UNIVERSITÄT DES SAARLANDES

Measuring Distribution Similarity in Context

[Dinu & Lapata 2010]

Computational Semantics Seminar: Summer 2012

Boris Kozlov

29 May 2012

boris.linguist@gmail.com

Outline

- Vector-based models
  - standard approach
  - distribution similarity with latent senses
- Latent Semantic Analysis
  - observation data explained by hidden structure: Hoffman[1989]
  - latent semantic classes
- Parametrizations
  - PLSA
  - NMF
  - LDA
- Evaluations and conclusions

Vector-based models: standard

- Distributional models of meaning = vector-based models
  - Firth (1957): “You shall know a word by a company it keeps!”
- Term-content matrix:

	aardvark	computer	data	pinch	result	sugar
apricot	0	0	0	1	0	1
pineapple	0	0	0	1	0	1
digital	0	2	1	0	1	0
information	0	1	6	0	4	0

  - word meaning is defined by a vector of context words

Classic vector-based models

Cosine for computing similarity

$$\cos(v, w) \frac{\bar{v} * \bar{w}}{|\overline{v}||\overline{w}|}$$

Vector based models melt all meanings of a word into one vector:

big: *"My bigger sister…"*

(a) size

(b) age

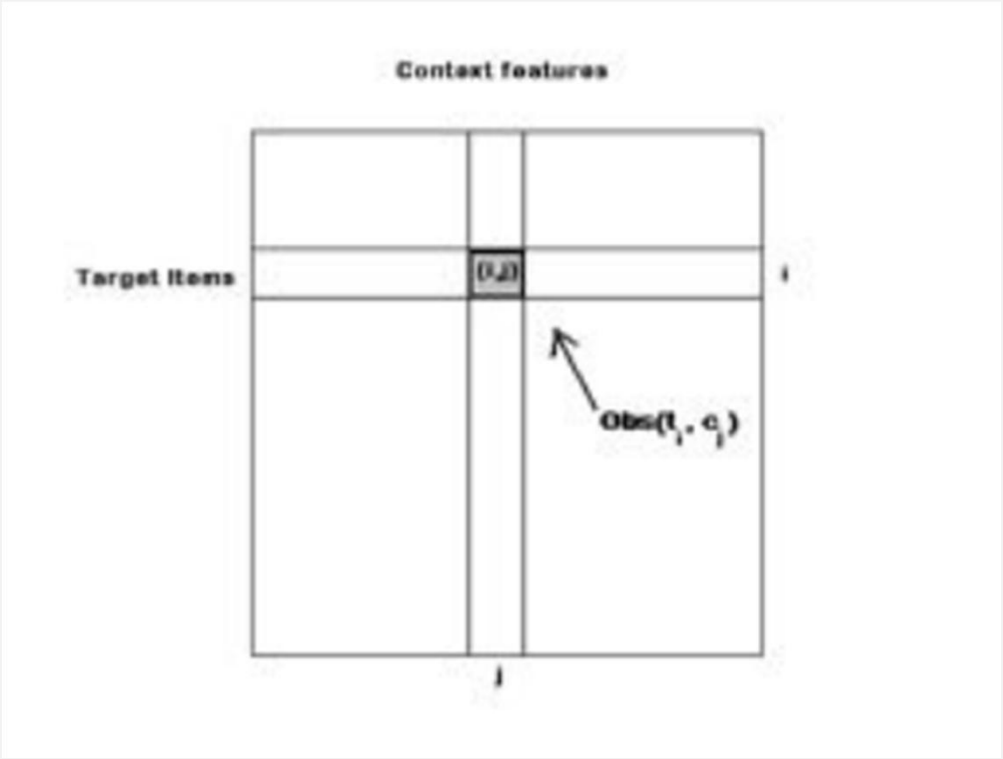
Latent Semantic Analysis

Latent class model:

- explains the structure of the observation data

Data:

- co-occurrence matrix of target words and context features



Latent Semantic Analysis

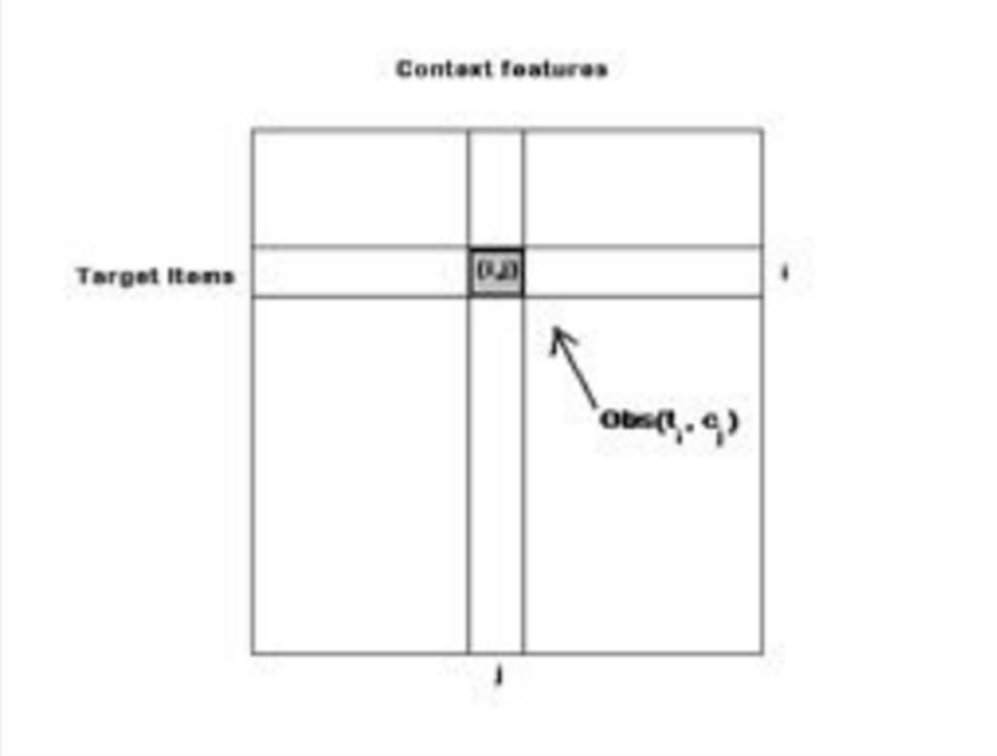
Hoffman [1989]:

finite set of aspects $A = \{a_1 \dots a_k\}$

each observation is assigned a class a_i

Observation data:

$$P(x, y) = \sum_a P(x, y, a)$$



x and y are independent:

$$P(x, y) = \sum_a P(x)P(a|x)P(y|a)$$

MLE: choose a structure that maximizes P(x,y)

Representation over latent senses

Assume :

$$Z = \{z_1 \dots z_k\}[latent\ semantic\ senses]$$

Substituting this

$$\mathrm{P}(x, y) = \sum_a P(x)P(a|x)P(y|a)$$

the same but with linguistic observation data:

$$\mathrm{P}(t_i, c_j) = \sum_k P(t_i)P(z_k|t_i)P(c_j|z_k)$$

MLE: choose a structure that maximizes P(x,y)

Meaning of a word as a distribution over the latent senses

$$v(t_i) = (P(z_1|t_i), \dots, P(z_k|t_i))$$

Representation for a target word t_i with a context feature c_j

$$v(t_i, c_j) = \left(P(z_1|t_i, c_j), \dots, P(z_k|t_i, c_j)\right)$$

Models for latent meaning induction

Where did latent senses come from??

Models for latent meaning induction

Different probabilistic models for induction of latent senses:

- Probabilistic Principal Components Analysis: Tipping and Bishop [1999]
- Probabilistic latent semantic analysis: Hofmann [2001]
- Non-negative Matrix Factorization: Lee and Seug [2000]
- Latent Dirichlet Allocation: Blei [2003]

The latter two are used in the paper

PLSA

Model:

$S = D\ x\ W$ – observation data

$D = \{d_1 \ldots d_I\}$ - set of document labels

W = $\{w_1 \ldots w_J\}$ - set of words

or:

d = $\{w_{d1} \ldots w_{dN}\}$ - document as words in it

PLSA

It is assumed that data is generated by the random sampling process:

- Select a document with a probability P(d)
- For each word position dn in a document d:
  1. Generate a latent topic z with probability P($z_{dn}|d$)
  2. Generate a word w with a probability P($w_{dn}|z_{dn}$)

Two assumptions here:

1. the pairs (d,w) are generated independently
2. Words and document are conditionally independent given a latent topic

   P(w|z, d) = $P(w|z)$

PLSA

Under the independence assumption of (d,w)

$$P(\mathcal{S}) = \prod_{d} \prod_{dn=d1}^{dN} P(d, w_{dn}) = \prod_{d} \prod_{dn=d1}^{dN} \sum_{k=1}^{K} P(d, w_{dn}, z_k)$$

the goal: reveal the underlying structure that maximizes P(S)

Under the independence assumption of P(w|z, d) = $P(w|z)$

$$P(d, w) = \sum_{z} P(d, w, z) = P(d) \sum_{z} P(z|d) P(w|z)$$

asymmetric (defined already)

$$P(d, w) = \sum_{z} P(d, w, z) = \sum_{z} P(z) P(d|z) P(w|z)$$

symmetric

Non-Negative Matrix Factorization

Non-negative Matrix Factorization (Lee & Seung [2000])

Main idea: *find non-negative matrices W and H*

$$V_{I,J} \approx W_{I,K} H_{K,J}$$

Where:

$$V_{i,j} = P(t_i, c_j) \quad W_{i,k} = P(t_i, z_k) \quad H_{k,j} = P(c_j | z_k)$$

Cost Function

$$J_{NMF-KL}(V||WH) = \sum_{i,j} V_{ij} log \frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij}$$

find W and H which minimize the whole equation

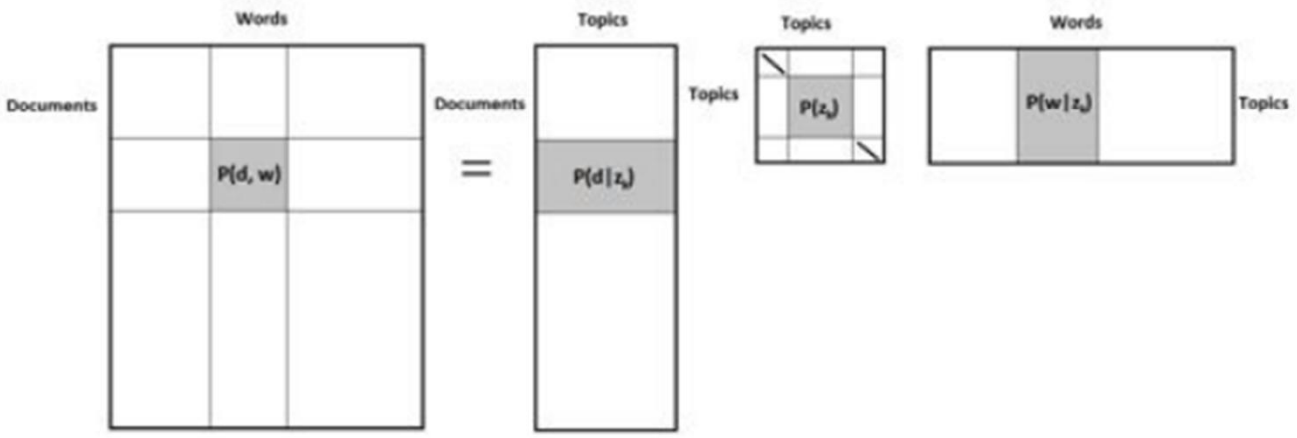
constraint: $W, H \geq 0$

Non-Negative Matrix Factorization

Symmetric parametrization: $P(d, w) = \sum_k P(d|z_k)P(z_k)P(w|z_k)$

1. Create diagonal matrix B where $B_{kk} = \sum_j H_{kj}$
2. Create diagonal matrix A where $A_{kk} = \sum_i W_{ki}$
3. $WH = (WA)(A^{-1}B)(B^{-1}H)$

WA	$(I \times K)$	$P(d\|z)$	$\sum_i P(d_i\|z) = 1$
$A^{-1}B$	$(K \times K)$	$P(z)$	$\sum_k P(z_k) = 1$
$B^{-1}H$	$(K \times J)$	$P(w\|z)$	$\sum_j P(w_j\|z) = 1$

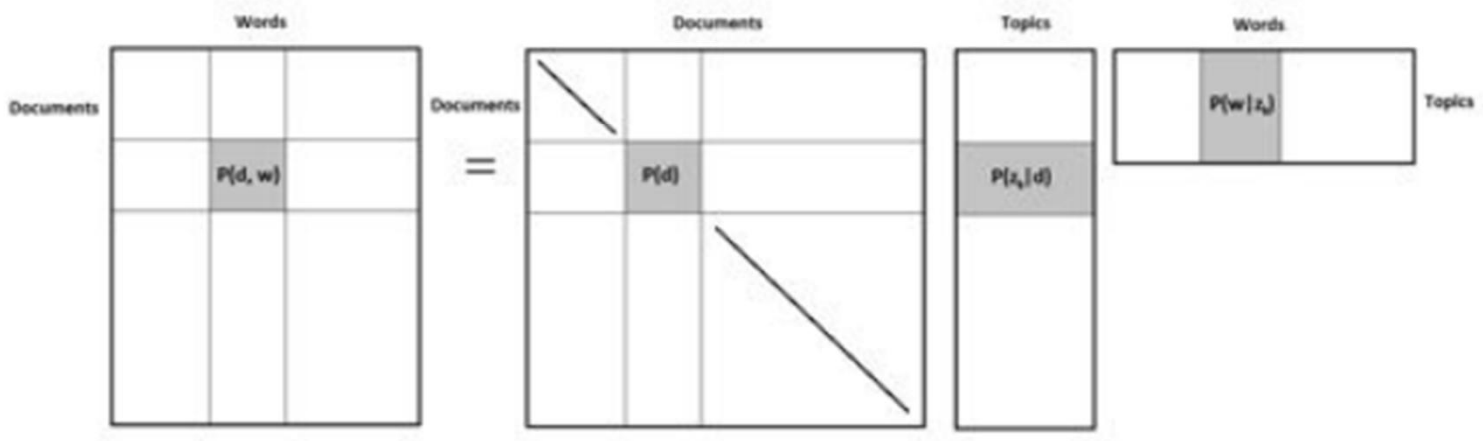


Non-Negative Matrix Factorization

Asymmetric parametrization $P(d, w) = \sum_k P(d)P(z_k|d)P(w|z_k)$

1. Create diagonal matrix B where $B_{kk} = \sum_j H_{kj}$
2. Create diagonal matrix A where $A_{ii} = \sum_k (WB)_{ik}$
3. $WH = (A)(A^{-1}WB)(B^{-1}H)$

A	$(I \times I)$	$P(d)$	$\sum_i P(d_i) = 1$
$A^{-1}WB$	$(I \times K)$	$P(z\|d)$	$\sum_k P(z_k\|d) = 1$
$B^{-1}H$	$(K \times J)$	$P(w\|z)$	$\sum_j P(w_j\|z) = 1$



Non-Negative Matrix Factorization

Asymmetric parametrization: inducing latent meaning

Representation for a target word t_i with a context feature c_j

$$v(t_i, c_j) = \left(P(z_1|t_i, c_j), \dots, P(z_k|t_i, c_j) \right)$$

$P(z_k|t_i, c_j)$ can be obtained as:

$$P(z_k|t_i, c_j) = \frac{P(t_i, z_k)P(c_j|z_k, t_i)}{\sum_k P(t_i, z_k)P(c_j|z_k, t_i)}$$

By assumption that $t_i \; and \; c_j$ are independent: $P(c_j|z_k, t_i) = P(c_j|z_k)$

$$P(z_k|t_i, c_j) = \frac{P(z_k|t_i)P(c_j|z_k)}{\sum_k P(z_k|t_i)P(c_j|z_k)} = \frac{(A^{-1}WB)_{ik}(B^{-1}H)_{kj}}{\sum_k (A^{-1}WB)_{ik}(B^{-1}H)_{kj}}$$

Latent senses: LDA

LDA: obtain data from a generative probabilistic process.

- for documents: hidden topic
- for words: hidden meaning
- put new data into the process: how does the document fits into the model.

Intuition: documents contain information about multiple topics.

Latent senses: Dirichlet distribution

The Dirichlet distribution is a K-dimensional distribution K ≥ 2

with parameters $a = (a_1 \dots a_k), a_1 \dots a_k > 0$

with density function

$$f(x_1, ..., x_K; \alpha_1, ..., \alpha_K) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$$

where $B(a)$ is a normalizing constant

$$B(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\prod_{i=1}^{K} \alpha_i)}$$

parameters control mean and variance of X

$$E[X_i] = \frac{\alpha_i}{\sum_{i=1}^{K} \alpha_k}$$

$$Var[X_i] = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}$$

Latent Senses: Dirichlet Distribution

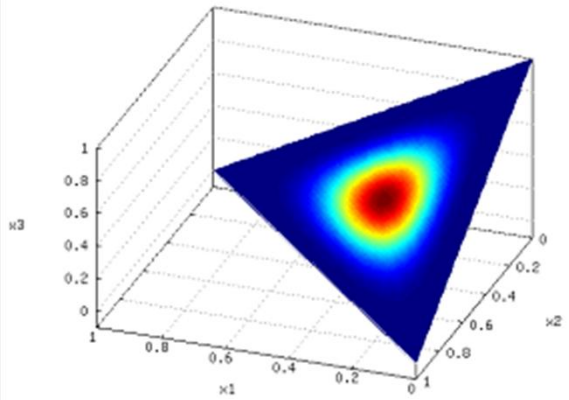


Figure A.1: $\alpha = (5, 5, 5)$

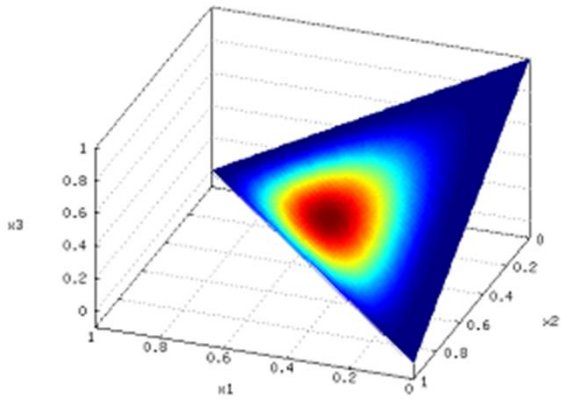


Figure A.2: $\alpha = (4, 3, 2)$

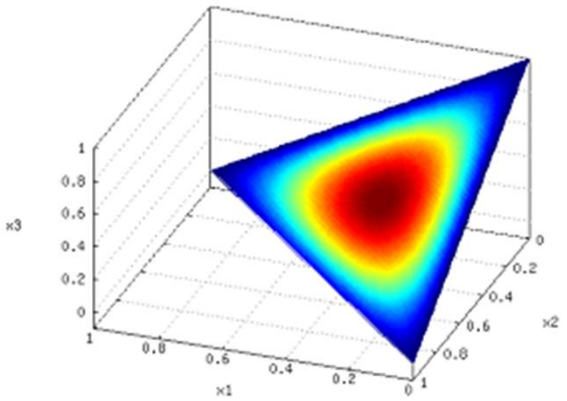


Figure A.3: $\alpha = (2, 2, 2)$

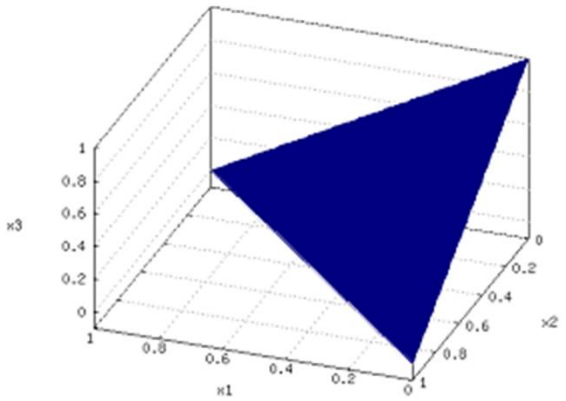
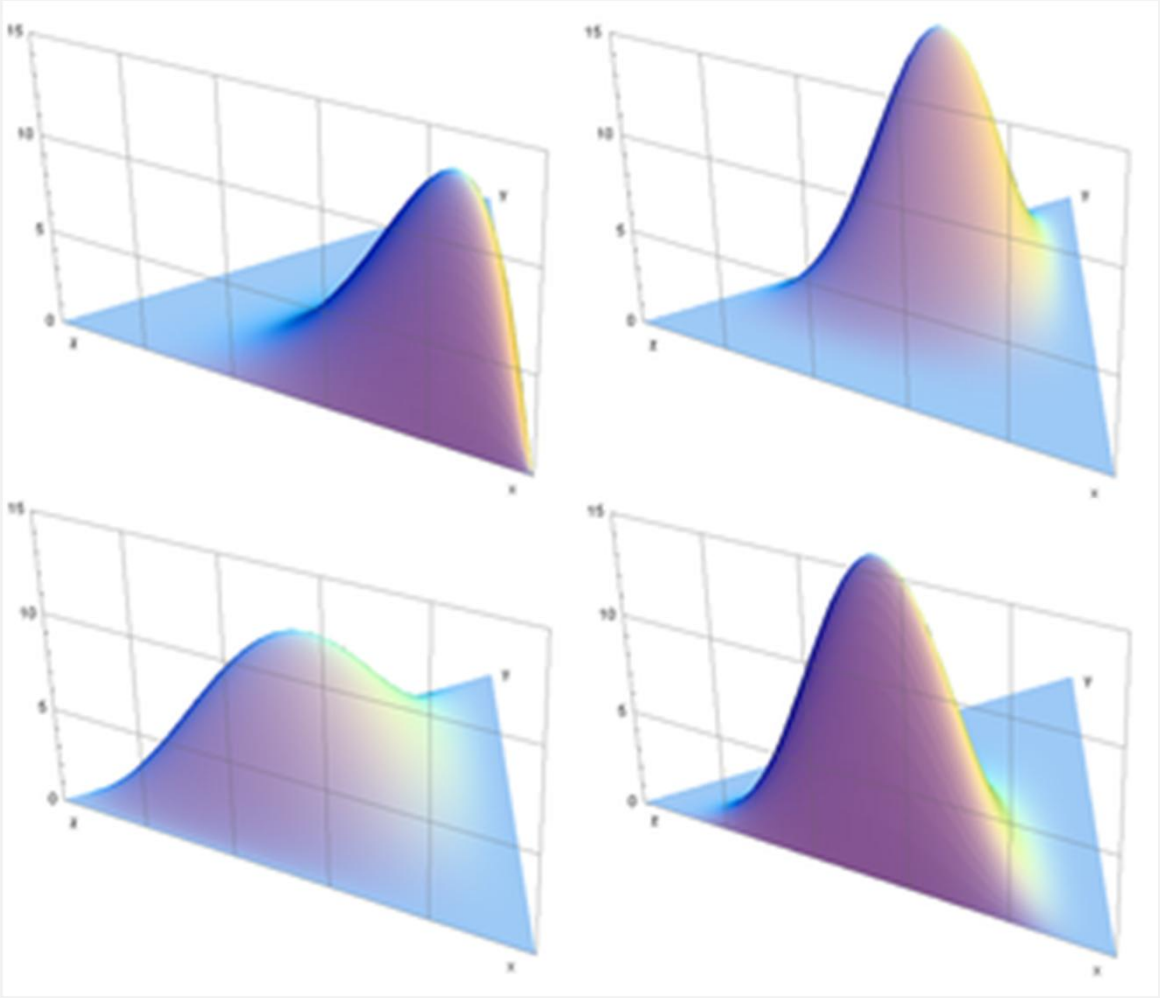


Figure A.4: $\alpha = (1, 1, 1)$

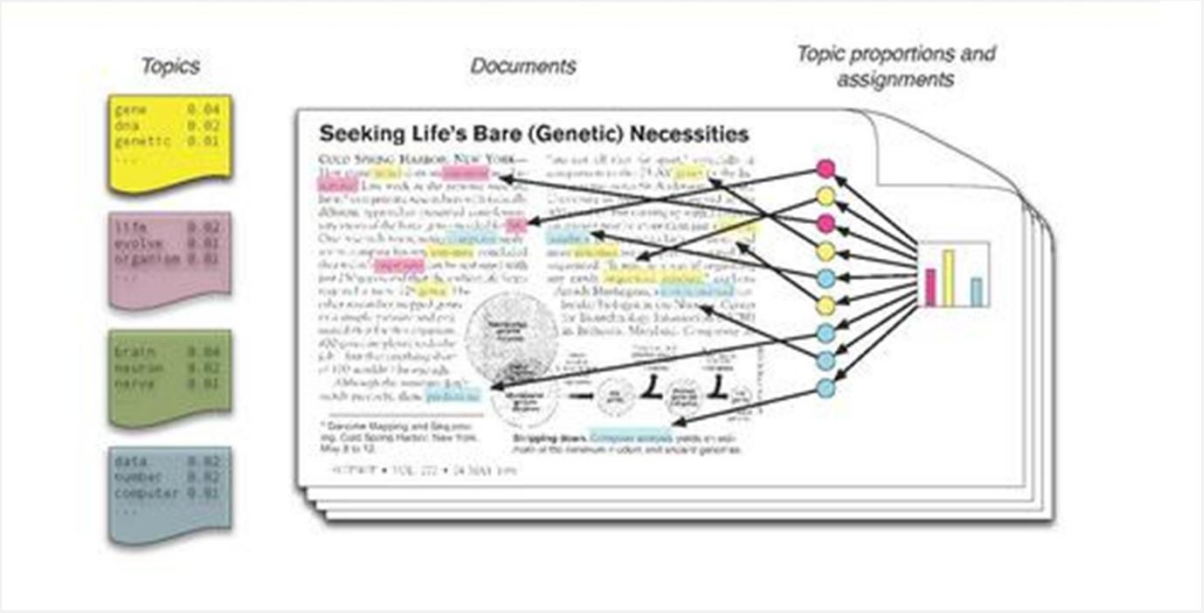
Latent Senses: Dirichlet Distribution



Latent senses: LDA

Assume: there are number of topics that exist outside of text data.

- each topic is a distribution over vocabulary:
- each document is a mixture of topics
- each word is taken from one of the topics

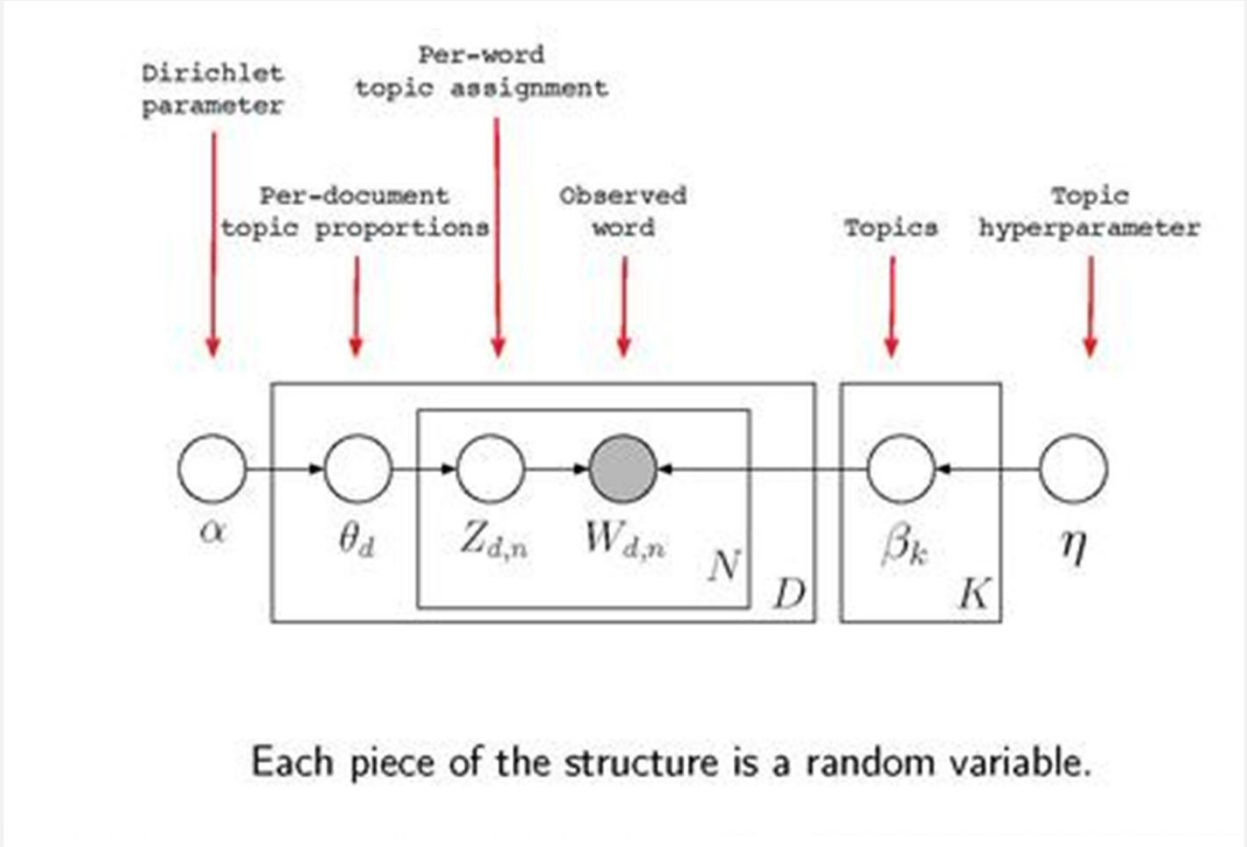


Latent senses: LDA

$\beta - distribution\ over\ terms\ of\ vocabulary\ (k\ of\ them)$

$\Theta_d - topic\ proportions, one\ for\ each\ document$

$Z_{dn} - topic\ assignment\ for\ each\ word$



Each piece of the structure is a random variable.

Latent senses: LDA

Generative model for each document:

- generate a distribution over topics Θ^d: $Dirichlet(a)$
- for each word *dn* in document *d*:
  a. generate a latent topic z_{dn} with probability P($z_{dn}|\Theta^d$)
  b. generate a distribution over words $\Phi^{z_{dn}}$: $Dirichlet(\beta)$
  c. generate a word w_{dn} with probability P($w_{dn}|\phi^{z_{dn}}$)

The joint probability of a document collection:

$$P(\mathcal{S}, \boldsymbol{\theta}, \boldsymbol{\phi}|\alpha, \beta) = \prod_{d} P(\theta^d|\alpha) \prod_{dn=d1}^{dN} P(z_{dn}|\theta^d) P(\phi^{z_{dn}}|\beta) P(w_{dn}|\phi^{z_{dn}})$$

Computing meaning representation:

$$P(z_k|t_i) = \theta_i^k$$

$$P(z_k|t_i, c_j) = \frac{P(z_k|t_i)P(c_j|z_k)}{\sum_k P(z_k|t_i)P(c_j|z_k)} = \frac{\theta_i^k \phi_j^k}{\sum_k \theta_i^k \phi_j^k}$$

Experiments and Evaluation

Experiments:

1. word-similarity task
2. lexical substitution task

Measure similarity by

scalar product:

$$sp\ (v,w) = < v,w > = \sum_i v_i w_i$$

cosine:

$$\cos(v,w) = \frac{<v,w>}{||v||||w||}$$

inverse Jensen-Shannon divergence:

$$JS(v,w) = \frac{1}{2}KL(v|m) + \frac{1}{2}KL(w|m) \qquad \text{m} = \frac{1}{2}(v+w)$$

$$KL(v|m) = \sum_i v_i \log(\frac{v_i}{w_i}) \qquad IJS(v,w) = \frac{1}{JS(v,w)}$$

Experiments and Evaluations

Task: judging similarity of two words out of context.

Evaluation: Spearman's ρ between similarity values from the modes and mean rating done by human participants. SVS as baseline

NMF: k =1000

LDA: k = 1200 $a = \frac{50}{k}$

NMF & LDA performs significantly better

($p < 0.01$) than LSA and SVS

Model	Spearman ρ
SVS	38.35
LSA	49.43
NMF	**52.99**
LDA	**53.39**
LSA_{MIX}	49.76
NMF_{MIX}	51.62
LDA_{MIX}	51.97

Table 2: Results on out of context word similarity using a simple co-occurrence based vector space model (SVS), latent semantic analysis, non-negative matrix factorization and latent Dirichlet allocation as individual models with the best parameter setting (LSA, NMF, LDA) and as mixtures (LSA_{MIX}, NMF_{MIX}, LDA_{MIX}).

Experiments and Evaluations

Task: contextualized lexical substitutions.

Evaluation: Kendall's τ rank correlation. SVS as baseline.

All models performs significantly better (p < 0.01) than SVS

Model	Kendall's τ_b
SVS	11.05
Add-SVS	12.74
Add-NMF	12.85
Add-LDA	12.33
Mult-SVS	14.41
Mult-NMF	13.20
Mult-LDA	12.90
Cont-NMF	14.95
Cont-LDA	13.71
Cont-NMF$_{\text{MIX}}$	**16.01**
Cont-LDA$_{\text{MIX}}$	**15.53**

Table 3: Results on lexical substitution using a simple semantic space model (SVS), additive and multiplicative compositional models with vector representations based on co-occurrences (Add-SVS, Mult-SVS), NMF (Add-NMF, Mult-NMF), and LDA (Add-LDA, Mult-LDA) and contextualized models based on NMF and LDA with the best parameter setting (Cont-NMF, Cont-LDA) and as mixtures (Cont-NMF$_{\text{MIX}}$, Cont-LDA$_{\text{MIX}}$).

Experiments and Evaluations

Evaluation:

Results of lexical substitutions for different parts of speech:

Model	Adv	Adj	Noun	Verb
SVS	22.47	14.38	09.52	7.98
Add-SVS	22.79	14.56	11.59	10.00
Mult-SVS	22.85	16.37	13.59	11.60
Cont-NMF$_{\text{MIX}}$	**26.13**	**17.10**	15.16	**14.18**
Cont-LDA$_{\text{MIX}}$	21.21	16.00	**16.31**	13.67

Table 4: Results on lexical substitution for different parts of speech with a simple semantic space model (SVS), two compositional models (Add-SVS, Mult-SVS), and contextualized mixture models with NMF and LDA (Cont-NMF$_{\text{MIX}}$, Cont-LDA$_{\text{MIX}}$), using Kendall's τ_b correlation coefficient.

Summing up

Paper proposes a solution to unsupervised context-sensitive word disambiguation.

Main idea: represent words out or in context as a probability distribution over set of induced latent senses.

Proposed methods include usage of NMF and LDA method for latent senses induction

Bibliography

Georgiana Dinu & Mirella Lapata. 2010. Measuring Distributional Similarity in Context. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 1162–1172*

Georgiana Dinu. 2011. *Word meaning in context: A probabilistic model and its application to Question Answering.* PhD thesis.

Wikipedia. 20 May 2012. Dirichlet Distributions. http://en.wikipedia.org/wiki/Dirichlet_distribution