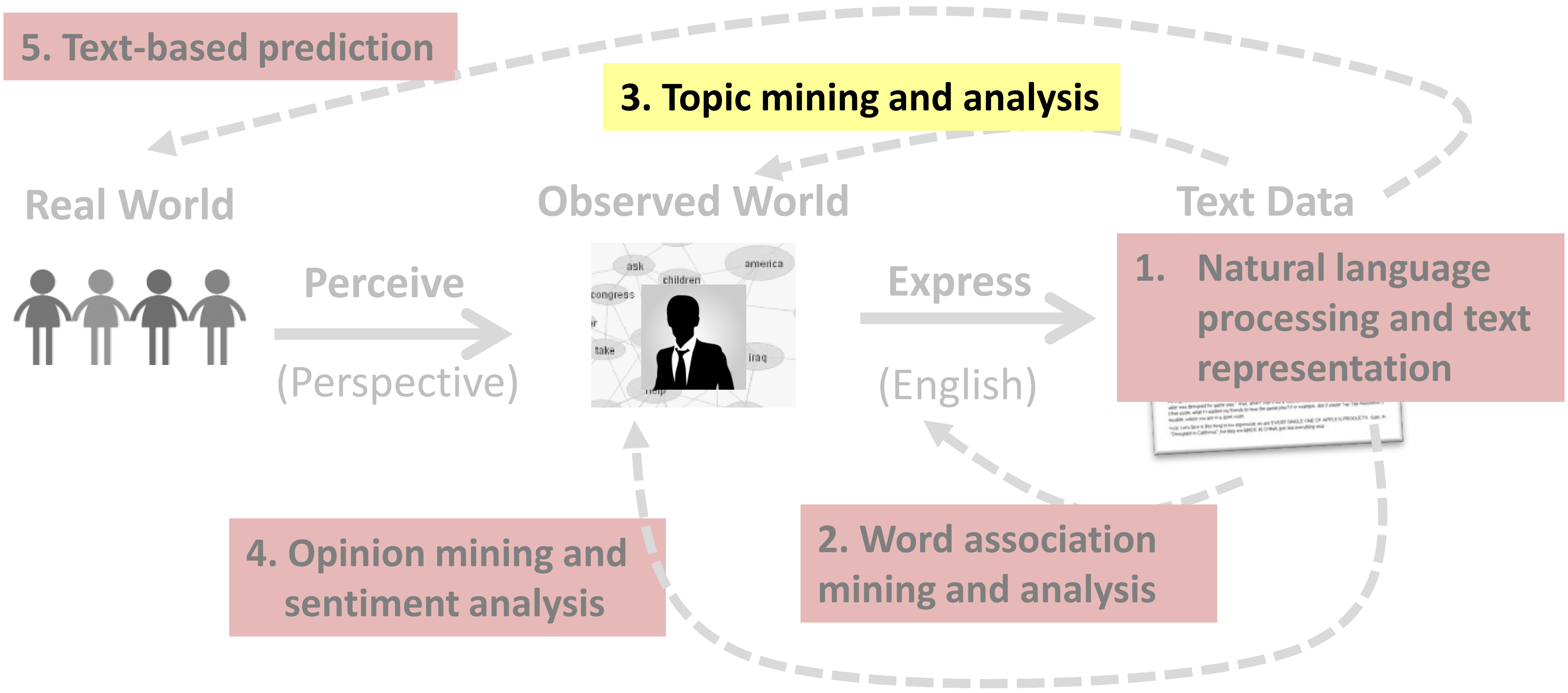
Probabilistic Topic Models: Mixture of Unigram Language Models

ChengXiang "Cheng" Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

Probabilistic Topic Models: Mixture of Unigram LMs



Factoring out Background Words

d

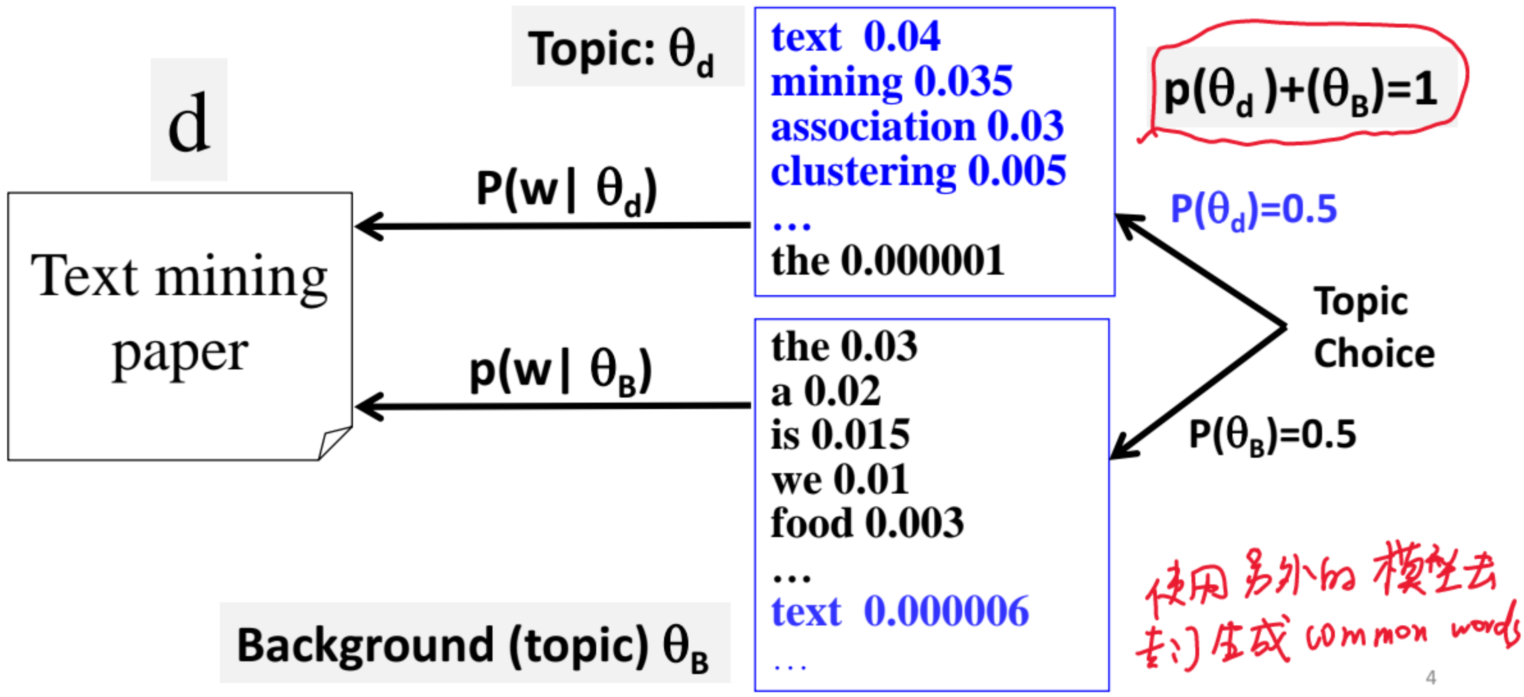
Text mining paper

p(w| θ)

the 0.031
a 0.018
…
text 0.04
mining 0.035
association 0.03
clustering 0.005
computer 0.0009
…
food 0.000001
…

How can we get rid of these common words?

Generate d Using Two Word Distributions



What's the probability of observing a word w?

该模型下生成指定单词的概率

选择模型的概率

Topic: θ_d

text 0.04
mining 0.035

$p(\theta_d)+(\theta_B)=1$

d

P("the")=$p(\theta_d)p(\text{"the"}|\theta_d) + p(\theta_B)p(\text{"the"}|\theta_B)$
= 0.5*0.000001+0.5*0.03

=0.5

the 0.000001

"the"?

"text"?

Topic
hoice

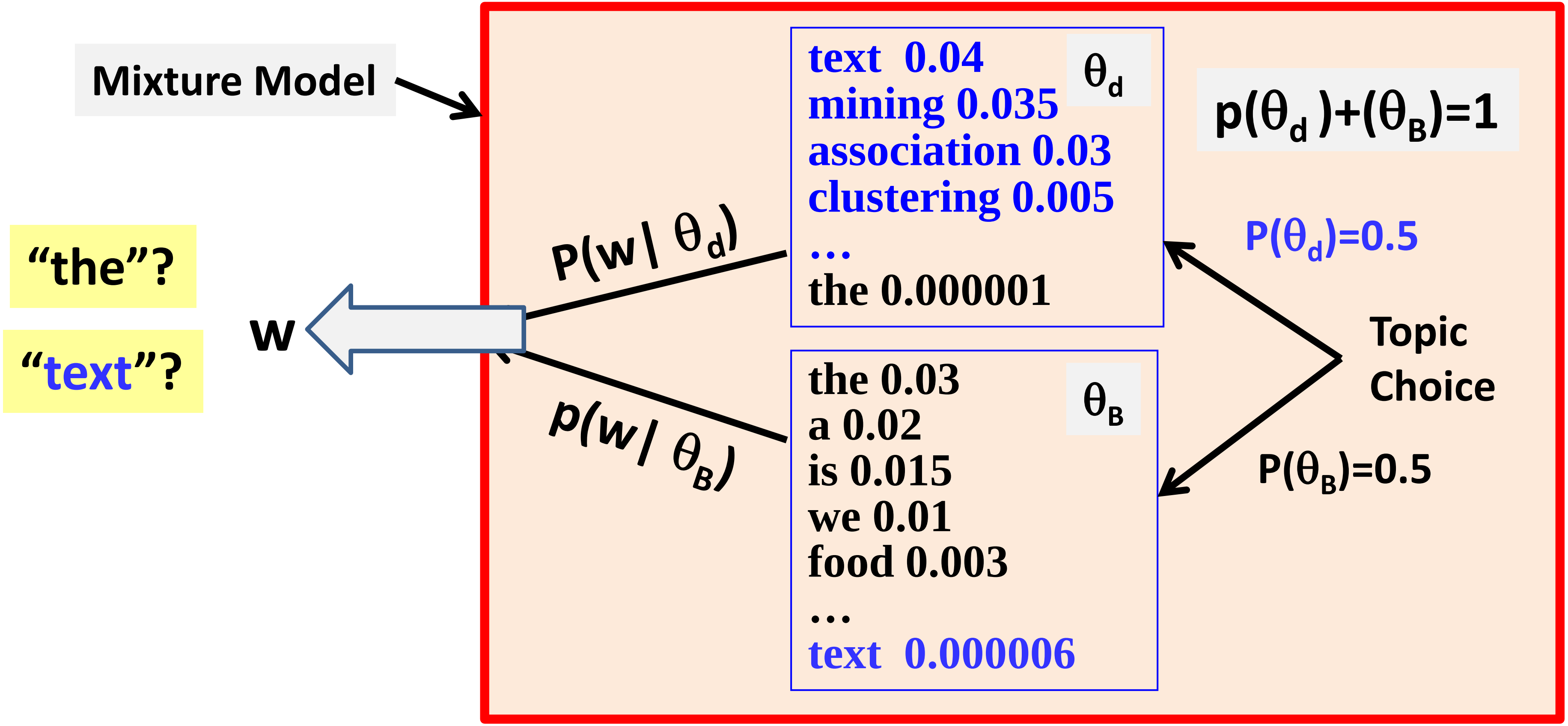
P("text")=$p(\theta_d)p(\text{"text"}|\theta_d) + p(\theta_B)\, p(\text{"text"}|\theta_B)$
= 0.5*0.04+0.5*0.000006

.5

we 0.01
food 0.003
…
text 0.000006
…

Background (topic) θ_B

The Idea of a Mixture Model



As a Generative Model…

text 0.04
mining 0.035
association 0.03
clustering 0.005

θ_d

$p(\theta_d)+(\theta_B)=1$

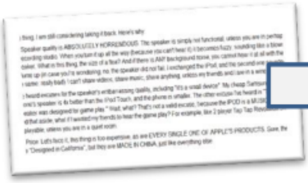
w

Formally defines the following generative model:

$$p(w)=p(\theta_d)p(w|\theta_d) + p(\theta_B)p(w|\theta_B)$$

θ_B

a 0.02



Estimate of the model "discovers"
two topics + topic coverage

…

What if $p(\theta_d)=1$ or $p(\theta_B)=1$? 只选择一种生成方法

Mixture of Two Unigram Language Models

- **Data**: Document d
- Mixture **Model**: **parameters** Λ=({p(w|θ_d)}, {p(w|θ_B)}, p(θ_B), p(θ_d))
  - Two unigram LMs: θ_d **(the topic of d);** θ_B **(background topic)**
  - Mixing weight (topic choice): **p(θ_d)+p(θ_B)=1**
- **Likelihood** function:

$$p(d \mid \Lambda) = \prod_{i=1}^{|d|} p(x_i \mid \Lambda) = \prod_{i=1}^{|d|} [p(\theta_d)p(x_i \mid \theta_d) + p(\theta_B)p(x_i \mid \theta_B)]$$
$$= \prod_{i=1}^{M} [p(\theta_d)p(w_i \mid \theta_d) + p(\theta_B)p(w_i \mid \theta_B)]^{c(w,d)}$$

- ML **Estimate:** $\Lambda^* = \arg\max_{\Lambda} p(d \mid \Lambda)$

Subject to $\sum_{i=1}^{M} p(w_i \mid \theta_d) = \sum_{i=1}^{M} p(w_i \mid \theta_B) = 1 \qquad p(\theta_d) + p(\theta_B) = 1$