prep_of/in-1#opening/NN

Example dimension 89

nouns	context words	dependency relations
virus/NN	security/NN	amod#malicious/JJ
software/NN	Microsoft/NNP	nn-1#vulnerability/NN
security/NN	Internet/NNP	conj_and/cc#worm/NN
firewall/NN	Windows/NNP	nn-1#worm/NN
spam/NN	computer/NN	nn-1#flaw/NN
Security/NNP	network/NN	nn#antivirus/NN
vulnerability/NN	attack/NN	nn#IM/NNP
system/NN	software/NN	prep_of/in#worm/NN
Microsoft/NNP	protect/VB	nn#Trojan/NNP
computer/NN	protection/NN	conj_and/cc#worm/NN

Example dimension 316

nouns	context words	dependency relations
virus/NN	brain/NN	dobj-1#infect/VB
disease/NN	animal/NN	nsubjpass-1#infect/VB
bacterium/NN	disease/NN	rcmod#infect/VB
infection/NN	human/JJ	nsubj-1#infect/VB
human/NN	blood/NN	prep_with/in-1#infect/VB
rat/NN	cell/NN	conj_and/cc#rat/NN
cell/NN	cancer/NN	prep_of/in#virus/NN
animal/NN	skin/NN	amod#infected/JJ
mouse/NN	scientist/NN	prep_of/in#flu/NN
cancer/NN	drug/NN	nn#monkey/NN

Evaluation

- Evaluated with SEMEVAL 2007 lexical substitution task
- find appropriate substitutes in context
- 200 target words (50 for each pos), 10 sentences each
- Paraphrase **ranking**: rank possible candidates, standard evaluation for unsupervised methods
 - Kendall's τ_b ranking coefficient
 - Generalized average precision
- Paraphrase **induction**: find candidates from scratch, not carried out before for unsupervised methods
 - Recall
 - Precision out-of-ten

Paraphrase ranking

model	τ_b	GAP
random	-0.61	29.98
vector _{dep}	16.57	45.08
EP09	–	32.2 ▼
EP10	–	39.9 ▼
TFP	–	45.94 ▼
DL	16.56	41.68
NMF _{context}	20.64**	47.60**
NMF _{dep}	22.49**	48.97**
NMF _{c+d}	22.59**	49.02**

Paraphrase induction

model	R_{best}	P_{10}
vector _{dep}	8.78	30.21
DL	1.06	7.59
NMF _{context}	8.81	30.49
NMF _{dep}	7.73	26.92
NMF _{c+d}	8.96	29.26

Conclusion

Beneficial to consider language as a multi-way co-occurrence problem

- Tensor space
 - novel model to investigate three-way (up to n -way) co-occurrence data
 - Possible to generalize over co-occurrence data with appropriate factorization models
 - Applicable to and beneficial for three-way selectional preference induction
- Pairwise matrices
 - 'makeshift' multi-way co-occurrence modeling
 - Useful when tensor approach is not feasible
 - Applicable to and beneficial for computation of word meaning in context