

# LSA.308 Computational Psycholinguistics, Class 6

LSA (Landauer and Dumais, 1997) and LDA (Griffiths et al., 2007)

July 24, 2007

## 1 Introduction

Big questions:

- How do children learn so many words every day for years on end?
- What is the right way to formalize semantic spaces of sense/topicality/association?

## 2 Latent Semantic Analysis

### 2.1 Model basics

- Start with a document/word matrix:

$$\begin{array}{c} \text{Words} \\ \left( \begin{array}{cccc} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mn} \end{array} \right) \end{array} = M$$

Contexts

- For each word, calculate the entropy across documents:

$$p_{ij} = \frac{c_{ij}}{\sum_j c_{ij}}$$
$$H_i = \sum_j p_{ij} \log \frac{1}{p_{ij}}$$

Now take the add-one log of  $M$  and divide through by the entropy for each word to create our starting matrix  $X$ :

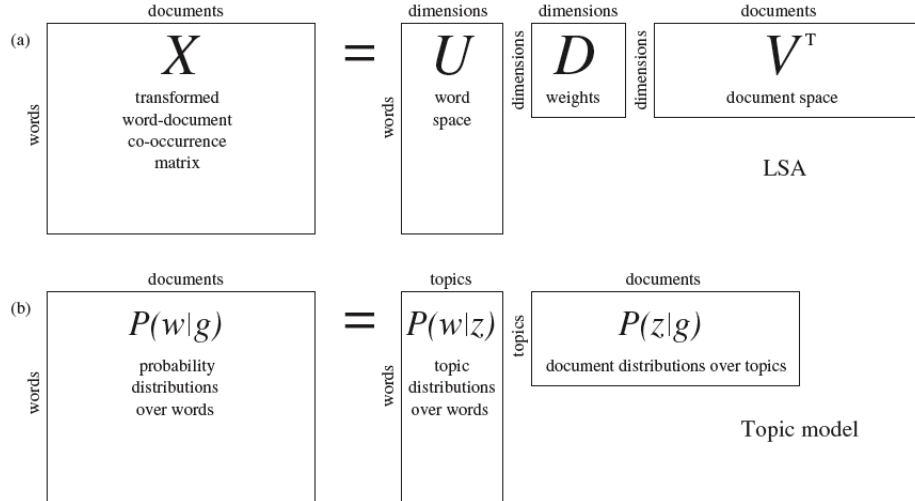


Figure 1: Matrix decompositions of Latent Semantic Analysis and the topics model

$$\begin{matrix} & & \text{Contexts} & & \\ & & \left( \begin{array}{cccc} \frac{\log(c_{11}+1)}{H_1} & \frac{\log(c_{12}+1)}{H_1} & \dots & \frac{\log(c_{1n}+1)}{H_1} \\ \frac{\log(c_{21}+1)}{H_2} & \frac{\log(c_{22}+1)}{H_2} & \dots & \frac{\log(c_{2n}+1)}{H_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\log(c_{m1}+1)}{H_m} & \frac{\log(c_{m2}+1)}{H_m} & \dots & \frac{\log(c_{mn}+1)}{H_m} \end{array} \right) & = & X \\ \text{Words} & & & & \end{matrix}$$

- Finally, find the singular-value decomposition of  $X$ :

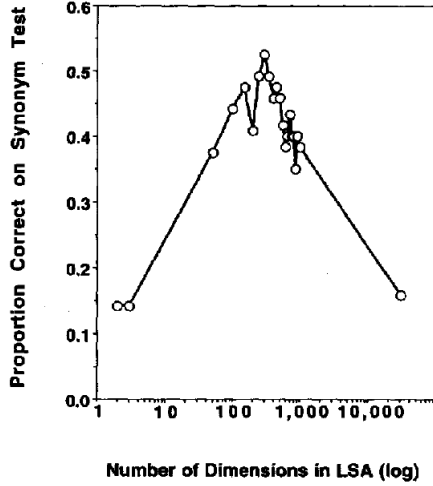
$$X = UDV$$

where  $D$  is a diagonal matrix,  $U$  has orthonormal columns (a spatial representation for words), and  $V$  has orthonormal rows (a spatial representation for documents). By convention,  $D$  is arranged so its largest values are

- Crucially, you now want to *throw out the dimensions with small singular values* (see slides)
- This is a case of *dimensionality reduction*—can be thought of as
  - Smoothing
  - Generalization
  - Inductive Bias

## 2.2 Results

- **Word knowledge acquisition.** TOEFL results based on 4.6M words:



- An intermediate (relatively small) number of dimensions ( $\sim 300$ ) performs best
  - **Direct versus indirect effects of learning.** Independently varied (i) the number of occurrences of a nonsense word in the corpus, and (ii) the size of the corpus.

Pick (2,4,8,16,32) occurrences of *hill*

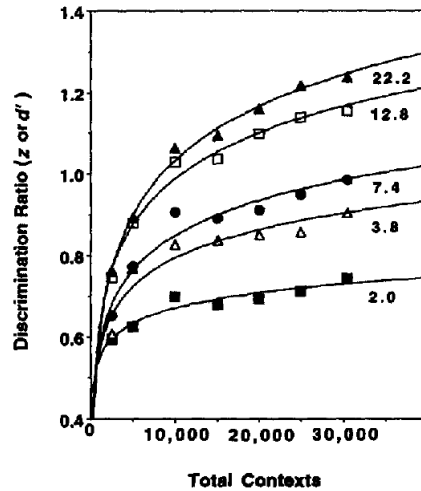
For these occurrences, *hill*  $\rightarrow$  *ghajq*

Include all these occurrences in the new corpus

Complete the corpus to size (2.5,5,10,15,20,25,30.5)K documents

Estimate the performance on TOEFL question for *ghajq*

All Results:

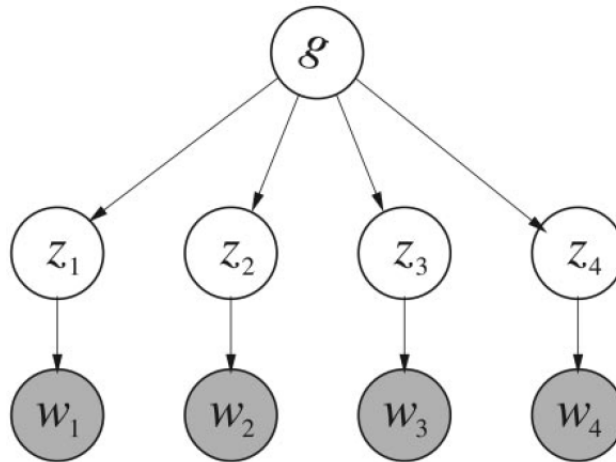


- Even after all occurrences of a word have been seen, extra context helps.

### 3 Topics Model (Latent Dirichlet Allocation; LDA)

#### 3.1 Model basics

- Simple model of semantic structure: A sequence of observed words  $w_{1\dots n}$  has some latent structure  $l$ , consisting of:
  1. the *gist*  $g$  of that sequence of words; and
  2. the *sense* of each word,  $z_{1\dots n}$ .
- Three problems of semantic knowledge:
  - **Prediction:** predict  $w_{n+1}$  from  $w_{1\dots n}$ .
  - **Disambiguation:** infer  $z$  from  $w_{1\dots n}$ .
  - **Gist extraction:** infer  $g$  from  $w_{1\dots n}$ .
- A generative model for this simple semantic structure:



- The generative model is defined as follows:
  - $P(z_i|g) \sim \text{Multinomial}(\theta^{(d)})$
  - $P(w_i|z_i) \sim \text{Multinomial}(\phi^{(z_i)})$
- The parameters  $\theta, \phi$  are estimated using Bayesian inference:

$$\begin{aligned} P(\theta, \phi | \vec{w}) &= \frac{P(\vec{w} | \theta, \phi) P(\theta, \phi)}{P(\vec{w})} \\ &= \frac{\int_{\vec{z}} P(\vec{w} | \vec{z}, \phi) P(\vec{z} | \theta) P(\theta, \phi)}{P(\vec{w})} \end{aligned}$$

where the prior distribution  $P(\phi, \theta)$  is specified by a pair of *symmetric Dirichlet distributions*.

## 3.2 Results

- Word associations (Figures 8, 9; see slides)
- Triangle inequality violations (Figure 10; see slides)

## References

Griffiths, T. L., Steyvers, M., and Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, 114(2):211–244.

Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato’s problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.