Exploring Content Models for Multi-Document Summarization

Aria Haghighi

UC Berkeley

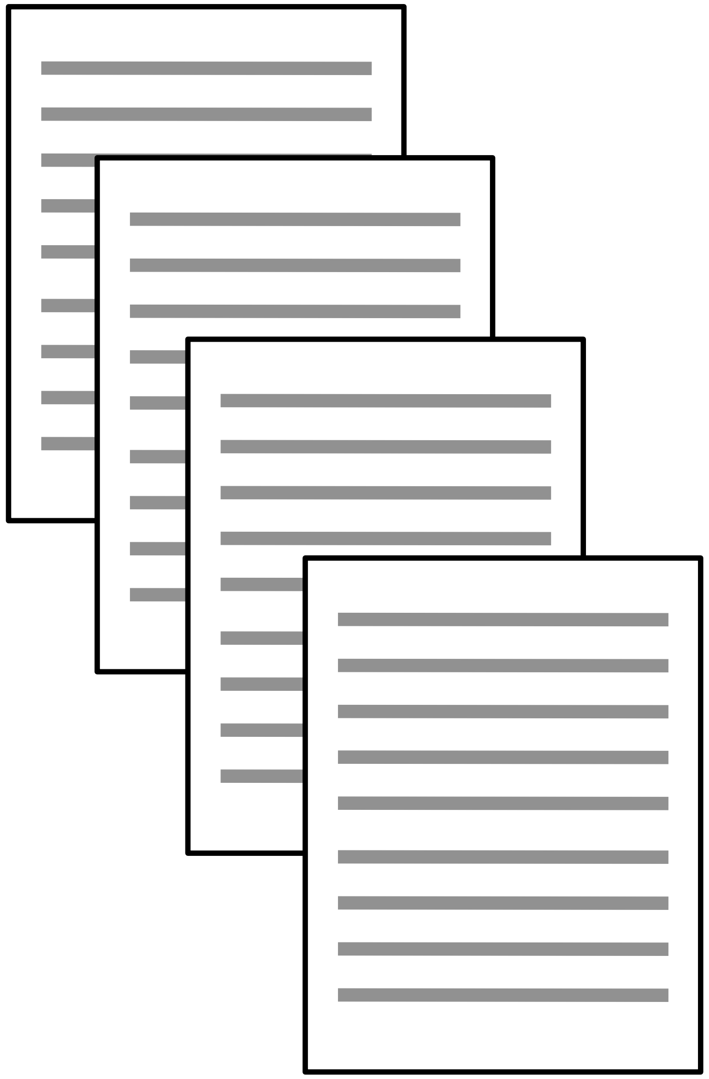
Lucy Vanderwende

MSR Redmond

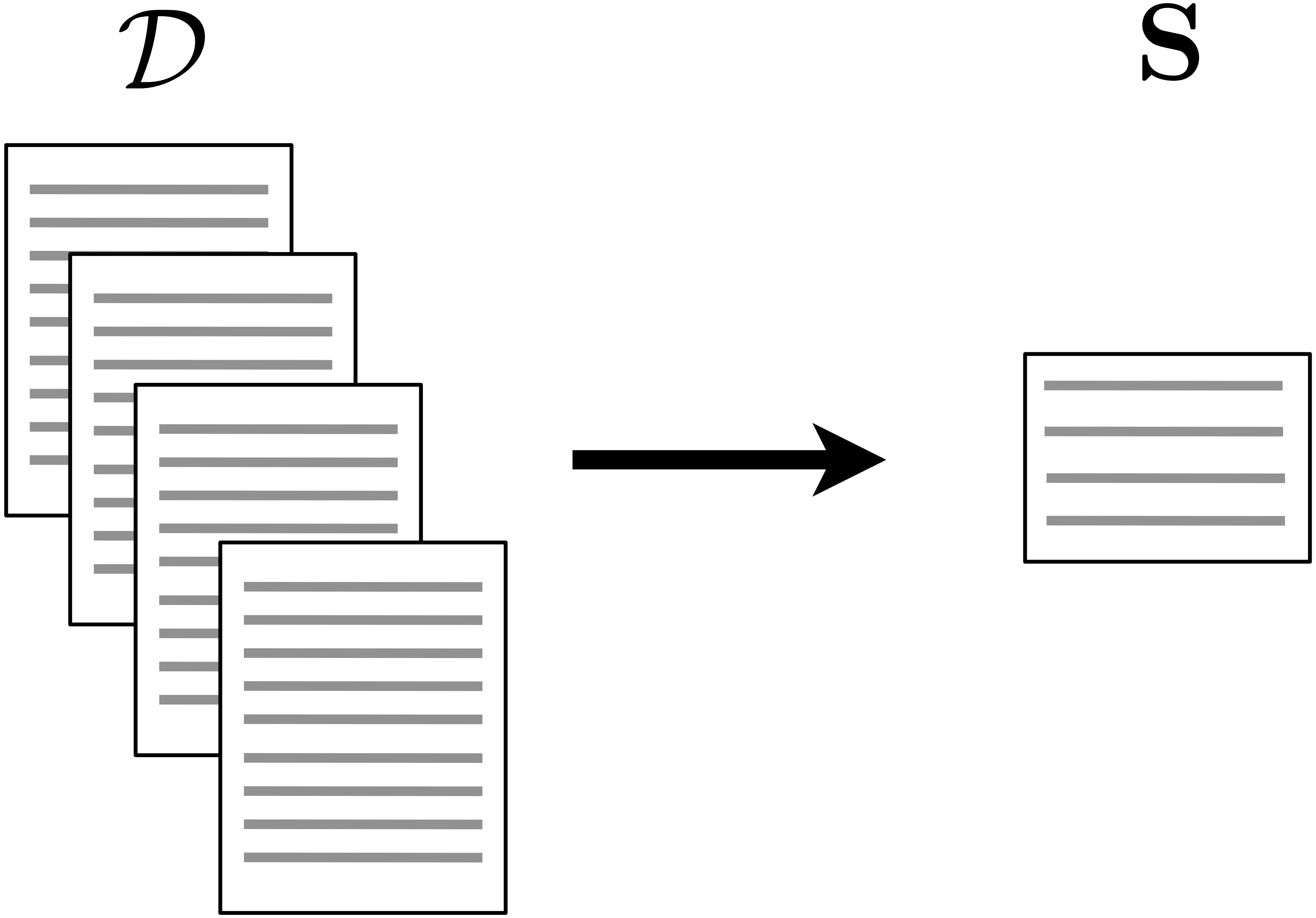
Summarization

Summarization

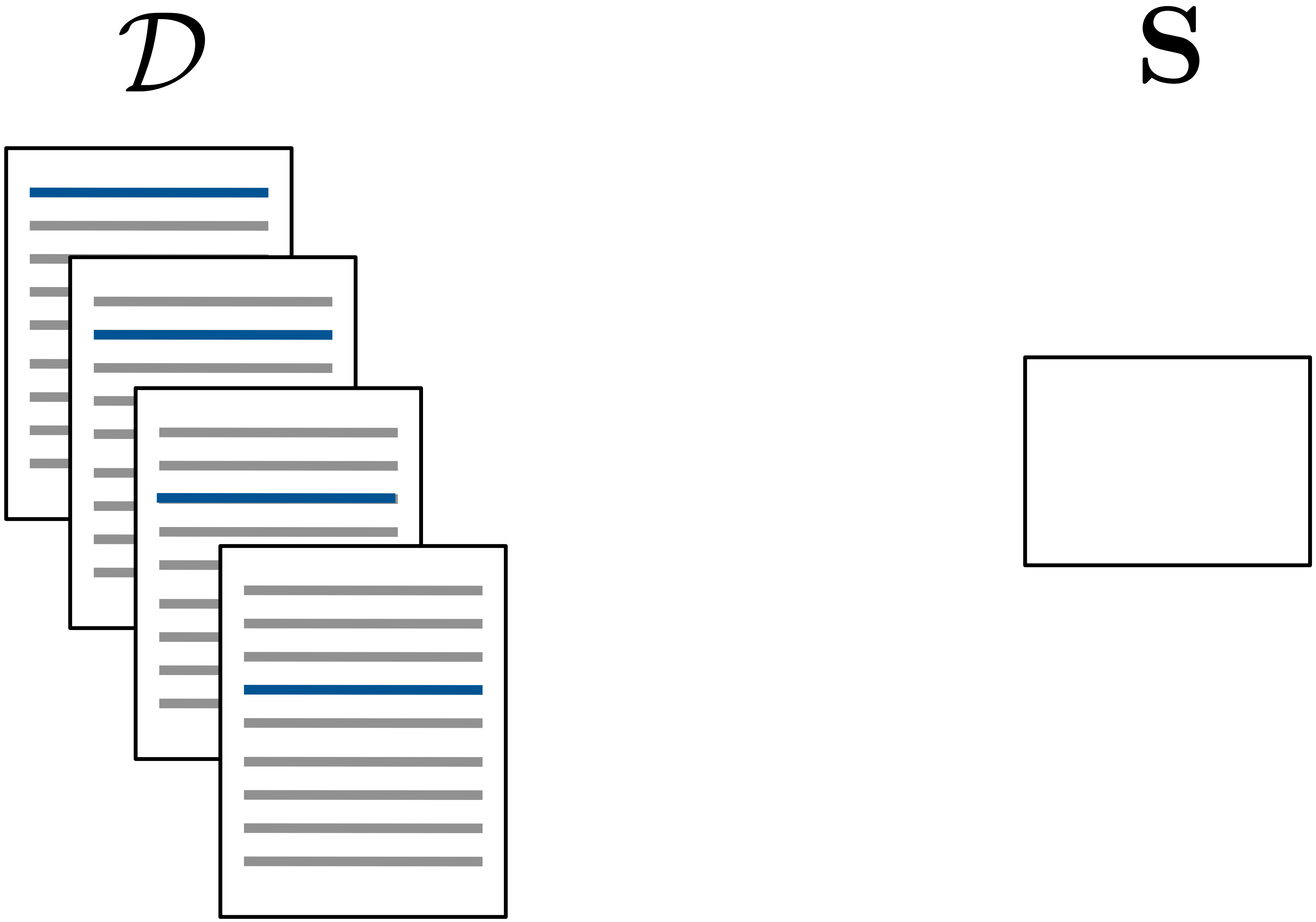
$\mathcal{D}$



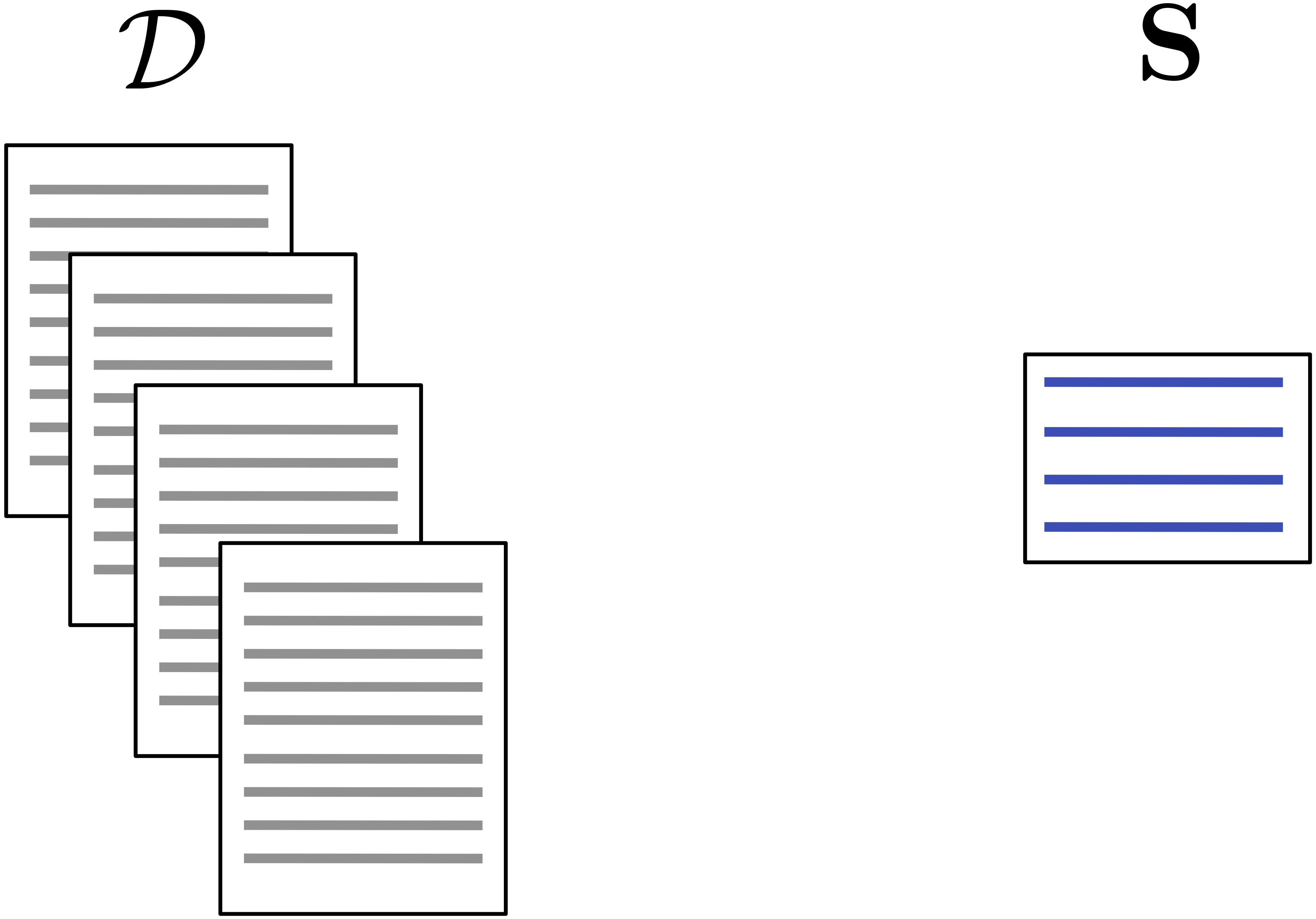
Summarization



Sentence Extraction



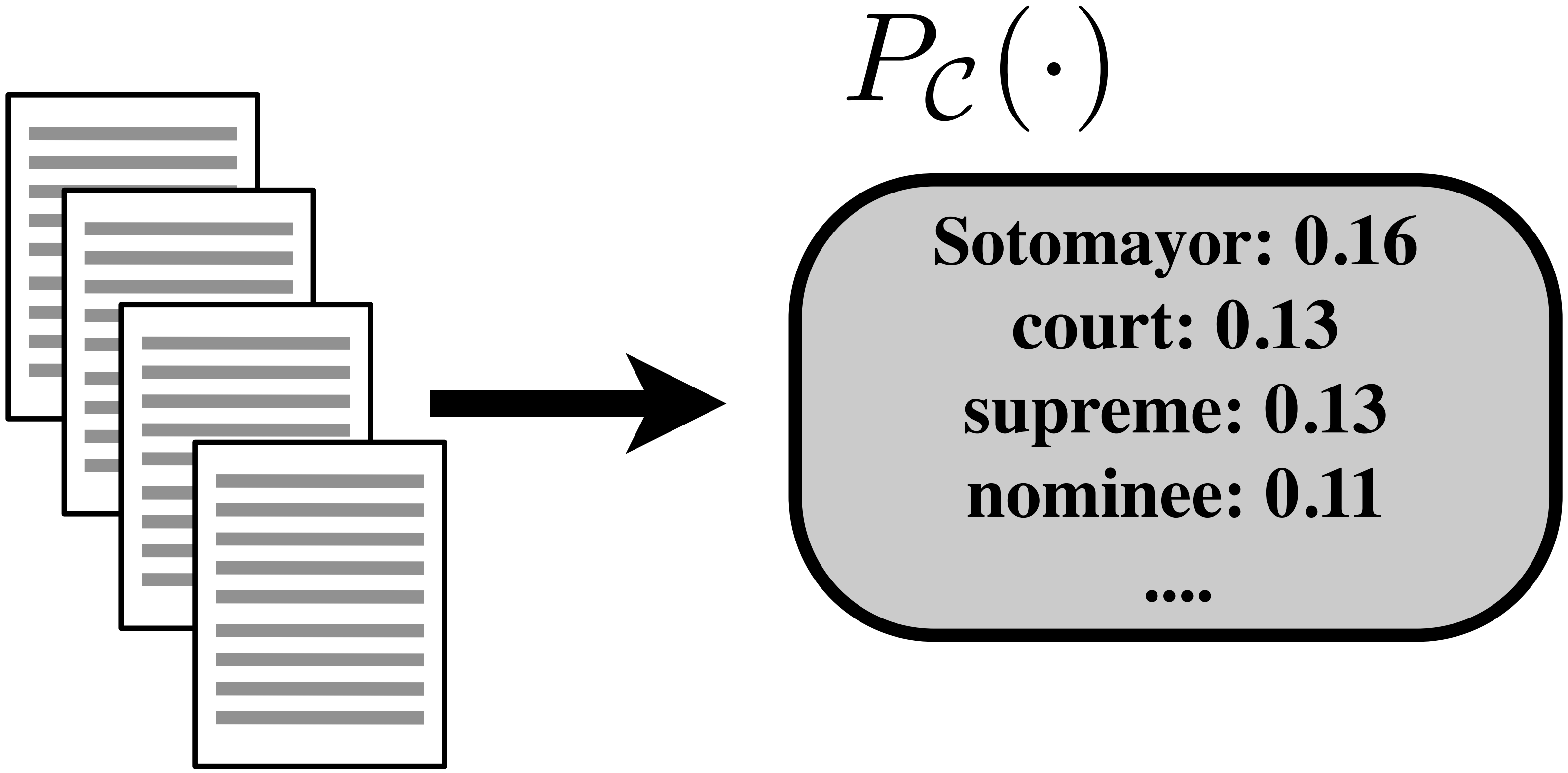
Sentence Extraction



Summarization

Summarization

Representation



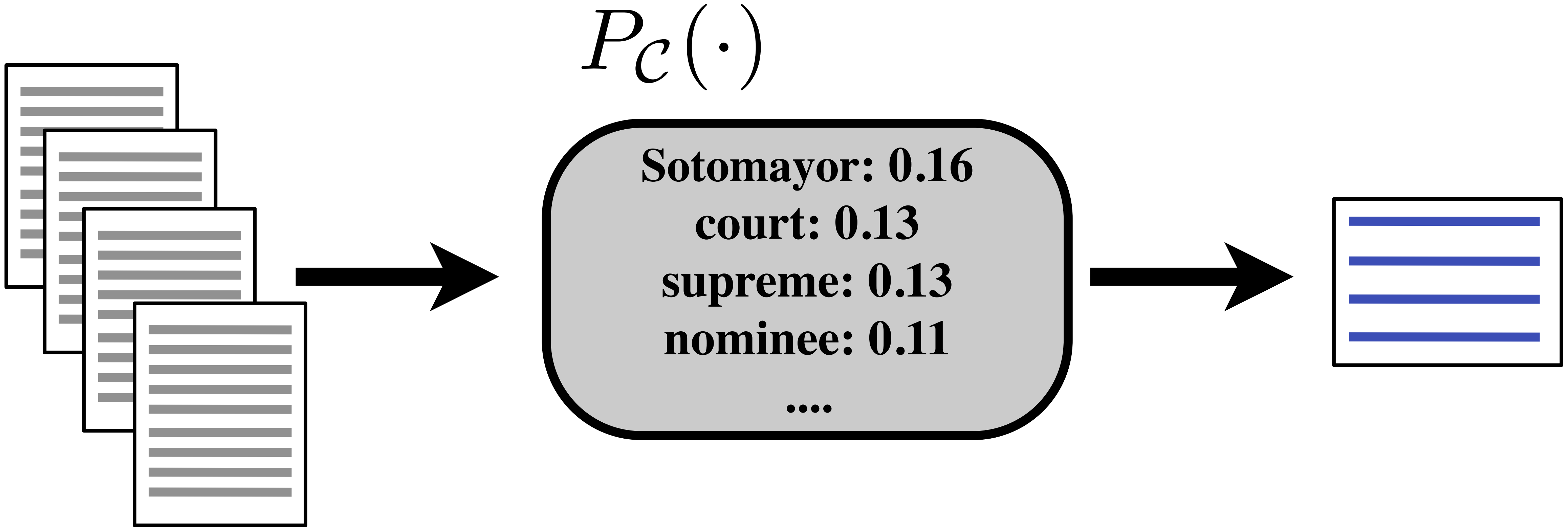
$P_C(\cdot)$

Sotomayor: 0.16
court: 0.13
supreme: 0.13
nominee: 0.11
....

Summarization

Representation

Extraction



$P_C(\cdot)$

Sotomayor: 0.16
court: 0.13
supreme: 0.13
nominee: 0.11
....

SumBasic: Representation

[Nenkova & Vanderwende] 2006

SumBasic: Representation

Simple Unigram MLE

[Nenkova & Vanderwende] 2006

SumBasic: Representation

Simple Unigram MLE

$$P_{\mathcal{C}}(w) = \hat{P}_{\mathcal{D}}(w)$$

Sotomayor: 0.15

Washington: 0.13

supreme: 0.12

Obama: 0.10

....

[Nenkova & Vanderwende] 2006

SumBasic: Extraction

$$\text{Score}(S) = \tfrac{1}{|S|} \sum_{w \in S} P_{\mathcal{C}}(w)$$

SumBasic: Extraction

Sentence Score

$$\text{Score}(S) = \frac{1}{|S|} \sum_{w \in S} P_C(w)$$

SumBasic: Extraction

Sentence Score

$$\mathrm{Score}(S) = \frac{1}{|S|} \sum_{w \in S} P_{\mathcal{C}}(w)$$

“**Obama announced** the **nomination** of **Sonia Sotomayor**”

SumBasic: Extraction

Sentence Score

$$\text{Score}(S) = \frac{1}{|S|}\sum_{w \in S} P_C(w)$$

“**Obama announced** the **nomination** of **Sonia Sotomayor**”

0.12 0.01 0.05 0.04 0.15

SumBasic: Extraction

Sentence Score

$$\text{Score}(S) = \frac{1}{|S|}\sum_{w\in S} P_C(w)$$

Score: 0.074

“**Obama announced** the **nomination** of **Sonia Sotomayor**”

0.12 0.01 0.05 0.04 0.15

SumBasic: Extraction

$$\text{Score}(S) = \tfrac{1}{|S|} \sum_{w \in S} P_{\mathcal{C}}(w)$$

→ $\mathbf{S} = \{\}$

SumBasic: Extraction

$$\mathrm{Score}(S) = \tfrac{1}{|S|} \textstyle\sum_{w \in S} P_{\mathcal{C}}(w)$$

```
$\mathbf{S} = \{\}$
→ while $words(\mathbf{S}) < L$:
```

SumBasic: Extraction

$$\text{Score}(S) = \frac{1}{|S|} \sum_{w \in S} P_{\mathcal{C}}(w)$$

```
$\mathbf{S} = \{\}$
while $words(\mathbf{S}) < L$:
    $S^* = \max_{S \notin \mathbf{S}} \text{Score}(S)$
```

SumBasic: Extraction

$$\text{Score}(S) = \frac{1}{|S|} \sum_{w \in S} P_{\mathcal{C}}(w)$$

```
$\mathbf{S} = \{\}$
while $words(\mathbf{S}) < L$:
    $S^* = \max_{S \notin \mathbf{S}} \text{Score}(S)$
    $\mathbf{S} = \mathbf{S} \cup S^*$
```

SumBasic: Extraction

$$\text{Score}(S) = \tfrac{1}{|S|} \textstyle\sum_{w \in S} P_{\mathcal{C}}(w)$$

```
$\mathbf{S} = \{\}$
while $words(\mathbf{S}) < L$:
    $S^* = \max_{S \notin \mathbf{S}} \text{Score}(S)$
    $\mathbf{S} = \mathbf{S} \cup S^*$
→   $P_{\mathcal{C}}(w) = P_{\mathcal{C}}(w)^2$, for $w \in S^*$
```

Experimental Results

Experimental Results

- DUC 2006

Experimental Results

- DUC 2006
  - 50 document sets, 25 docs each

Experimental Results

- DUC 2006
  - 50 document sets, 25 docs each
  - Max 250 tokens

Experimental Results

- DUC 2006
  - 50 document sets, 25 docs each
  - Max 250 tokens
- ROUGE-2

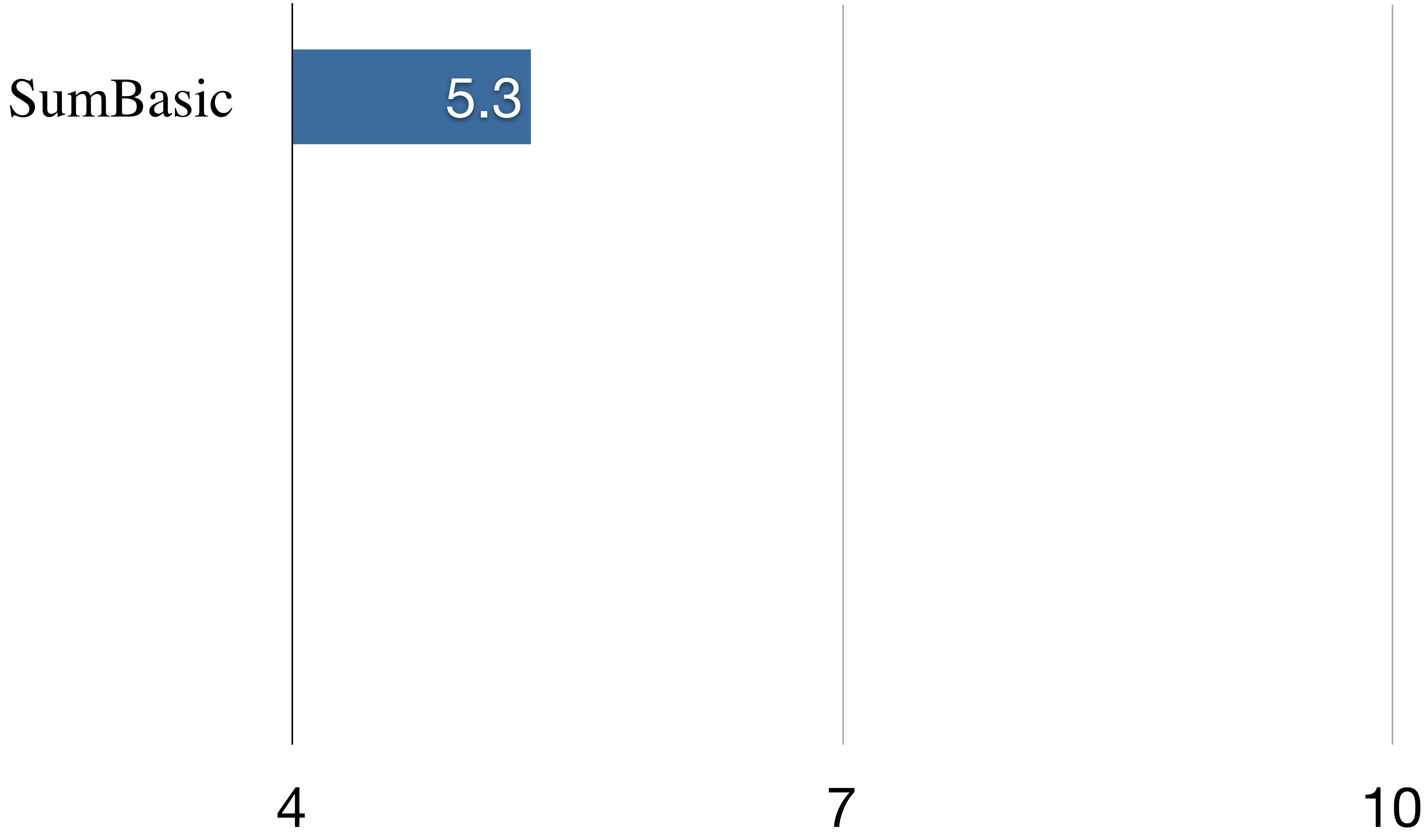
Experimental Results

- DUC 2006
  - 50 document sets, 25 docs each
  - Max 250 tokens
- ROUGE-2
  - Recall over bigrams w/o stop words against human summaries

Experimental Results

- DUC 2006
  - 50 document sets, 25 docs each
  - Max 250 tokens
- ROUGE-2
  - Recall over bigrams w/o stop words against human summaries
  - Bad at summary quality, decent at content

SumBasic: Performance



SumBasic: Extraction Issues

SumBasic: Extraction Issues

- What are we optimizing?

SumBasic: Extraction Issues

- What are we optimizing?
  - Without word length, when to stop?

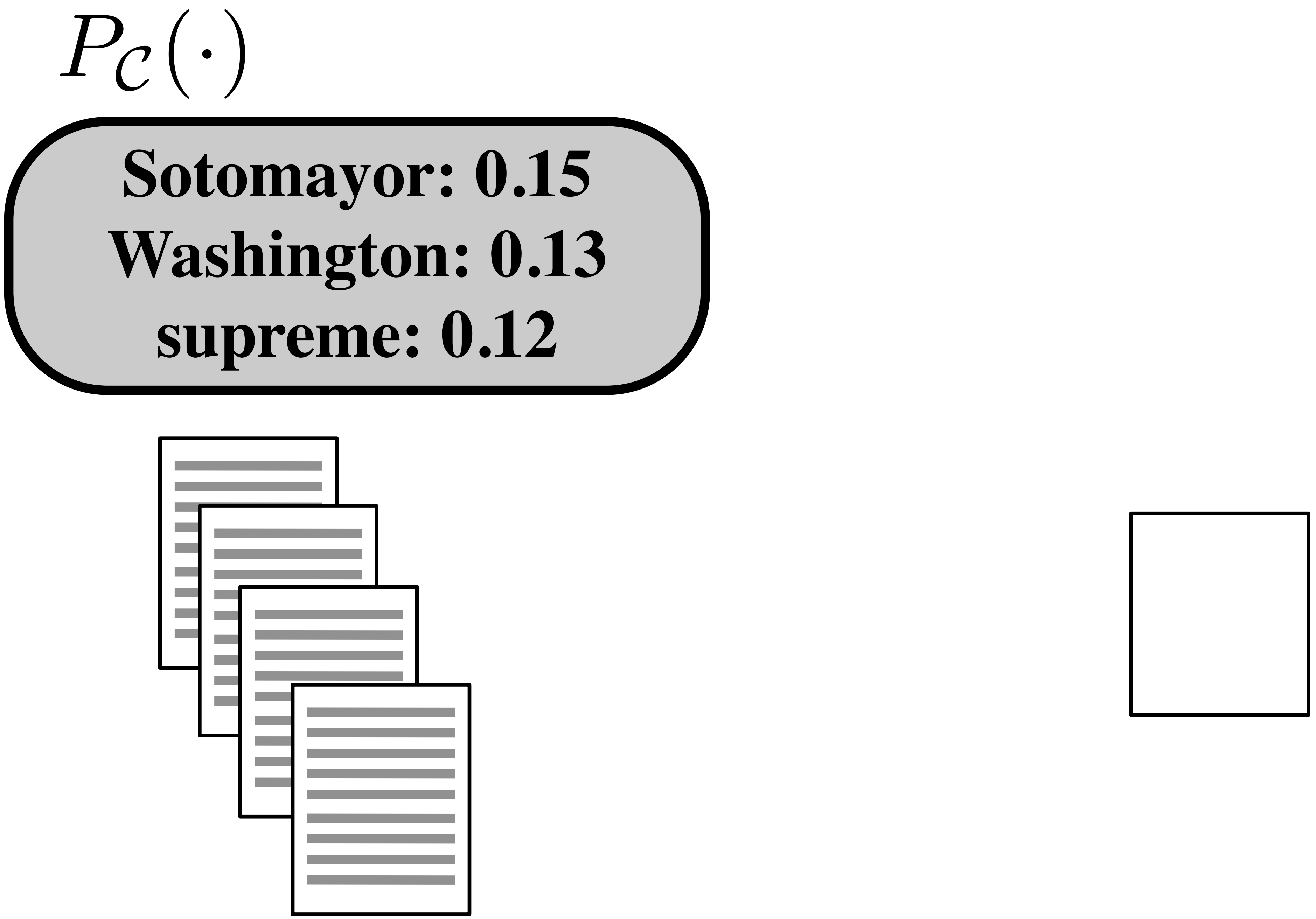
SumBasic: Extraction Issues

- What are we optimizing?
  - Without word length, when to stop?
- Not Recall Oriented

SumBasic: Extraction Issues

- What are we optimizing?
  - Without word length, when to stop?
- Not Recall Oriented
  - No direct penalty for missing freq. words

KLSum: Extraction



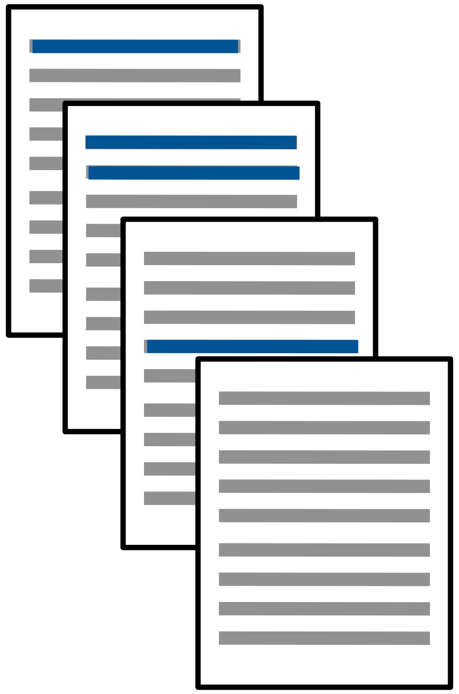
$P_C(\cdot)$

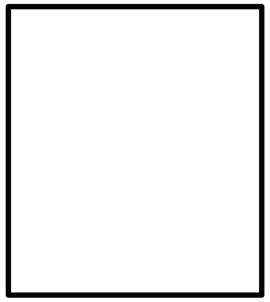
Sotomayor: 0.15
Washington: 0.13
supreme: 0.12

KLSum: Extraction

$P_C(\cdot)$

Sotomayor: 0.15
Washington: 0.13
supreme: 0.12





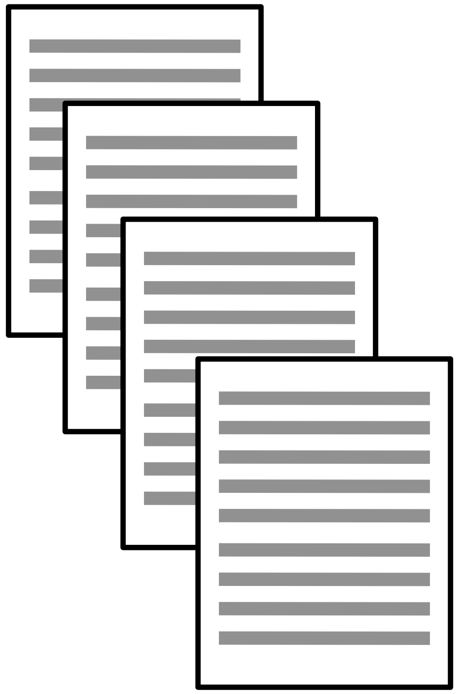
KLSum: Extraction

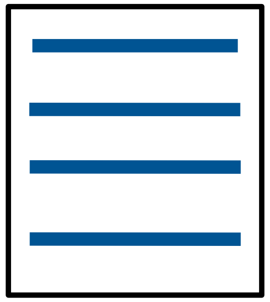
$P_{\mathcal{C}}(\cdot)$

Sotomayor: 0.15
Washington: 0.13
supreme: 0.12

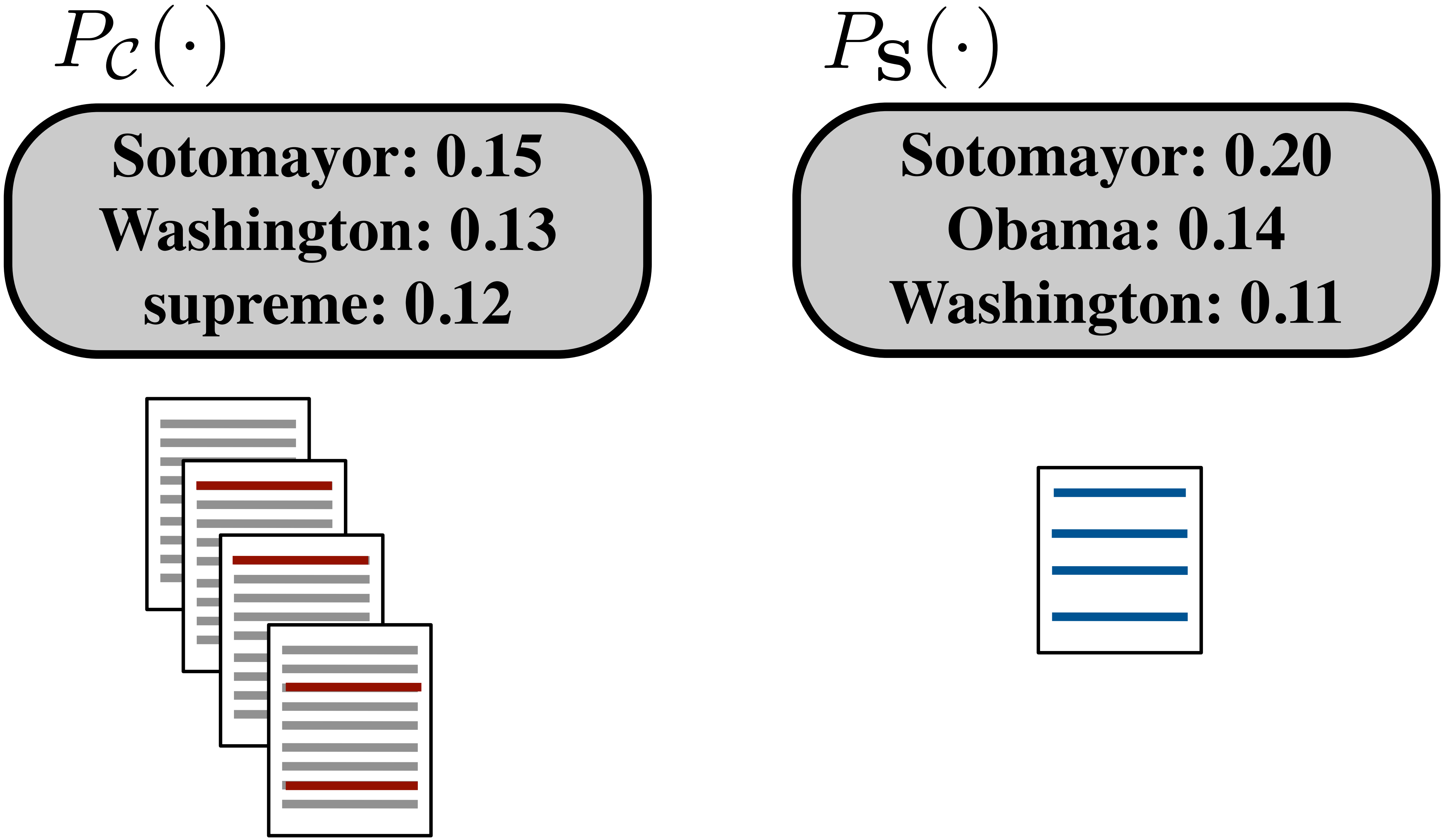
$P_{\mathbf{S}}(\cdot)$

Sotomayor: 0.20
Obama: 0.14
Washington: 0.11





KLSum: Extraction



$P_C(\cdot)$

Sotomayor: 0.15
Washington: 0.13
supreme: 0.12

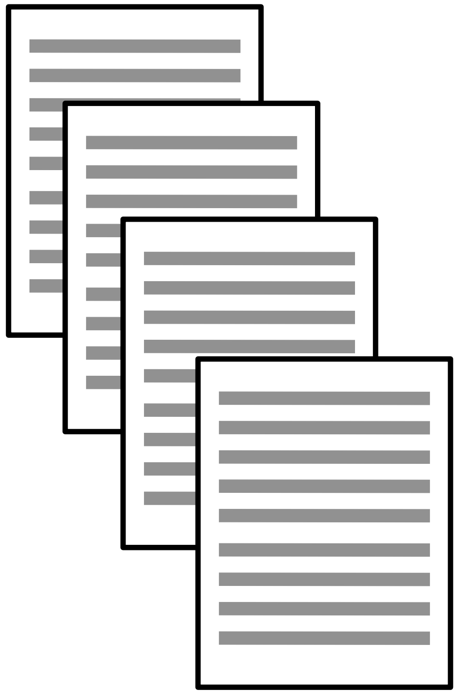
$P_{\mathbf{S}}(\cdot)$

Sotomayor: 0.20
Obama: 0.14
Washington: 0.11

KLSum: Extraction

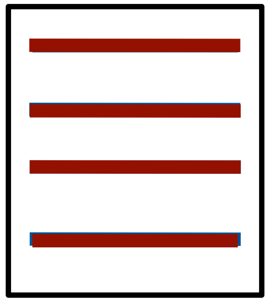
$P_C(\cdot)$

Sotomayor: 0.15
Washington: 0.13
supreme: 0.12



$P_{\mathbf{S}}(\cdot)$

Sotomayor: 0.18
Washington: 0.11
supreme: 0.10



KLSum: Extraction

$$\mathbf{S}^* = \min_{\mathbf{S}:words(\mathbf{S})\leq L} KL(P_{\mathcal{C}} \| P_{\mathbf{S}})$$

$P_{\mathcal{C}}(\cdot)$

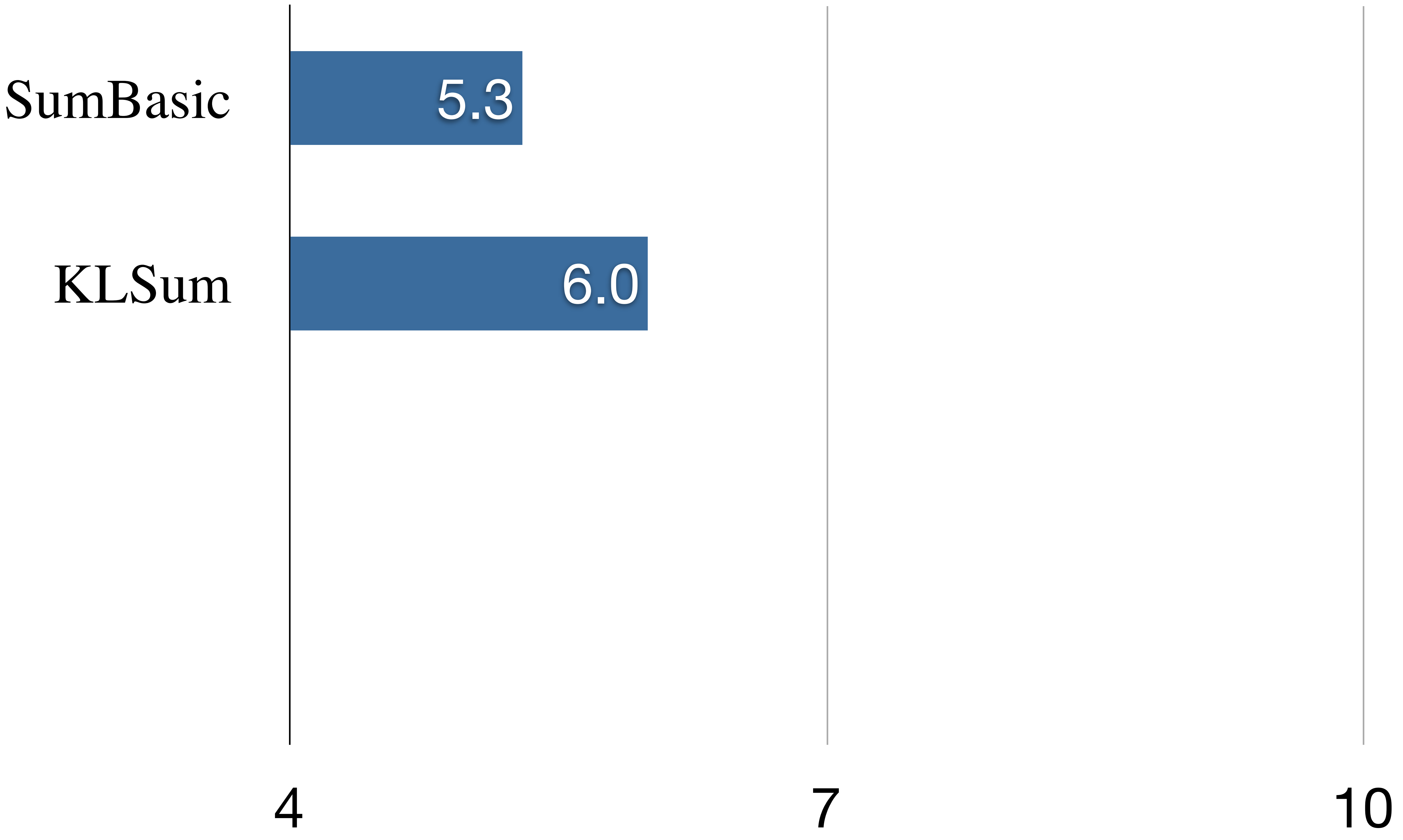
Sotomayor: 0.15
Washington: 0.13
supreme: 0.12

$P_{\mathbf{S}}(\cdot)$

Sotomayor: 0.18
Washington: 0.11
supreme: 0.10

See Paper For Details

KLSum: Performance



Improving Representation

Improving Representation

- Flexible stop words

Improving Representation

- Flexible stop words
  - e.g. “stock” in financial document sets

Improving Representation

- Flexible stop words
  - e.g. “stock” in financial document sets
- No pref. for multi-document usage

Improving Representation

- Flexible stop words
  - e.g. “stock” in financial document sets
- No pref. for multi-document usage
  - Many docs indicate content importance

Improving Representation

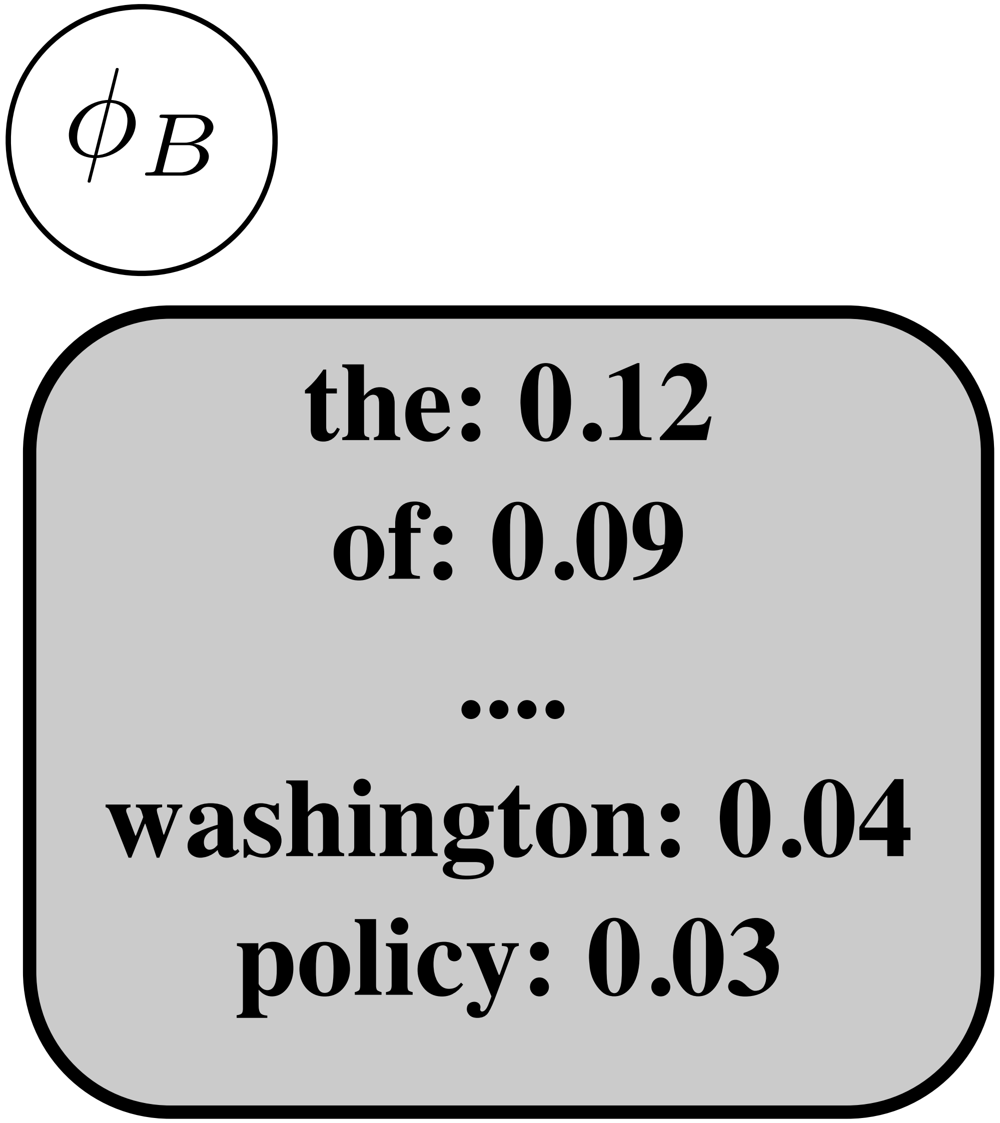
- Flexible stop words
  - e.g. “stock” in financial document sets
- No pref. for multi-document usage
  - Many docs indicate content importance
- Prefer words in early sentences

Adding Topics

Background Distribution

Adding Topics

Background Distribution



ϕ_B

the: 0.12

of: 0.09

....

washington: 0.04

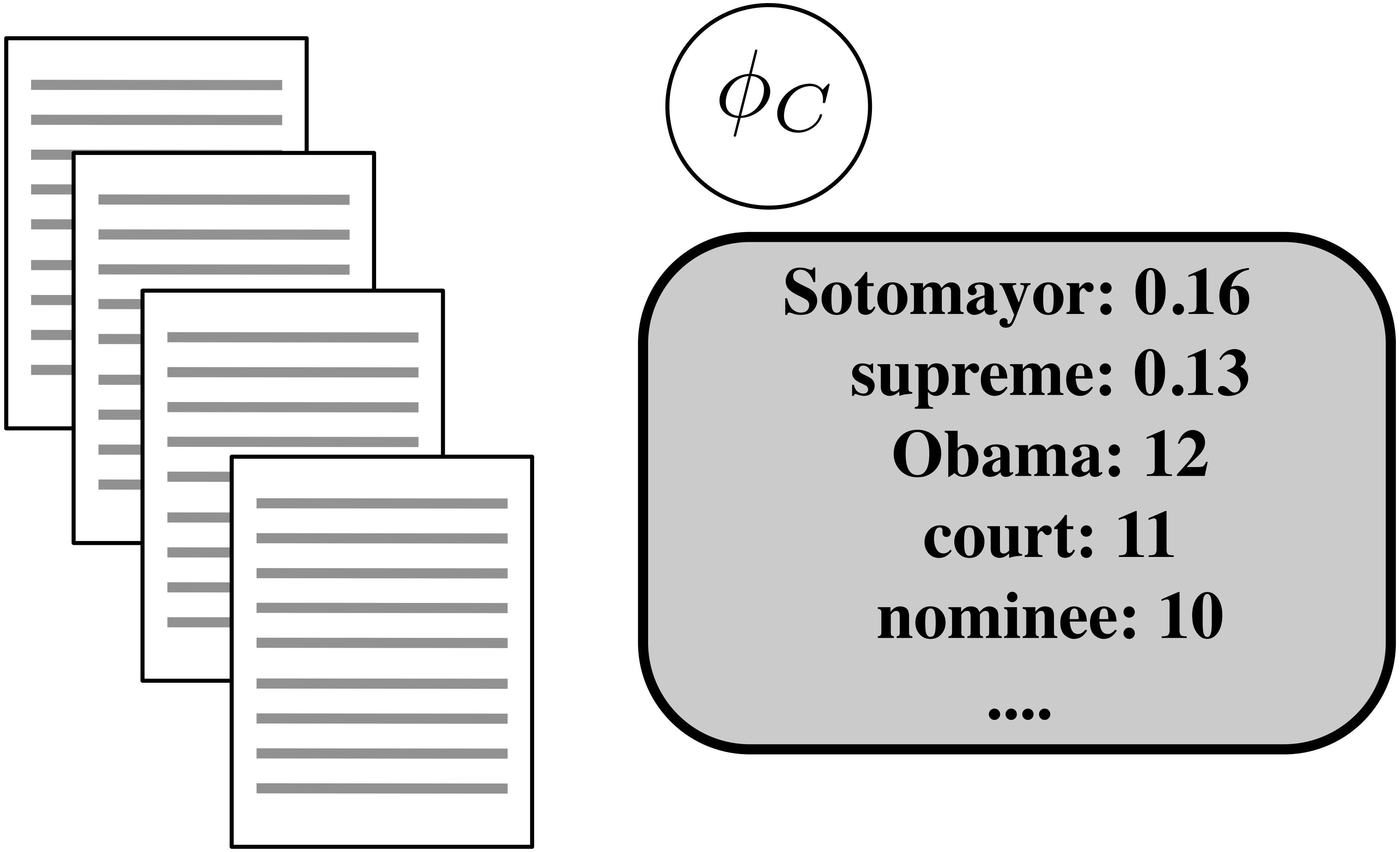
policy: 0.03

Adding Topics

<u>Content Distribution</u>

Adding Topics

Content Distribution



ϕ_C

Sotomayor: 0.16
supreme: 0.13
Obama: 12
court: 11
nominee: 10
....

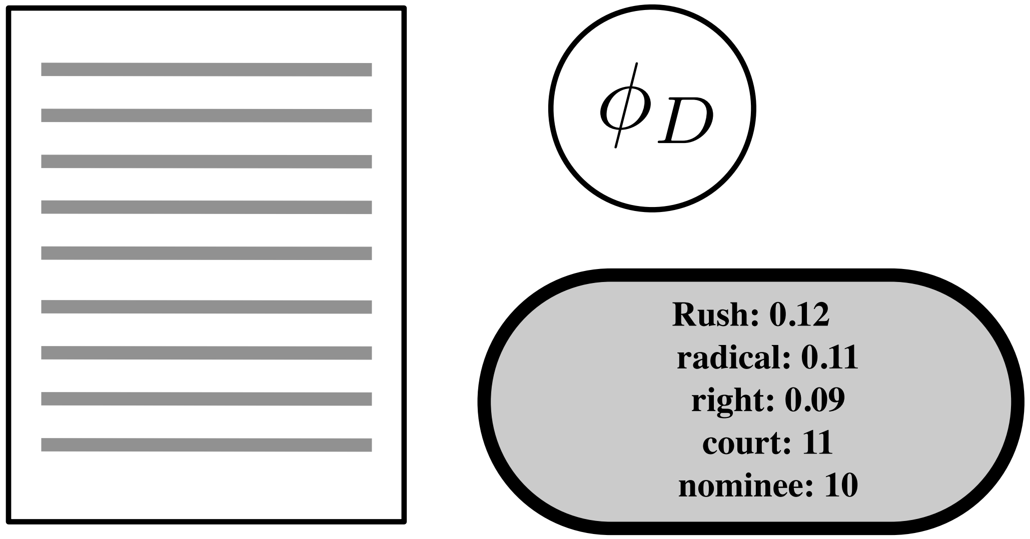
Adding Topics

Document-Specific Distribution

similar to [Daume & Marcu, 2006]

Adding Topics

Document-Specific Distribution



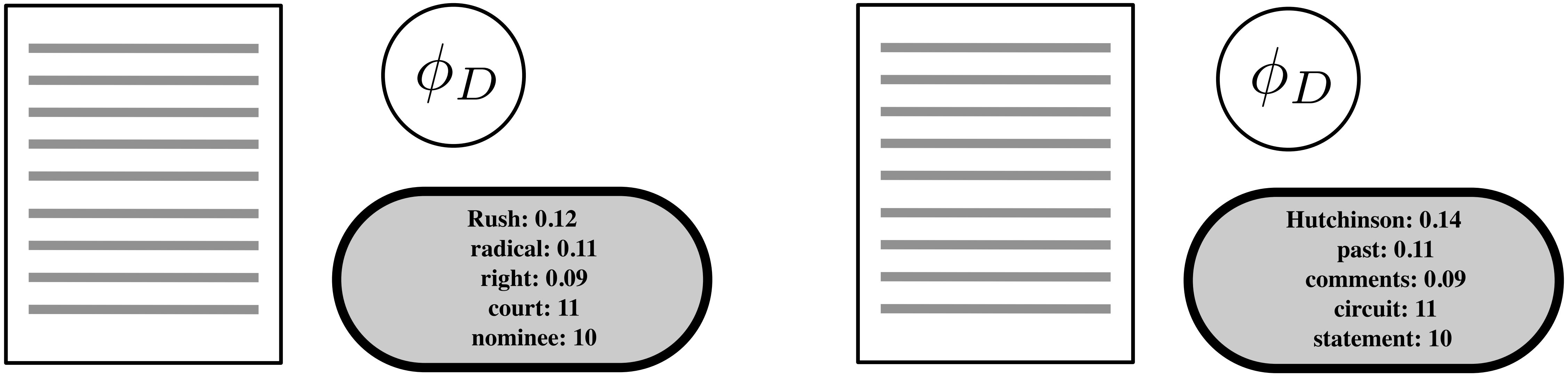
ϕ_D

Rush: 0.12
radical: 0.11
right: 0.09
court: 11
nominee: 10

similar to [Daume & Marcu, 2006]

Adding Topics

Document-Specific Distribution



ϕ_D

Rush: 0.12
radical: 0.11
right: 0.09
court: 11
nominee: 10

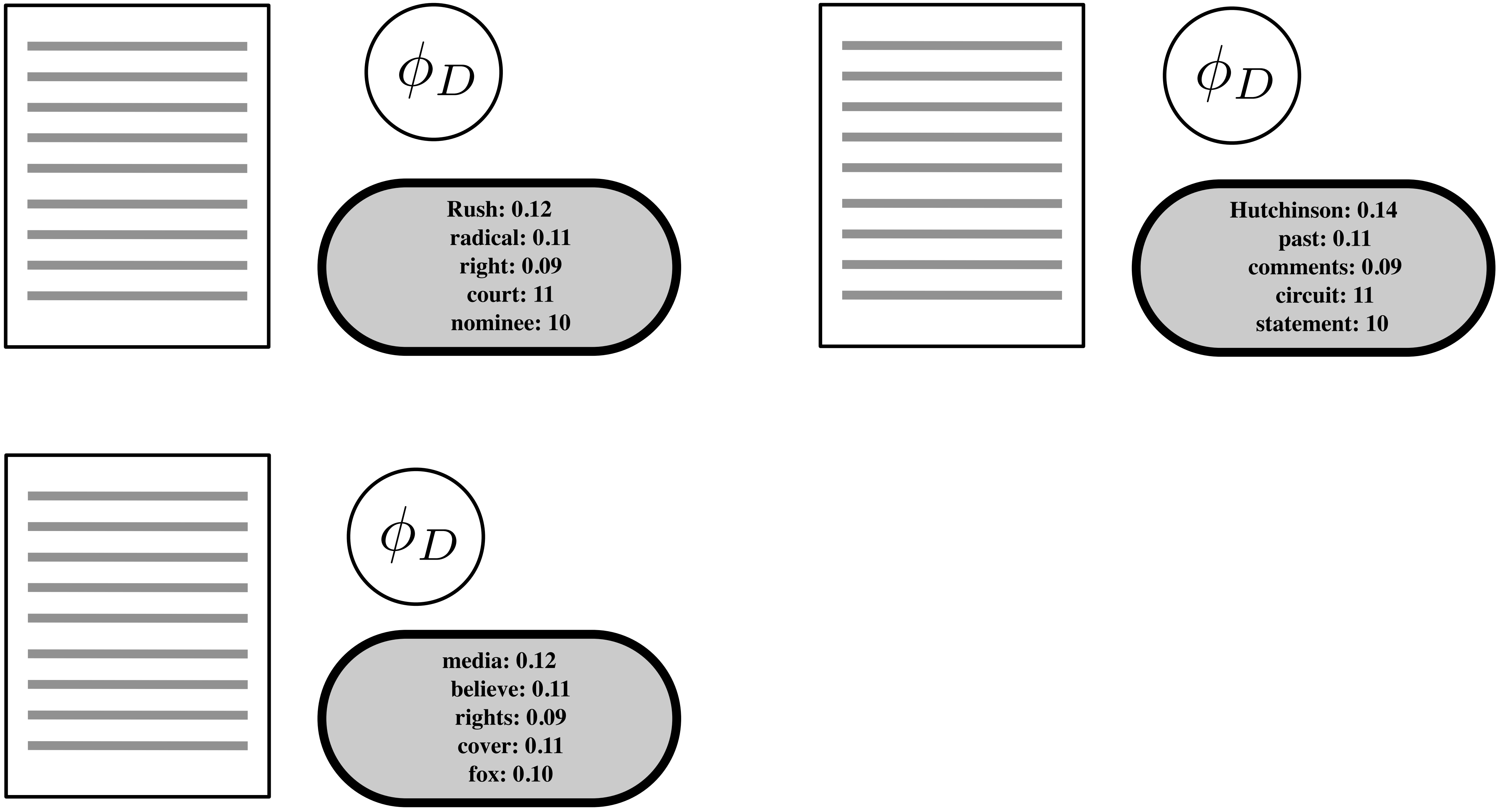
ϕ_D

Hutchinson: 0.14
past: 0.11
comments: 0.09
circuit: 11
statement: 10

similar to [Daume & Marcu, 2006]

Adding Topics

Document-Specific Distribution



ϕ_D

Rush: 0.12
radical: 0.11
right: 0.09
court: 11
nominee: 10

ϕ_D

Hutchinson: 0.14
past: 0.11
comments: 0.09
circuit: 11
statement: 10

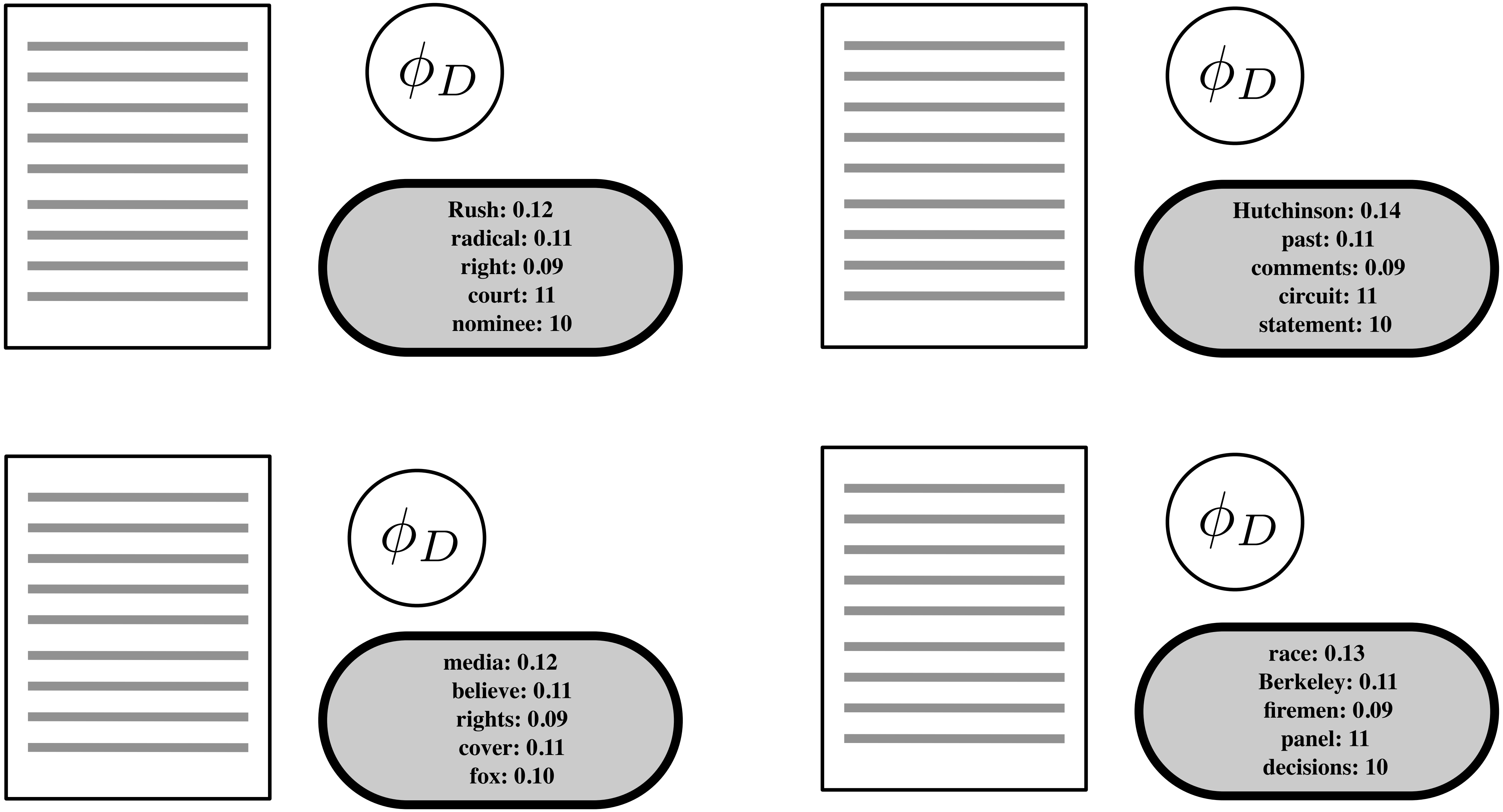
ϕ_D

media: 0.12
believe: 0.11
rights: 0.09
cover: 0.11
fox: 0.10

similar to [Daume & Marcu, 2006]

Adding Topics

Document-Specific Distribution



ϕ_D

Rush: 0.12
radical: 0.11
right: 0.09
court: 11
nominee: 10

ϕ_D

Hutchinson: 0.14
past: 0.11
comments: 0.09
circuit: 11
statement: 10

ϕ_D

media: 0.12
believe: 0.11
rights: 0.09
cover: 0.11
fox: 0.10

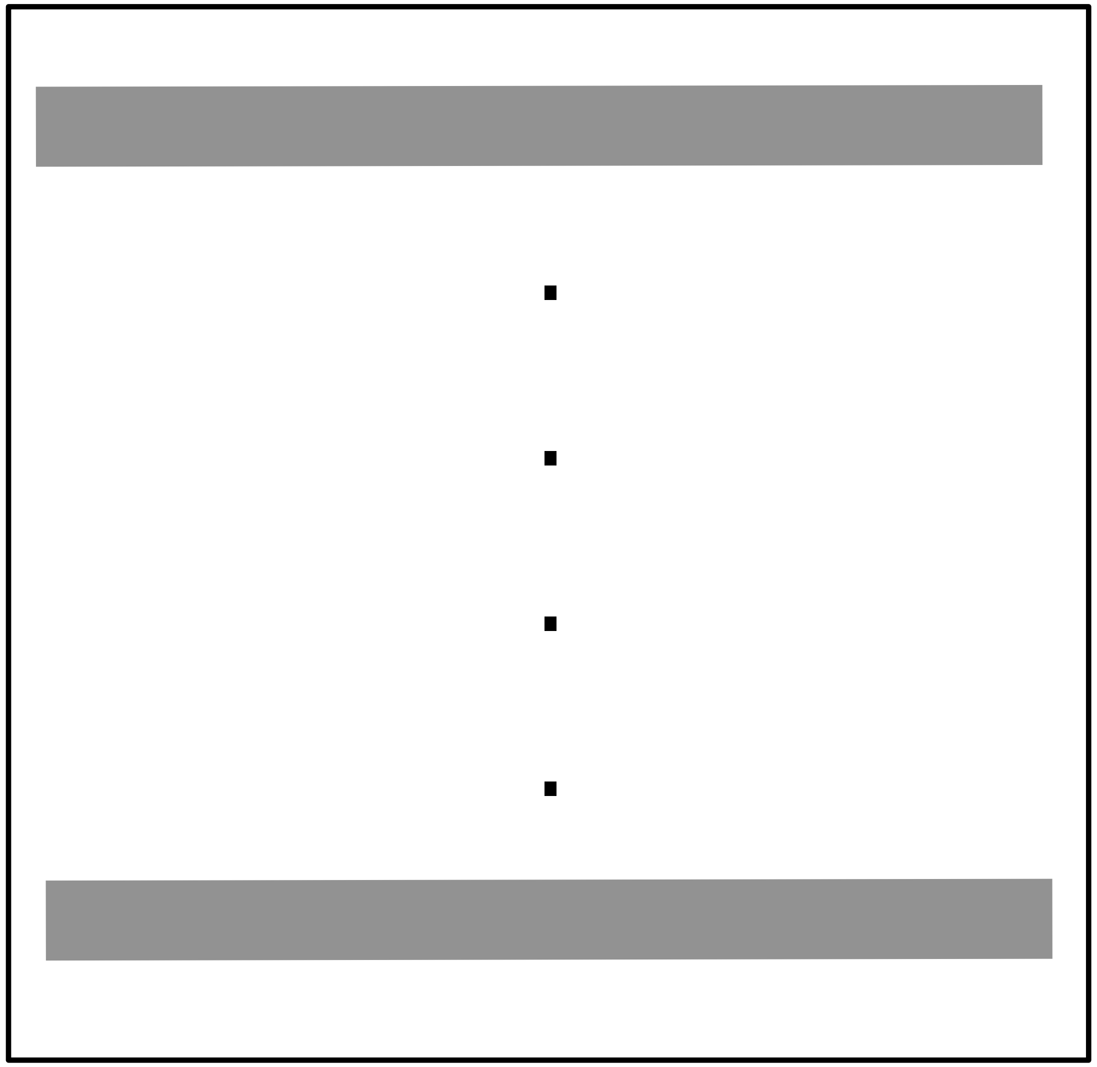
ϕ_D

race: 0.13
Berkeley: 0.11
firemen: 0.09
panel: 11
decisions: 10

similar to [Daume & Marcu, 2006]

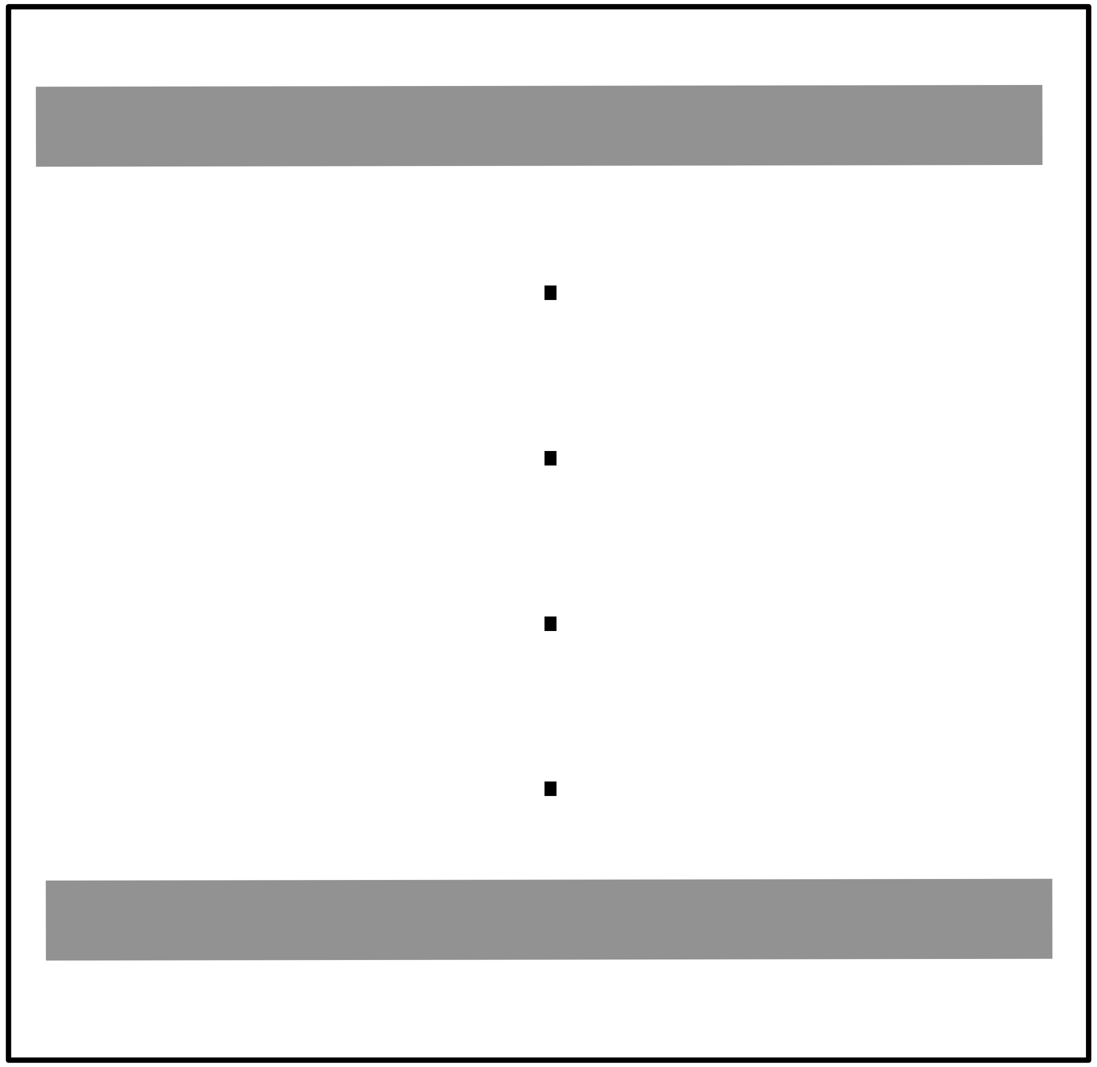
Adding Topics

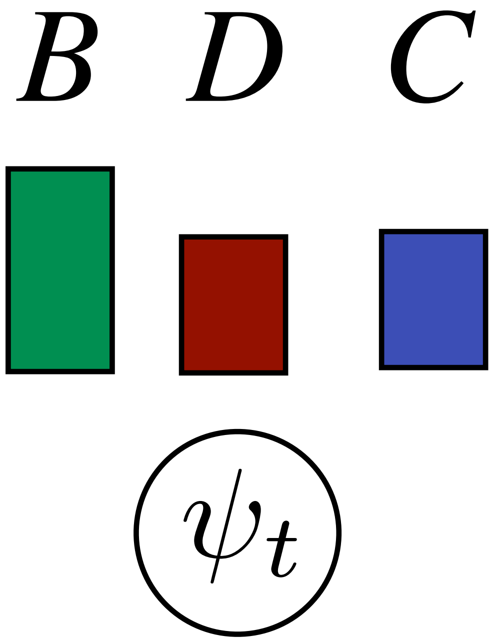
Document Level



Adding Topics

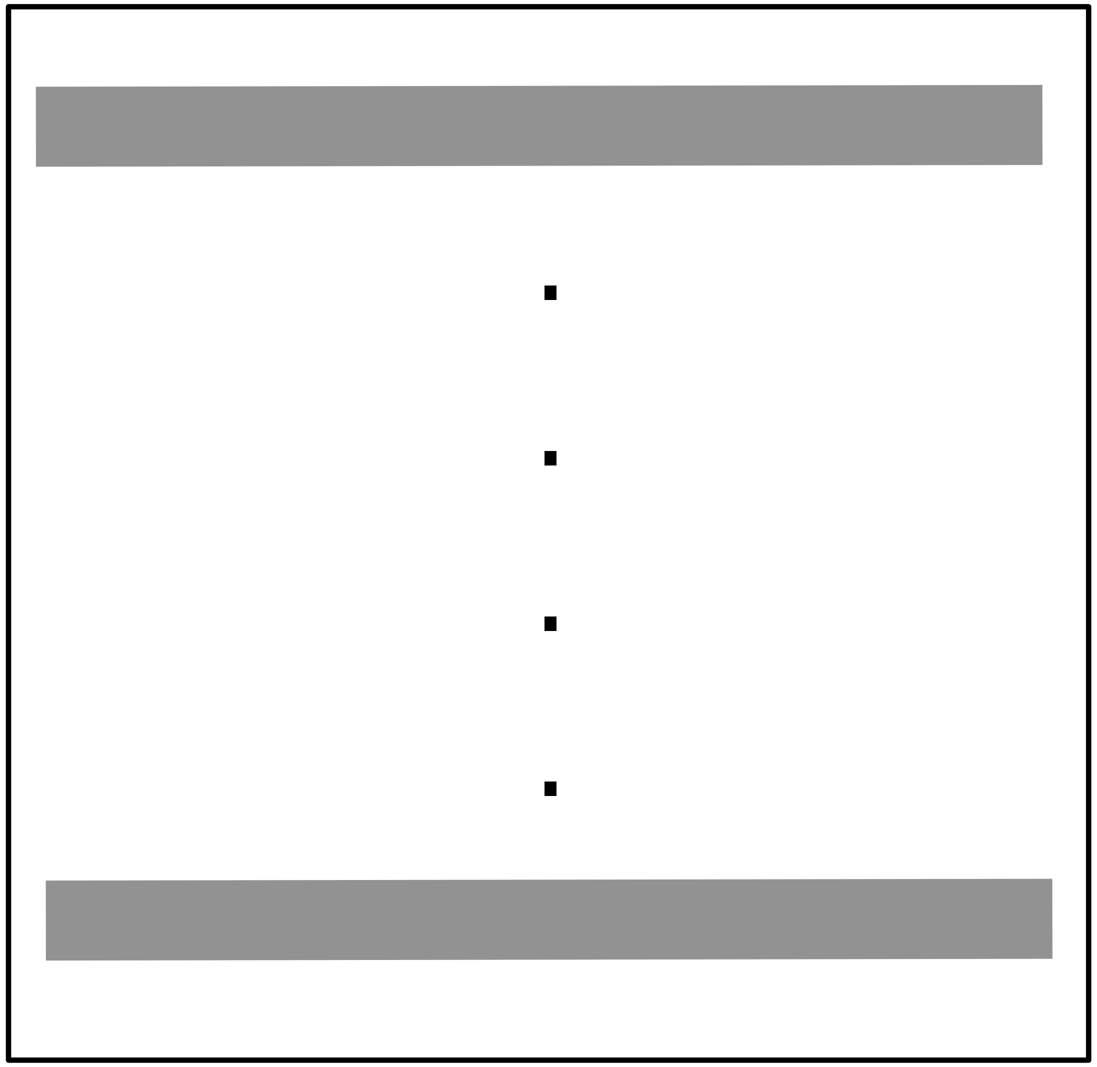
Document Level

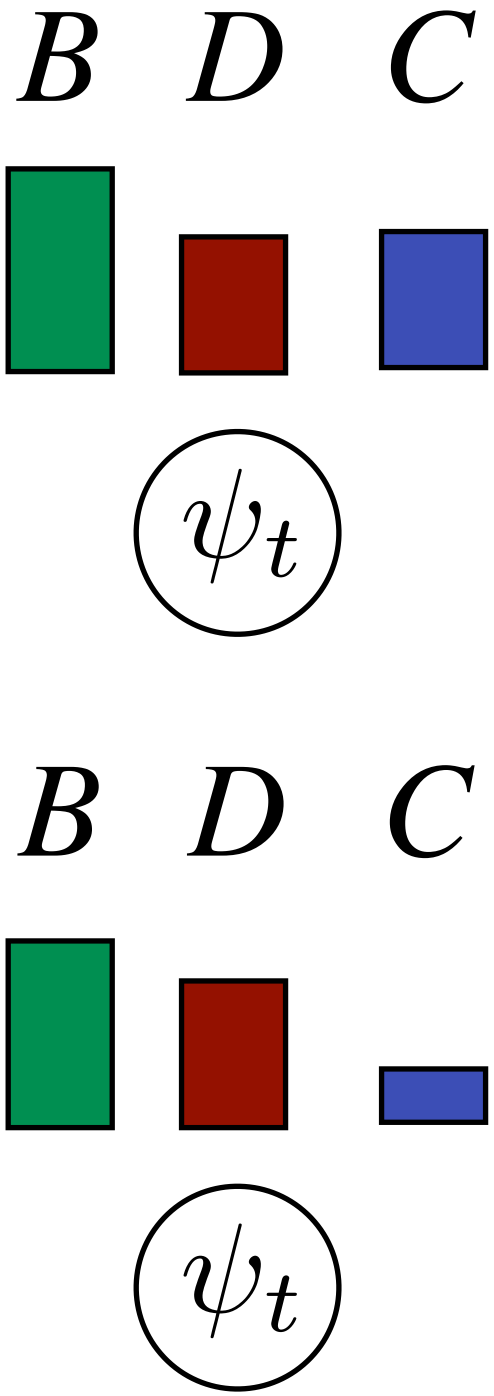




Adding Topics

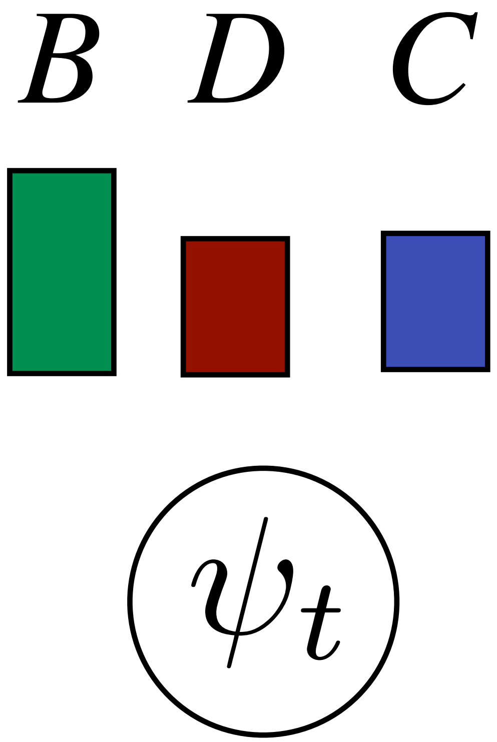
Document Level



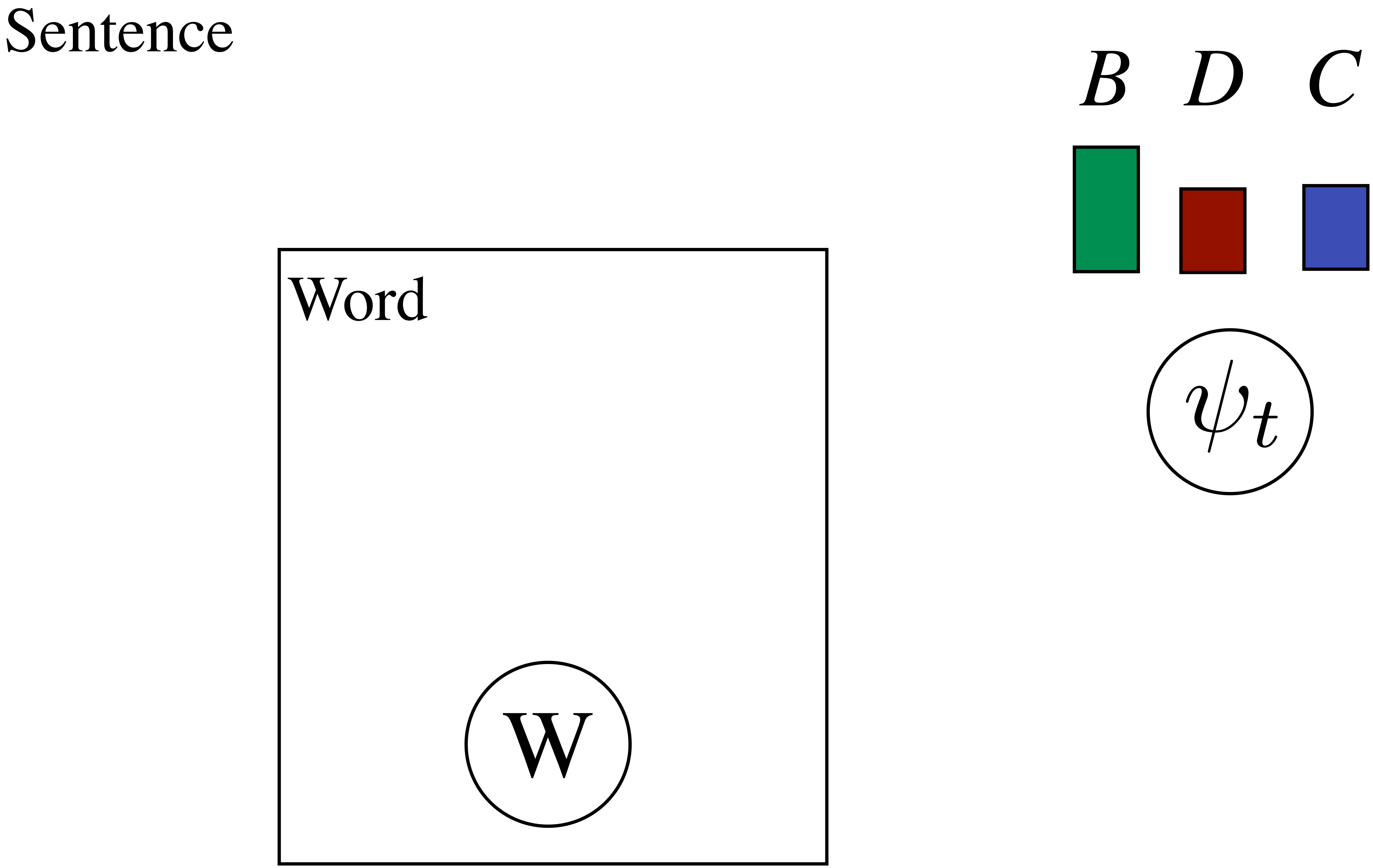


Adding Topics

Sentence



Adding Topics



Sentence

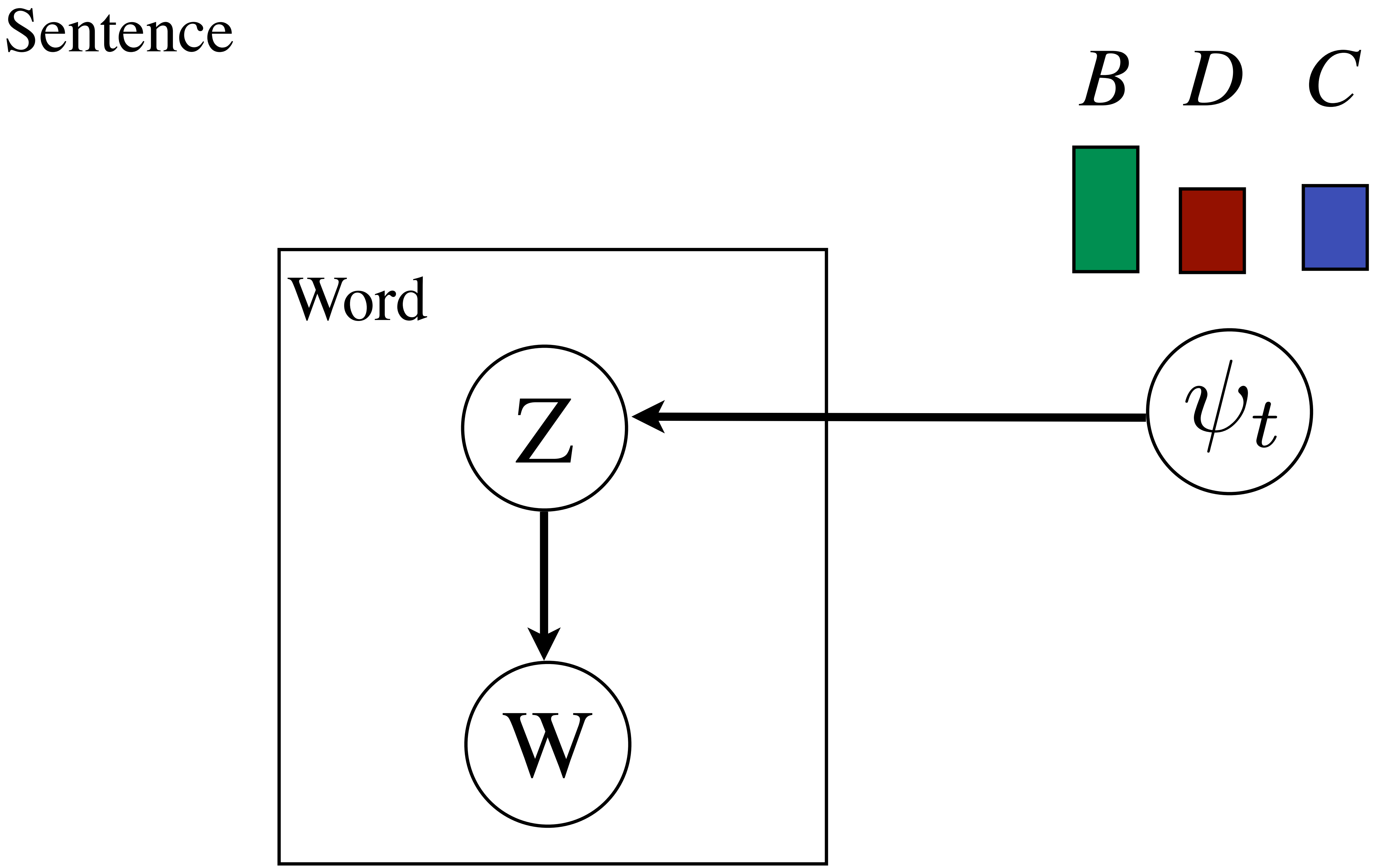
B D C

ψ_t

Word

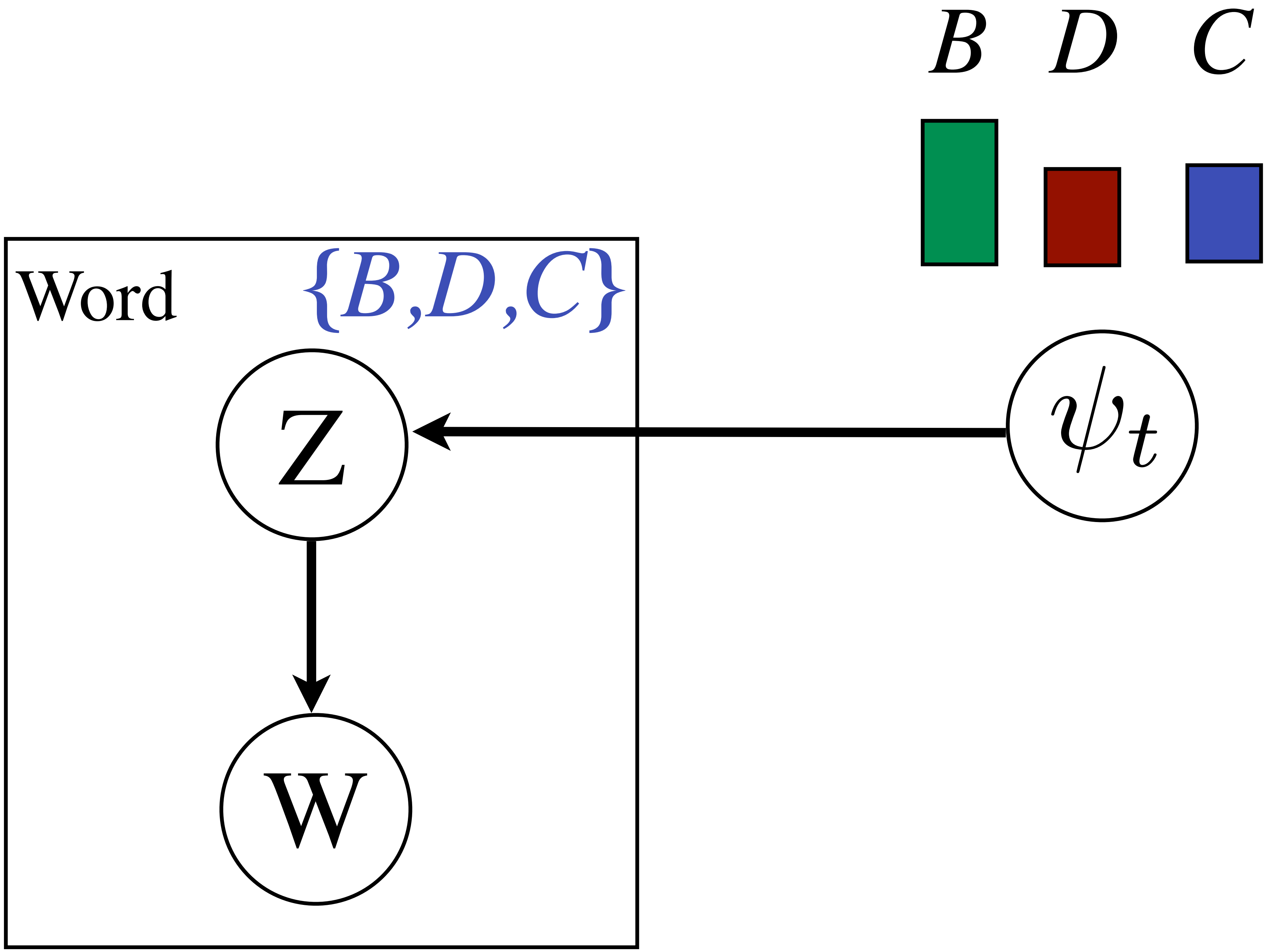
W

Adding Topics

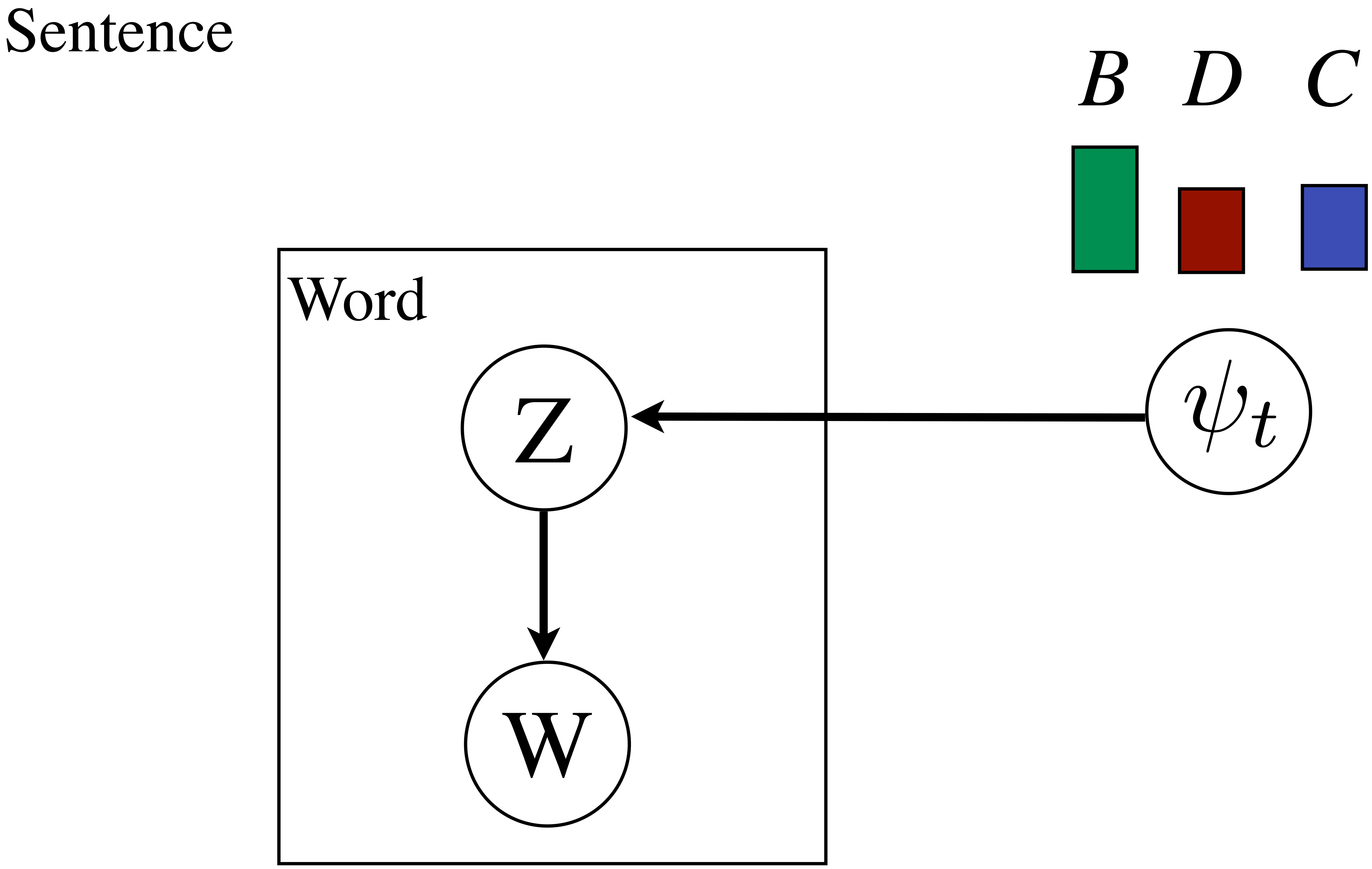


Adding Topics

Sentence



Adding Topics



Sentence

B D C

Word

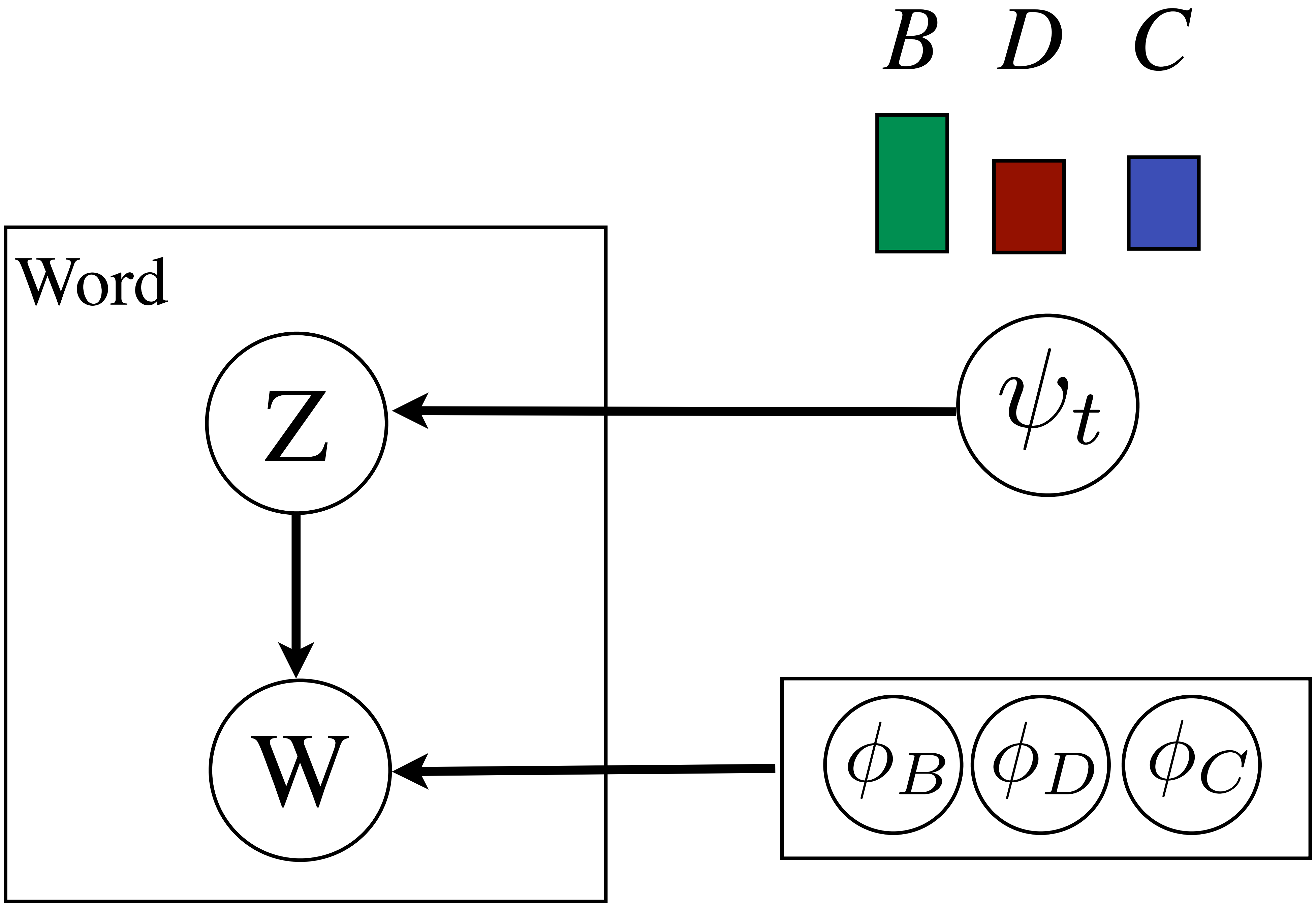
Z

W

ψ_t

Adding Topics

Sentence



B D C

Word

Z

W

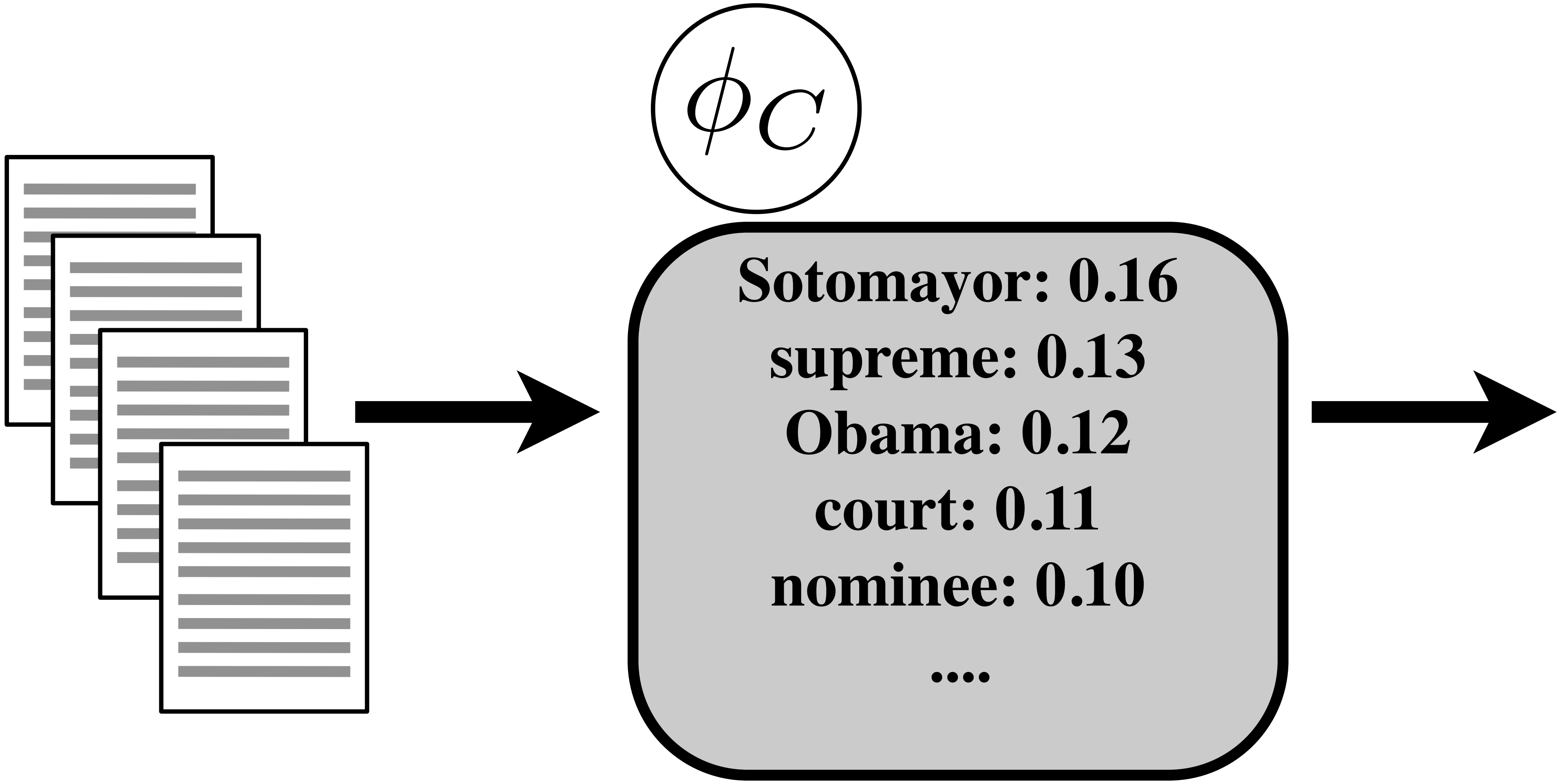
ψ_t

ϕ_B ϕ_D ϕ_C

Adding Topics

Representation

Extraction



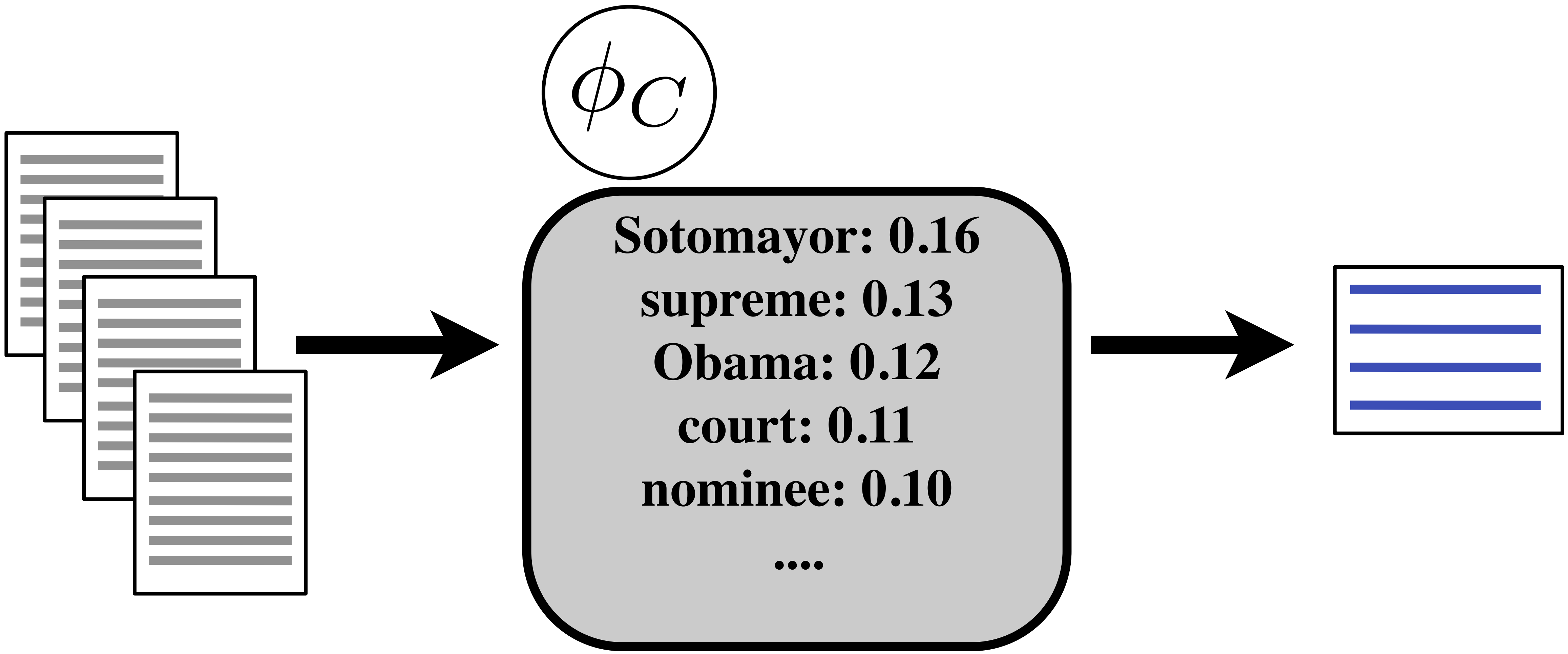
ϕ_C

Sotomayor: 0.16
supreme: 0.13
Obama: 0.12
court: 0.11
nominee: 0.10
....

Adding Topics

Representation

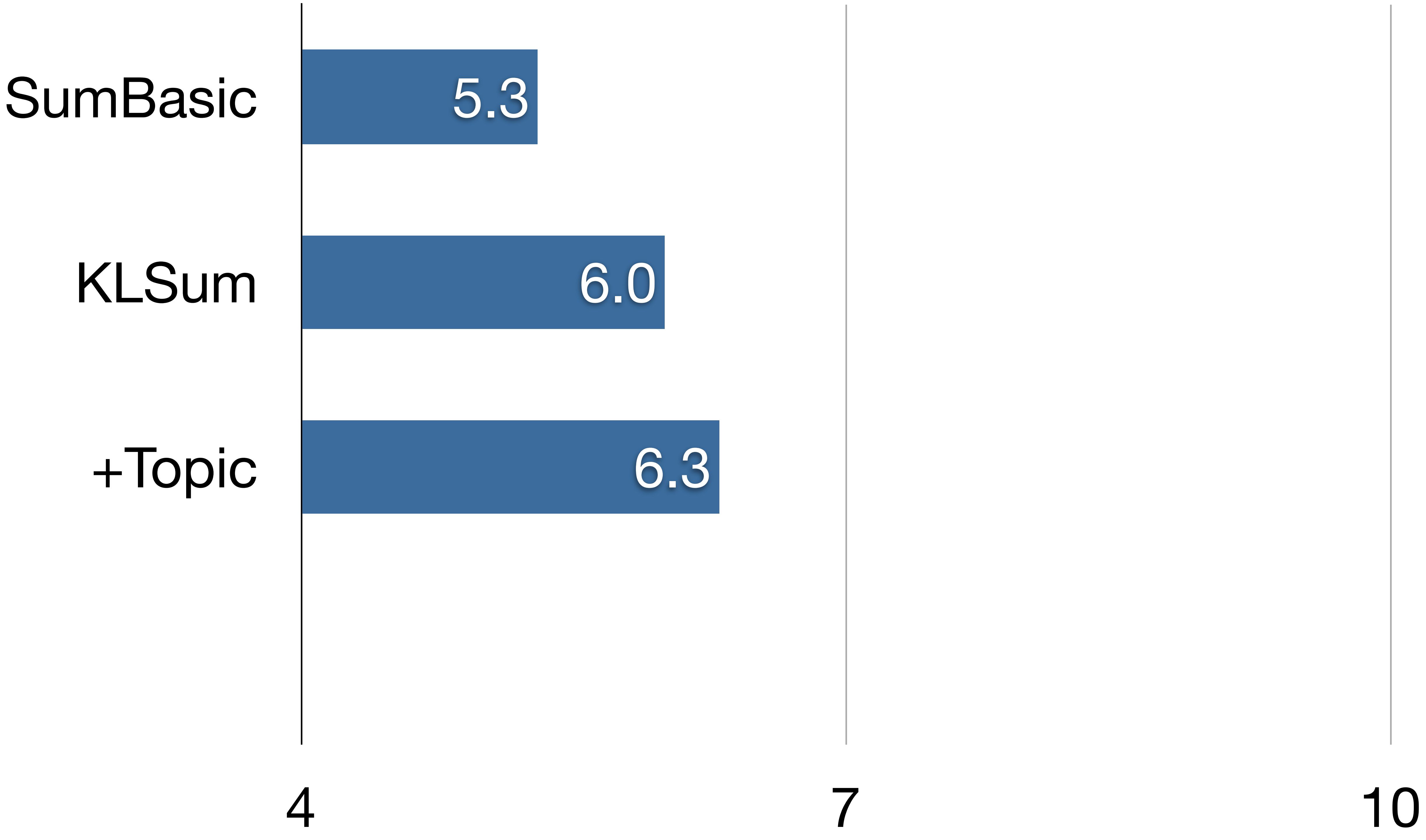
Extraction



ϕ_C

Sotomayor: 0.16
supreme: 0.13
Obama: 0.12
court: 0.11
nominee: 0.10
....

Performance

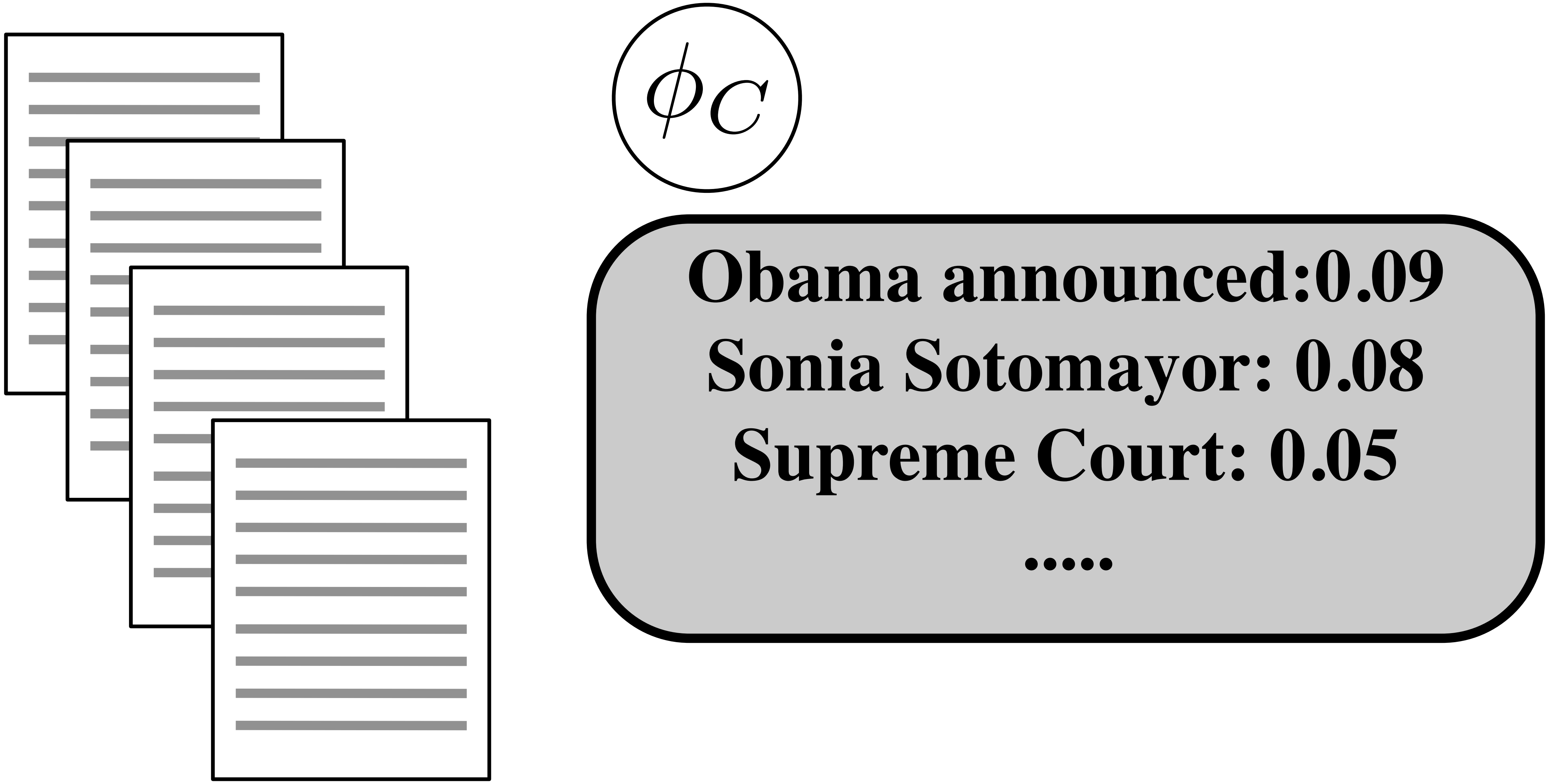


Adding Bigrams

Each sentence is a bag of bigrams

Adding Bigrams

Each sentence is a bag of bigrams



ϕ_C

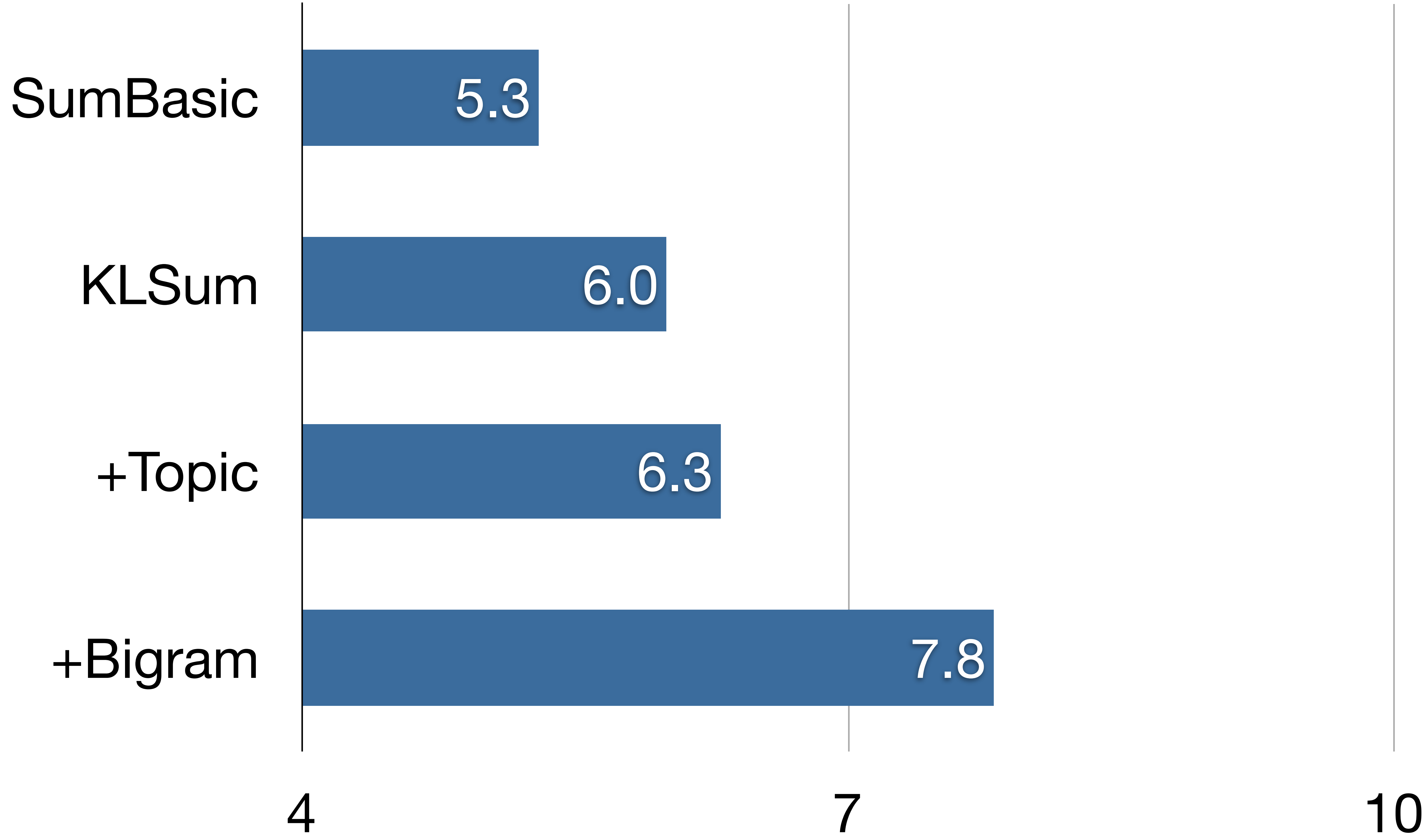
Obama announced:0.09

Sonia Sotomayor: 0.08

Supreme Court: 0.05

.....

Performance



Structured Content Models

Structured Content Models

General Content

Sotomayor: 0.16
supreme: 0.13
court: 0.12
....

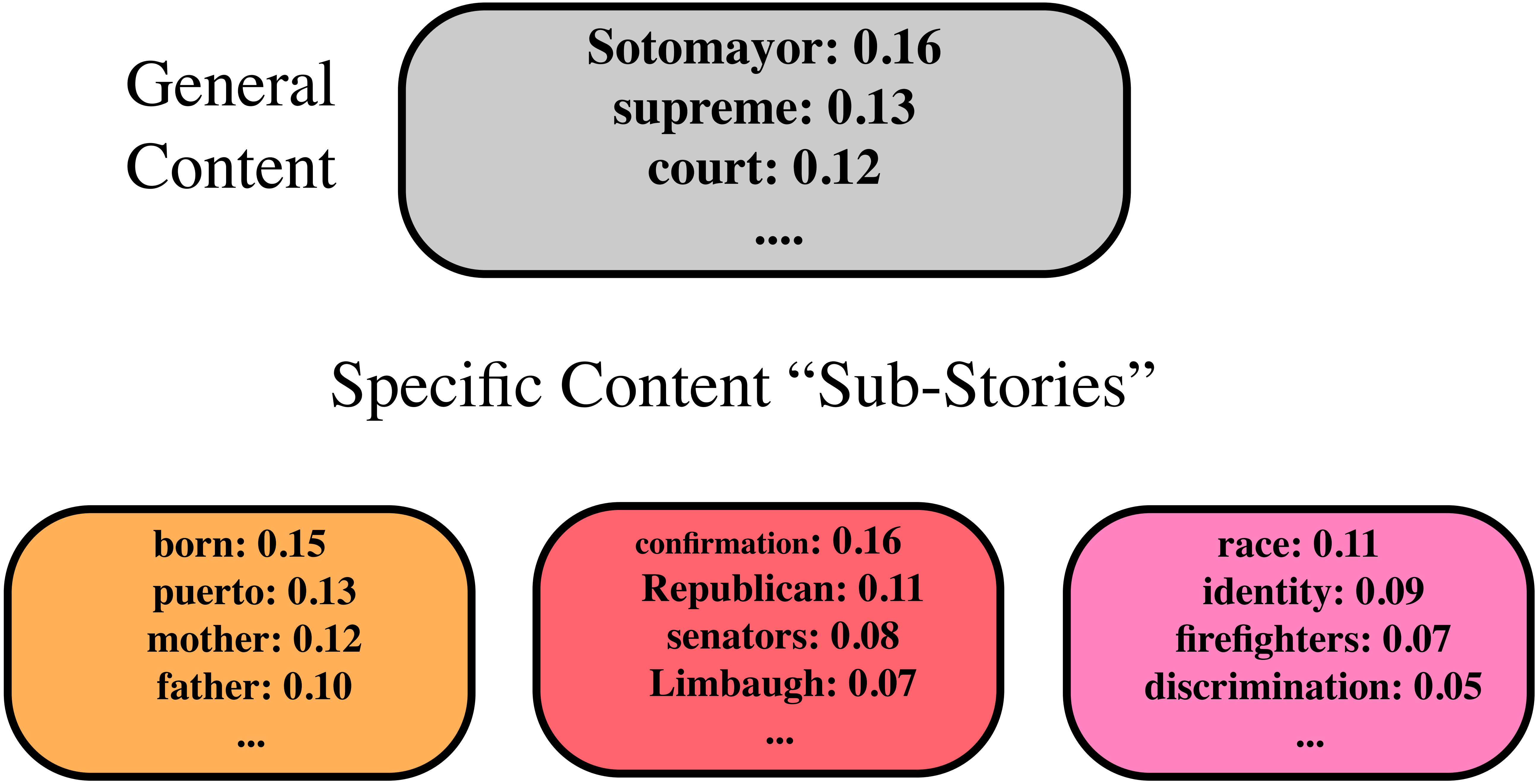
Structured Content Models

General Content

Sotomayor: 0.16
supreme: 0.13
court: 0.12
....

Specific Content “Sub-Stories”

Structured Content Models



General Content

Sotomayor: 0.16
supreme: 0.13
court: 0.12
....

Specific Content “Sub-Stories”

born: 0.15
puerto: 0.13
mother: 0.12
father: 0.10
...

confirmation: 0.16
Republican: 0.11
senators: 0.08
Limbaugh: 0.07
...

race: 0.11
identity: 0.09
firefighters: 0.07
discrimination: 0.05
...

Example Sub-Story

Biography

Sonia Sotomayor Born to Puerto Rican parents who moved to the Bronx in New York during World War Two.

Example Sub-Story

Biography

born: 0.15

puerto: 0.13

mother: 0.12

father: 0.10

...

Sonia Sotomayor Born to Puerto Rican parents who moved to the Bronx in New York during World War Two.

Example Sub-Story

Confirmation

Senate Republicans have combed over Sotomayor's record on the federal bench ahead of confirmation hearings.

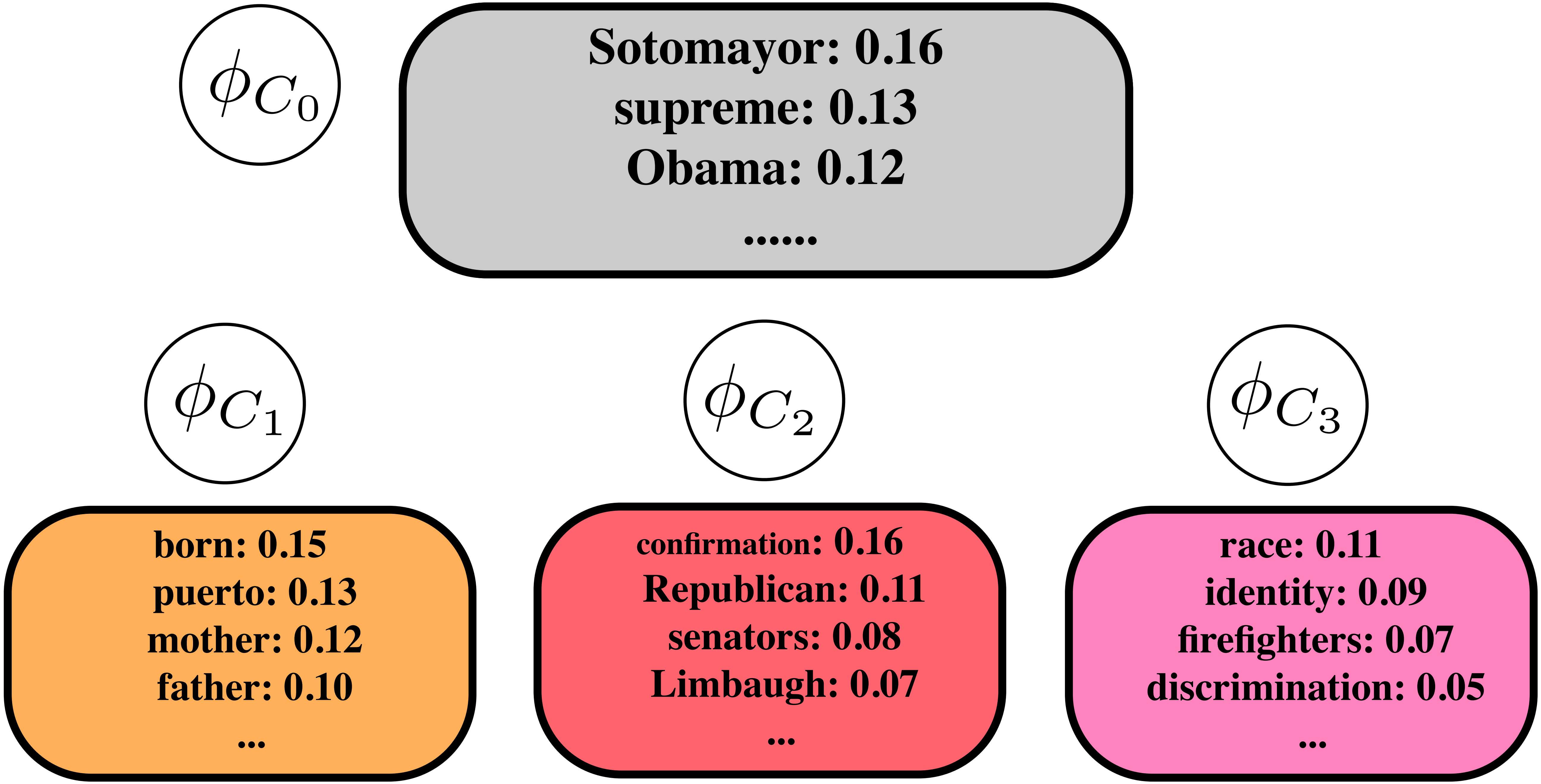
Example Sub-Story

Confirmation

confirmation: 0.16
Republican: 0.11
senators: 0.08
Limbaugh: 0.07
...

Senate Republicans have combed over Sotomayor's record on the federal bench ahead of confirmation hearings.

Structured Content Models



ϕ_{C_0}

Sotomayor: 0.16
supreme: 0.13
Obama: 0.12
......

ϕ_{C_1}

born: 0.15
puerto: 0.13
mother: 0.12
father: 0.10
...

ϕ_{C_2}

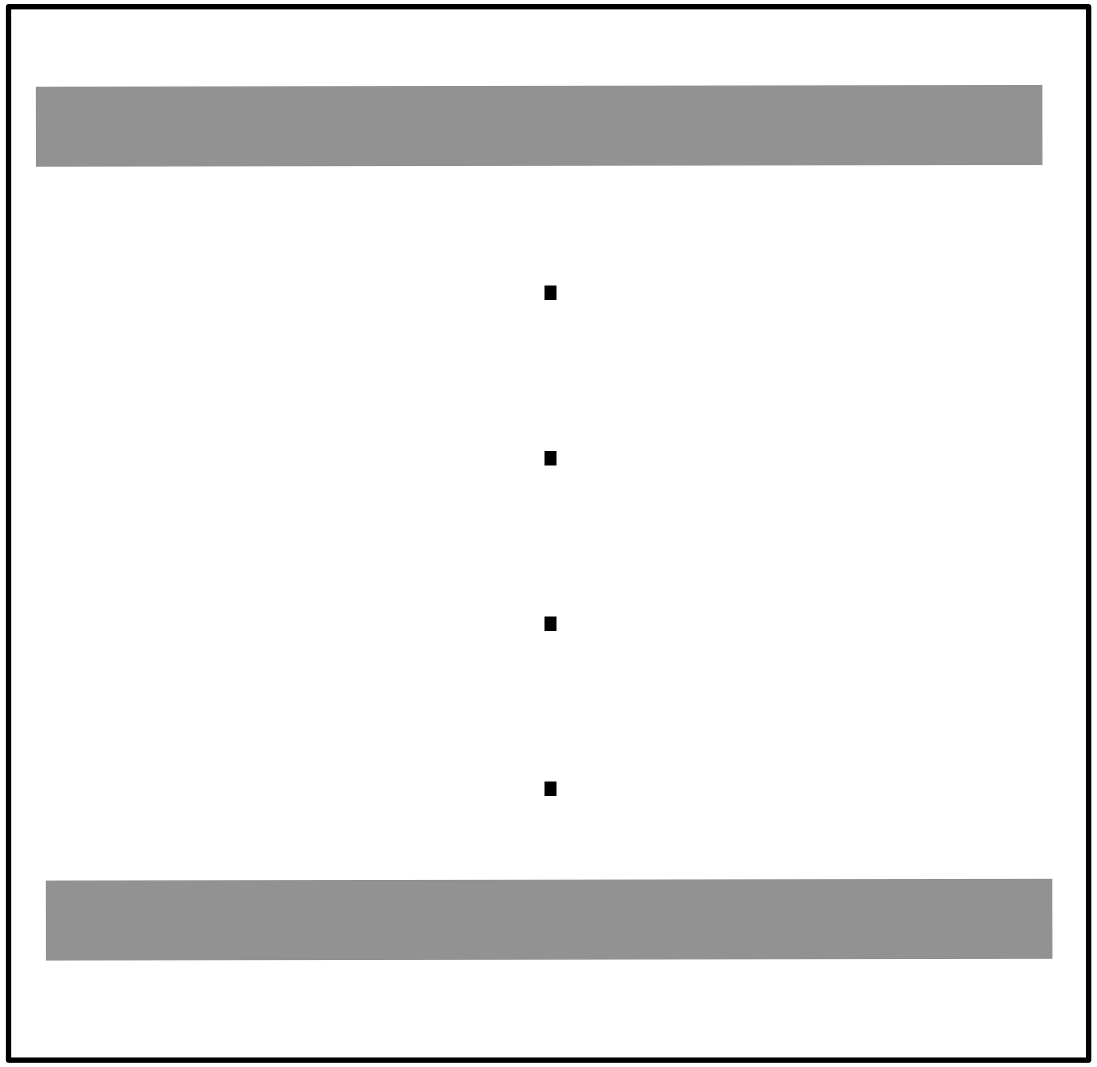
confirmation: 0.16
Republican: 0.11
senators: 0.08
Limbaugh: 0.07
...

ϕ_{C_3}

race: 0.11
identity: 0.09
firefighters: 0.07
discrimination: 0.05
...

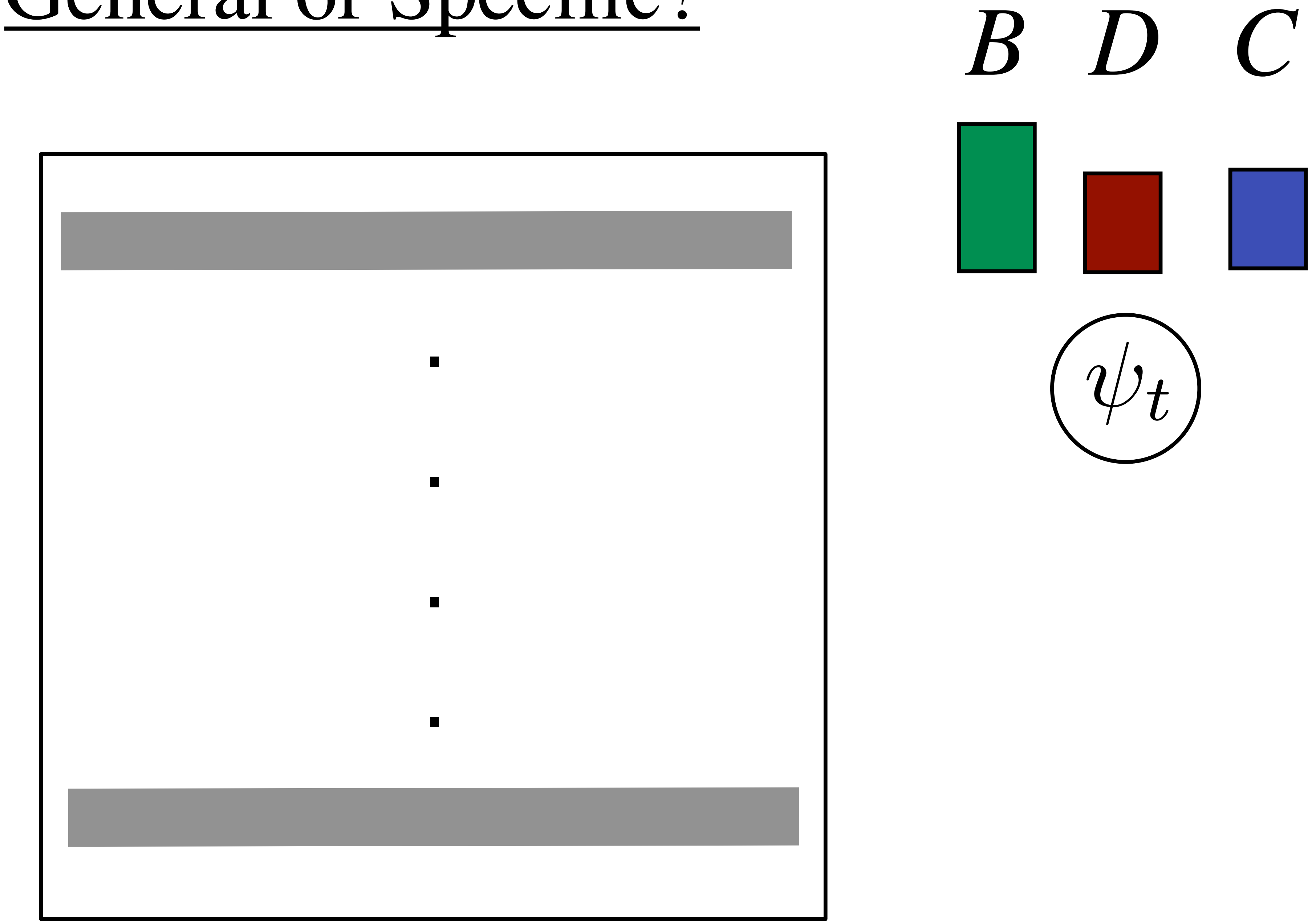
Structured Topics

General or Specific?



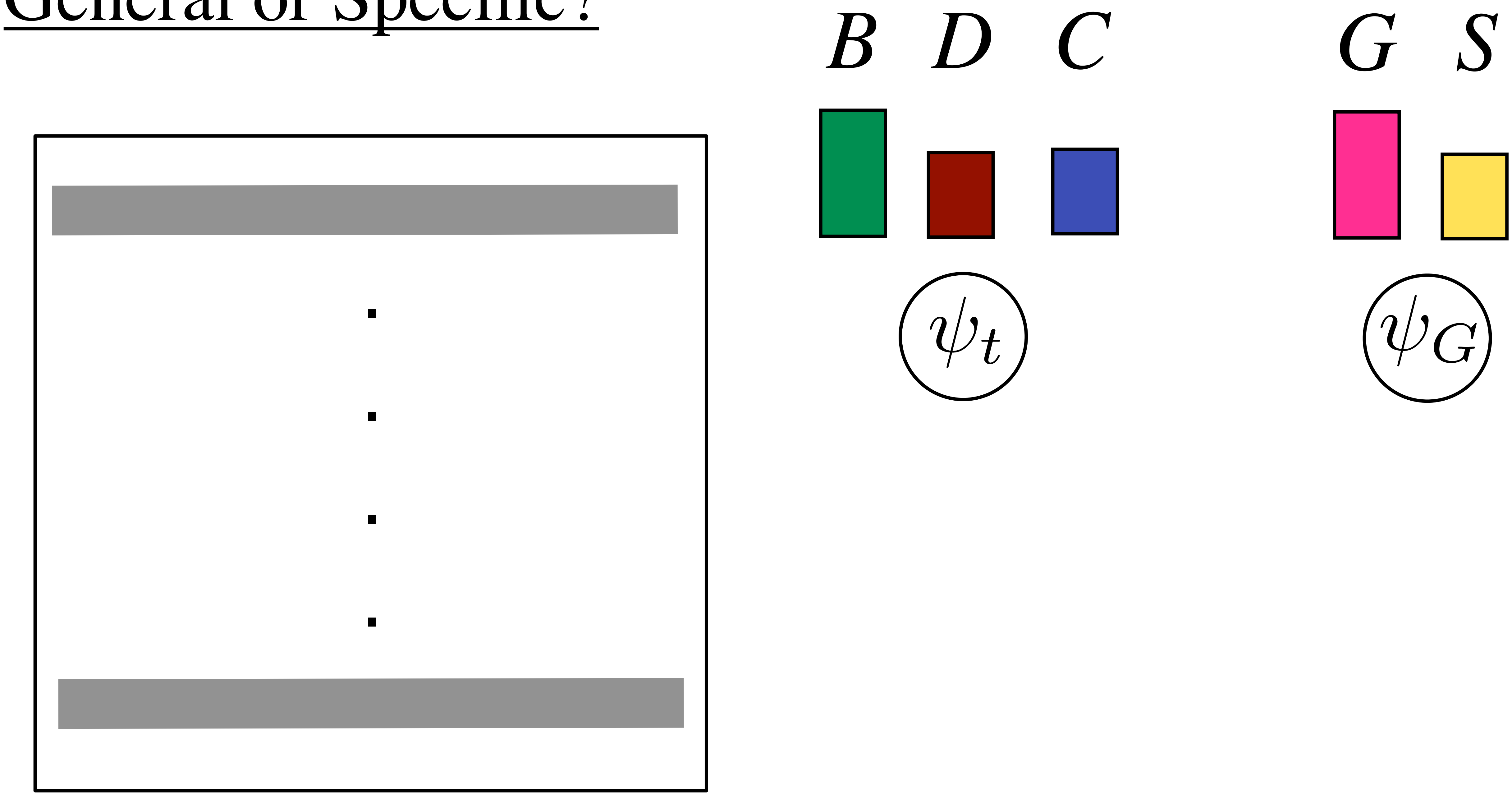
Structured Topics

<u>General or Specific?</u>



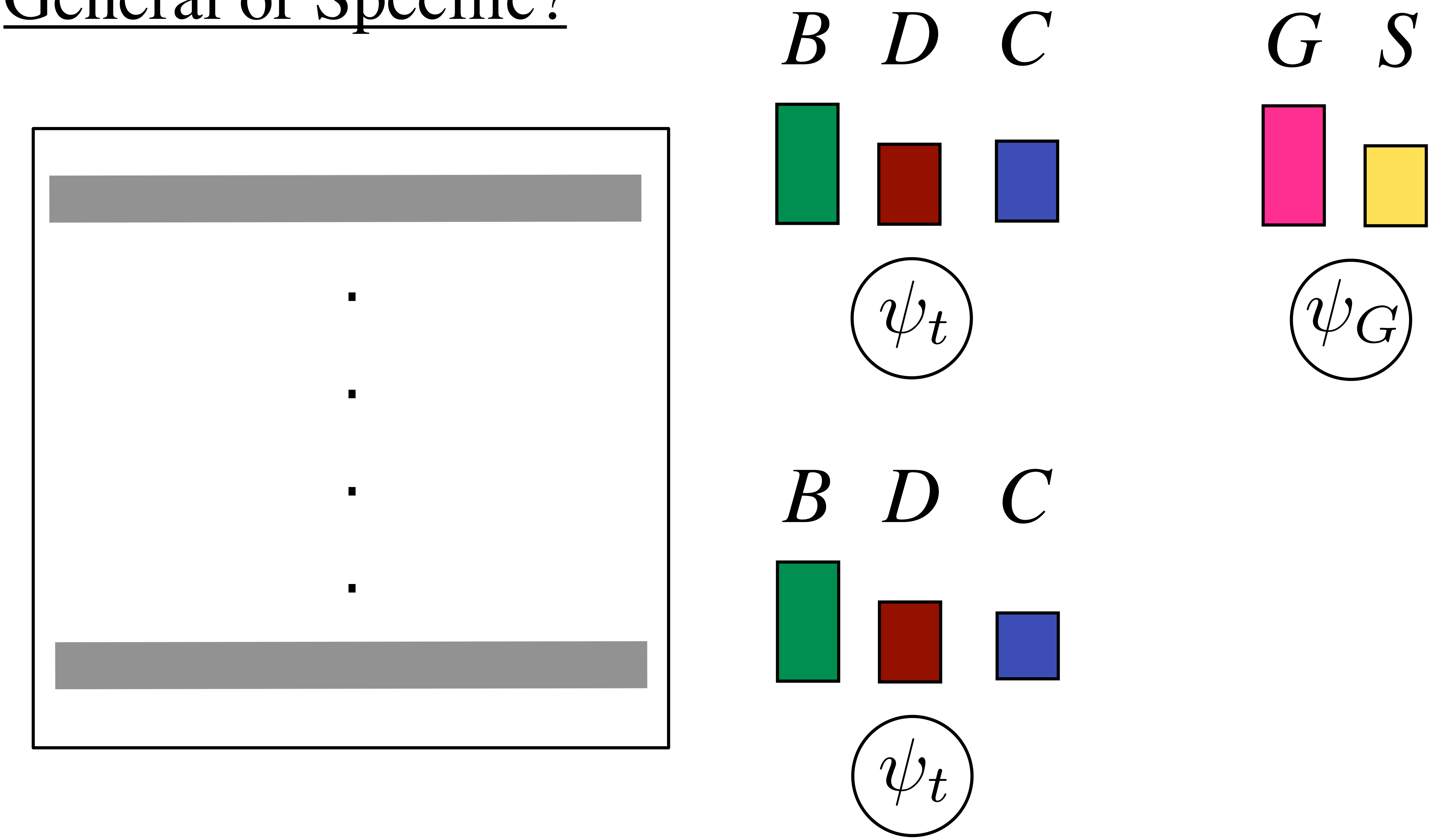
Structured Topics

General or Specific?



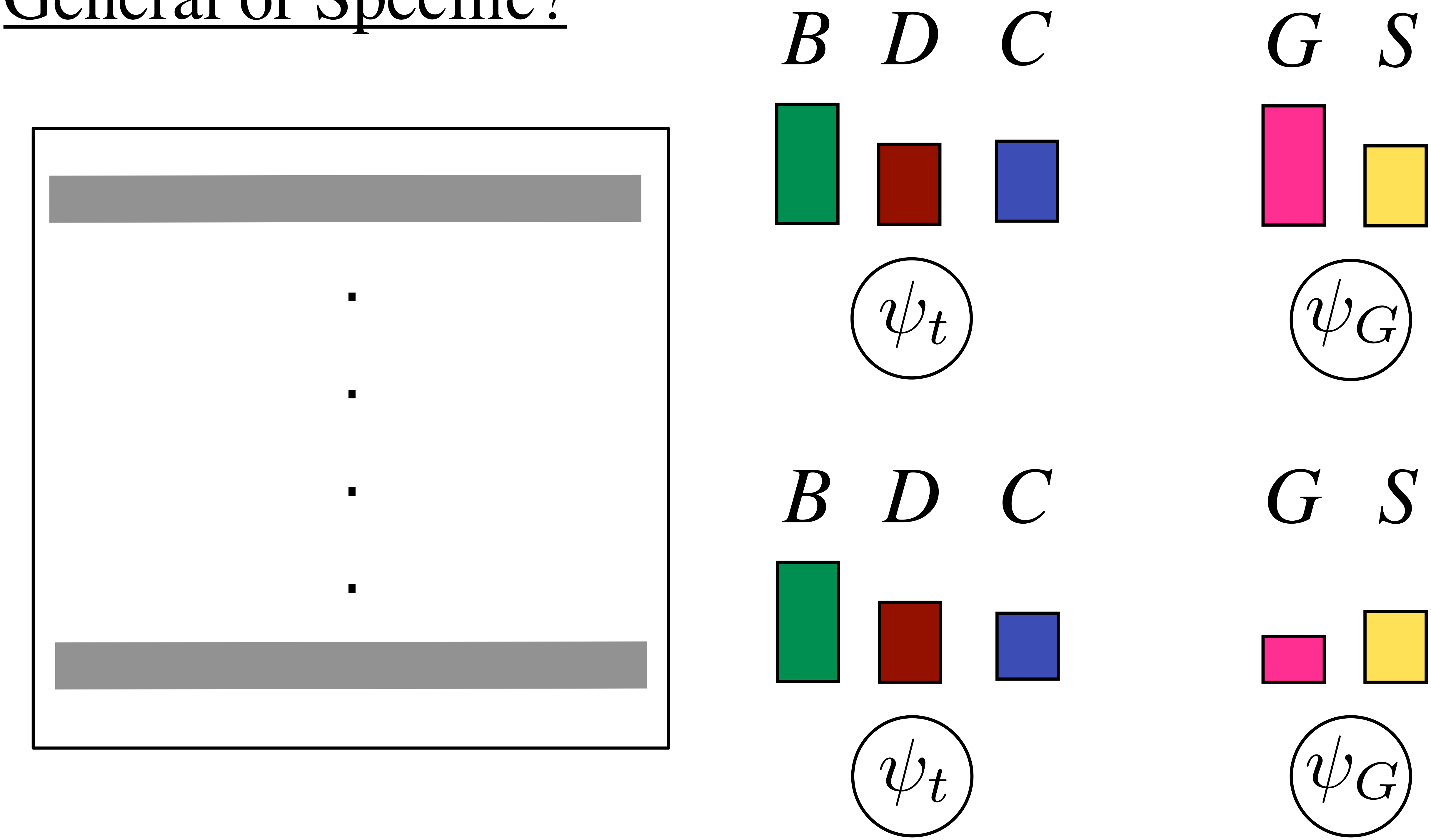
Structured Topics

General or Specific?



Structured Topics

<u>General or Specific?</u>

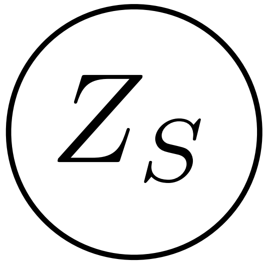


Structured Topics

What specific topic?

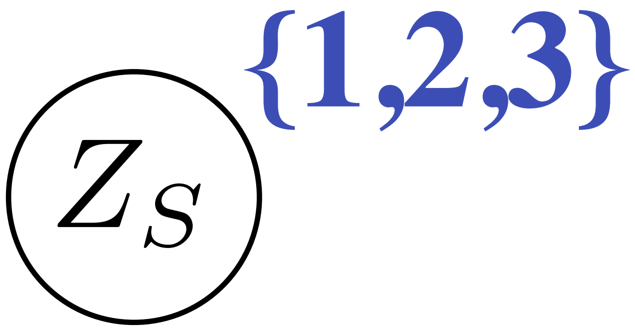
Structured Topics

What specific topic?



Structured Topics

<u>What specific topic?</u>

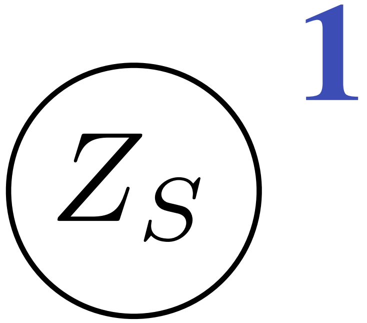


Structured Topics

What specific topic?

Structured Topics

What specific topic?

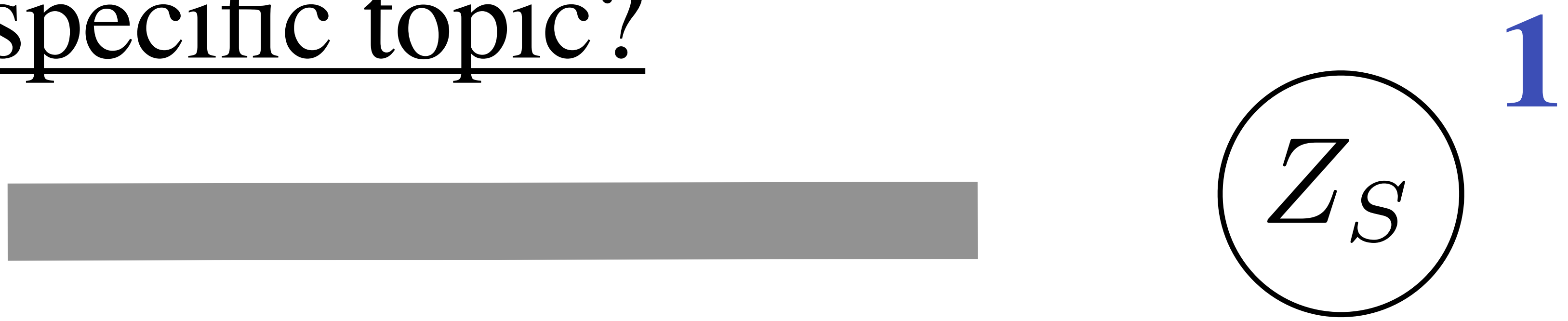


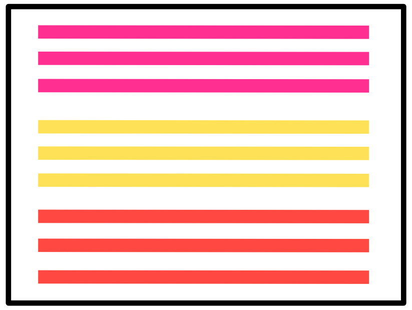
“Sticky” Specific Topics

$$P(Z'_S|Z_S) = \begin{cases} \sigma & \text{if } Z'_S = Z_S \\ (1-\sigma)\theta_{Z'_S|Z_S} & \text{o.w.} \end{cases}$$

Structured Topics

What specific topic?

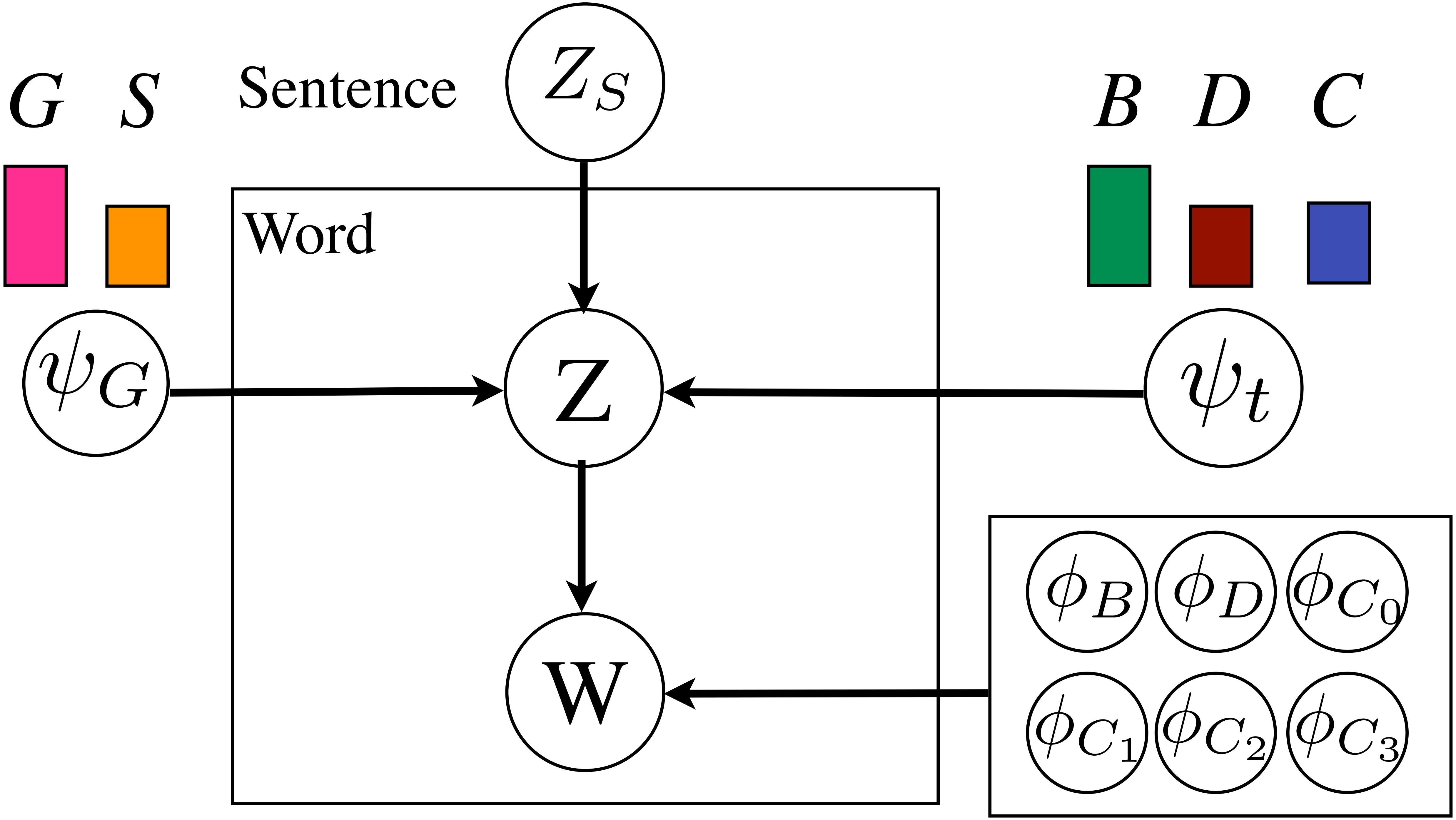




“Sticky” Specific Topics

$$P(Z'_S|Z_S) = \begin{cases} \sigma & \text{if } Z'_S = Z_S \\ (1-\sigma)\theta_{Z'_S|Z_S} & \text{o.w.} \end{cases}$$

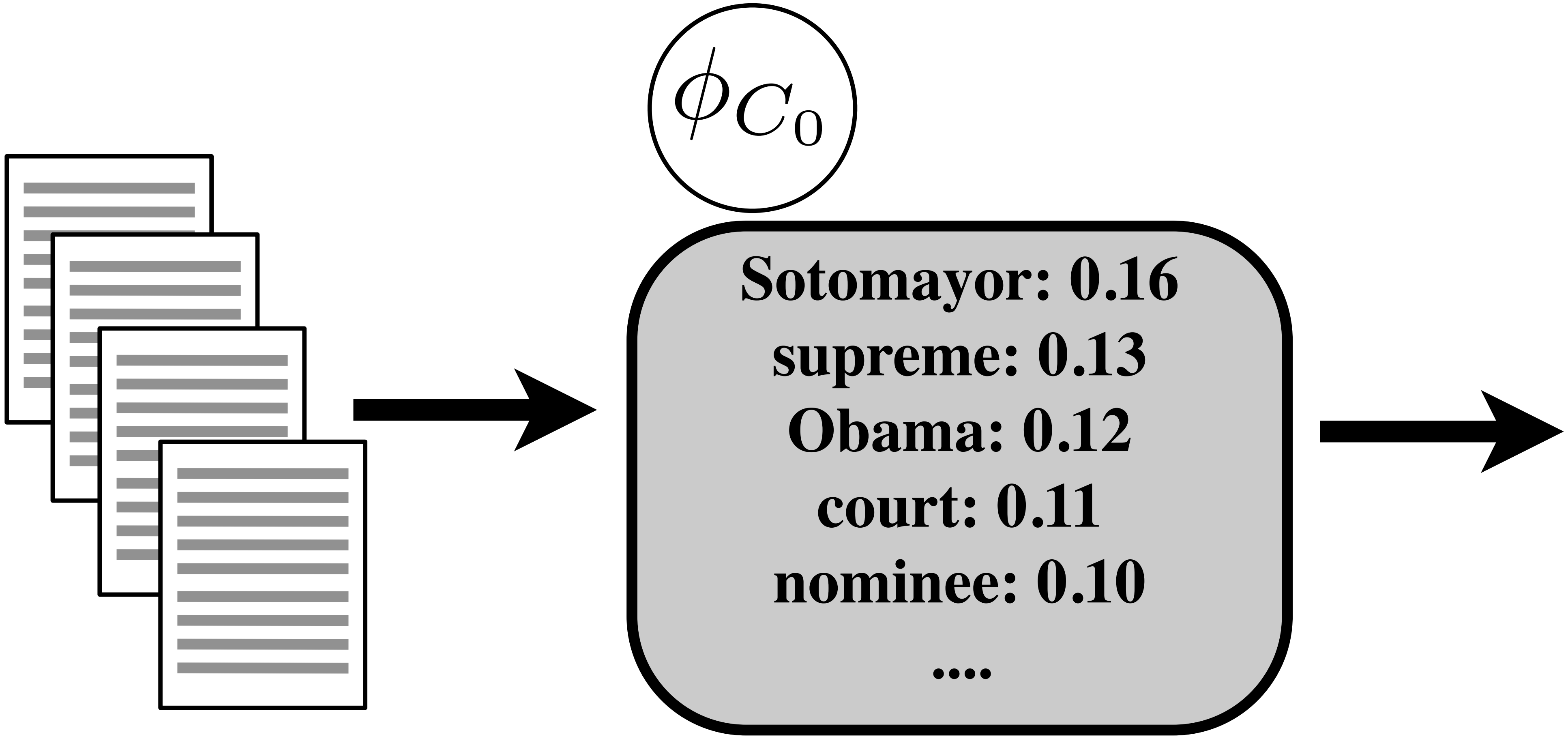
Structured Topics



Structured Topics

Representation

Extraction



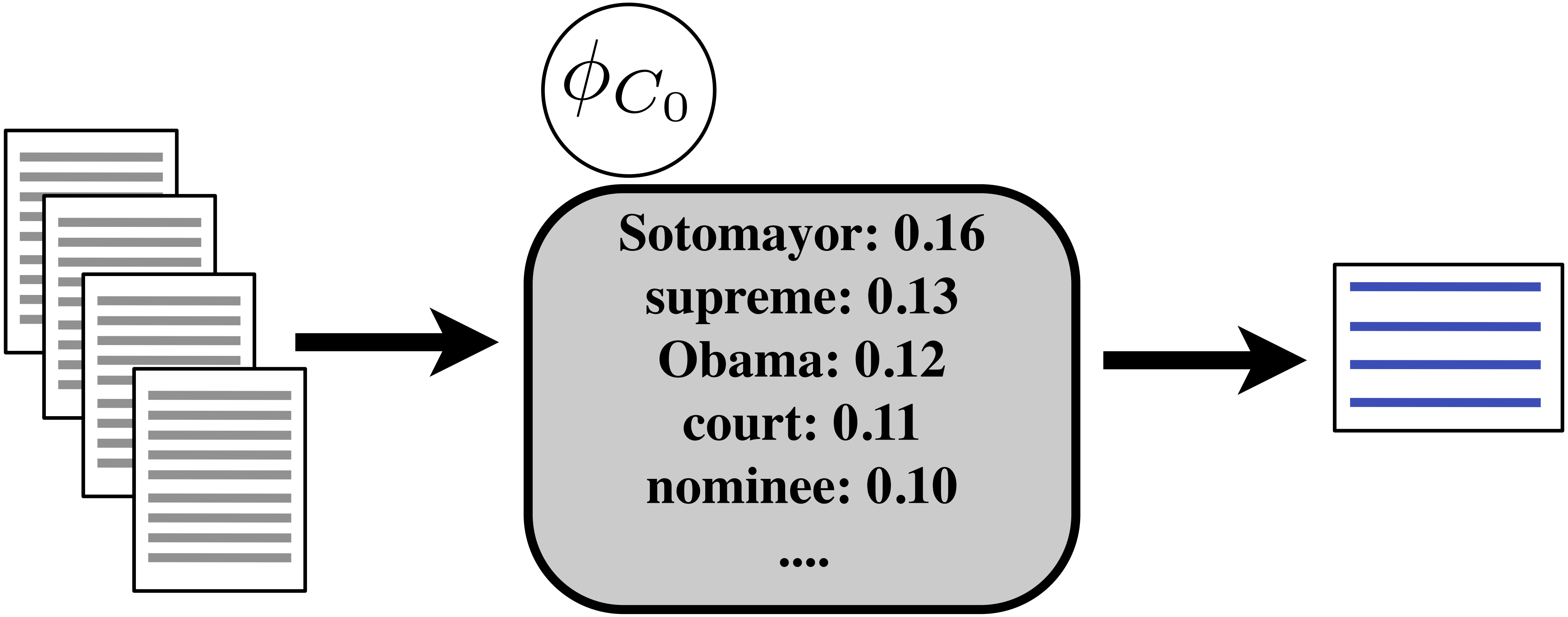
ϕ_{C_0}

Sotomayor: 0.16
supreme: 0.13
Obama: 0.12
court: 0.11
nominee: 0.10
....

Structured Topics

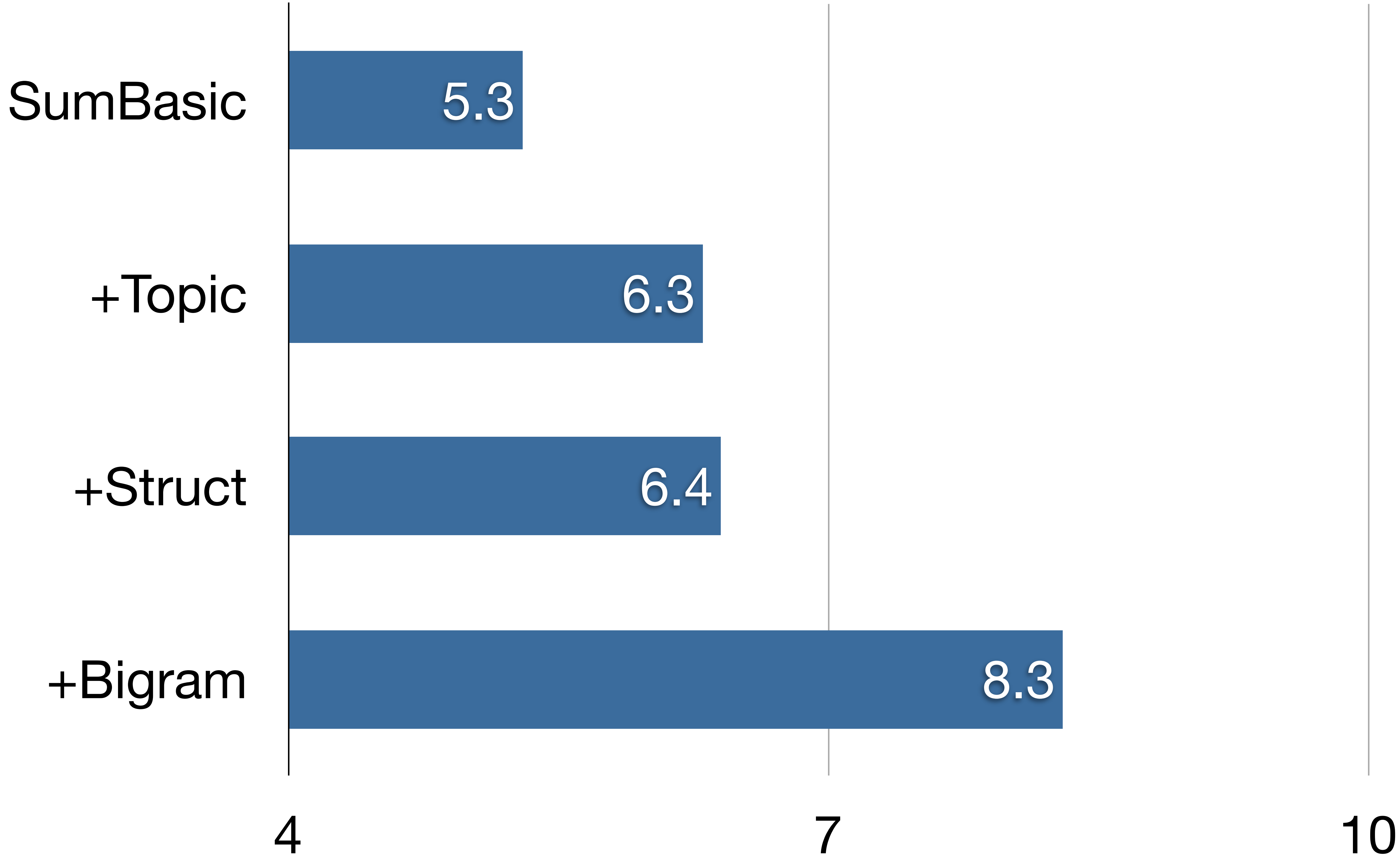
Representation

Extraction



Sotomayor: 0.16
supreme: 0.13
Obama: 0.12
court: 0.11
nominee: 0.10
....

Performance



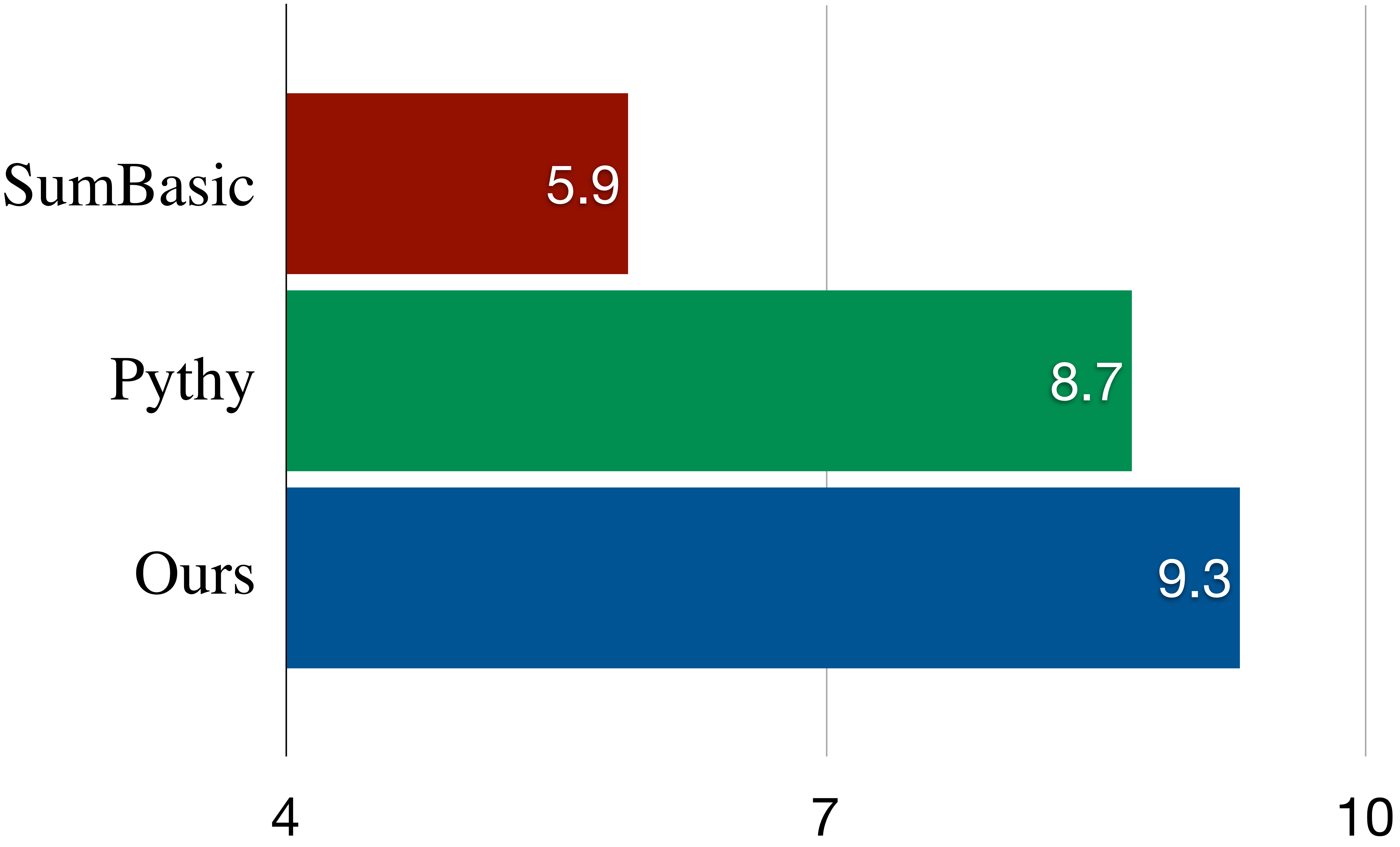
HierSum: Manual Evaluation

- Pairwise Comparison
  - 23 participants, 69 judgments
  - Each participant sees reference summary and two model summaries
  - Pythy: State-of-the-art discriminative system. Highest automatic score and high-performing on manual content evaluation

Manual Evaluation

Question	Pythy	Ours
Overall	20	49
Redundancy	21	48
Coherence	15	54
Focus	28	41

DUC '07 Results



Example General Summary

Former House Speaker Newt Gingrich is asking a judge to force his estranged wife to turn over money he says she is hoarding.

On Thursday, accusations of wrongdoing and the mining of dirt in the former U.S. House speaker's divorce case gave way to a secret settlement between Gingrich and his wife of 18 years, Marianne Gingrich.

Gingrich filed for divorce July 29 amid allegations he is having an affair with 33-year-old congressional aide Callista Bisek.

Newt Gingrich's attorney, Randy Evans, says Marianne Gingrich has refused to even discuss a settlement until she questions Bisek.

Example Topical Summary

Callista Bisek

Marianne Gingrich also wants to depose Callista Bisek, a congressional aide with whom the former U.S. House speaker has a relationship.

And in motions filed last week in Superior Court for the District of Columbia, Callista Bisek, a clerk for the House Agriculture Committee, asked a judge to overturn a Georgia court order requiring her to answer questions about her relationship with Gingrich.

Mayoue has said he intends to question Bisek about all aspects of her relationship with Newt Gingrich.

Example Topical Summary

Gingrich Bio / Post-Speaker Life

Gingrich is best known leading the Republican Party's takeover of the House in 1994. During that so-called Republican Revolution, Gingrich emphasized that "family values" should be a core pillar in American society.

Since resigning as speaker and from the congressional seat he held for 20 years, Gingrich has been making a living giving speeches, sitting on corporate boards, consulting and appearing as a political analyst on Fox News.

U.S. Rep. J.D. Hayworth (R-Ariz.) argued that Gingrich's new job as a political commentator for Fox News makes it inappropriate to include him in political gatherings. "Time marches on. He's gone on to other pursuits," Hayworth said.

Conclusion

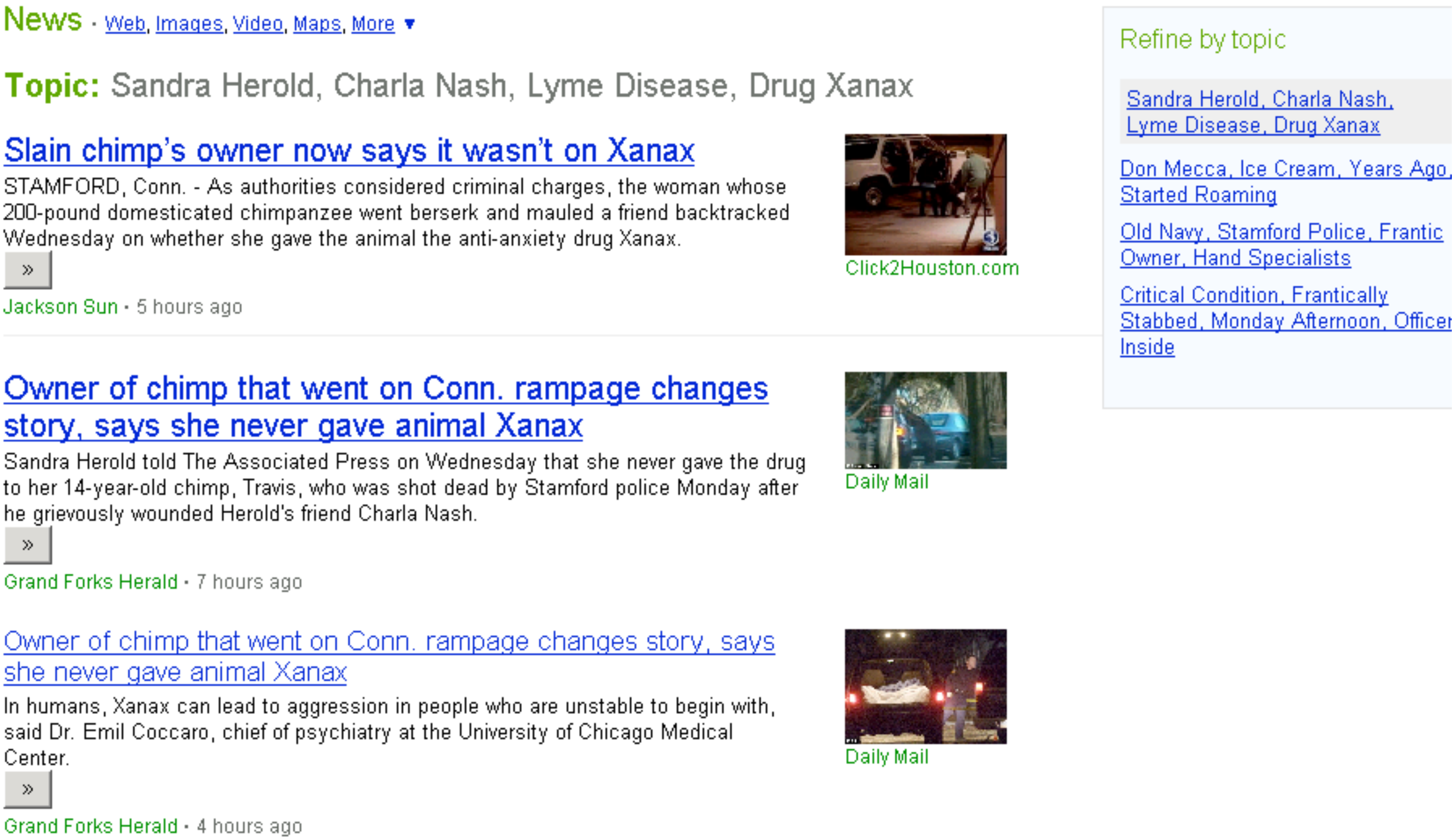
- KL objective for sentence extraction summarization
- Topic models can yield state-of-the-art automatic and manual summary eval
- Also structured content models for topical summarization

Thanks!

Questions?

If not, I have more examples..........

Topical Summarization



News · Web, Images, Video, Maps, More ▼

Topic: Sandra Herold, Charla Nash, Lyme Disease, Drug Xanax

Slain chimp's owner now says it wasn't on Xanax

STAMFORD, Conn. - As authorities considered criminal charges, the woman whose 200-pound domesticated chimpanzee went berserk and mauled a friend backtracked Wednesday on whether she gave the animal the anti-anxiety drug Xanax.

»

Jackson Sun · 5 hours ago

Click2Houston.com

Owner of chimp that went on Conn. rampage changes story, says she never gave animal Xanax

Sandra Herold told The Associated Press on Wednesday that she never gave the drug to her 14-year-old chimp, Travis, who was shot dead by Stamford police Monday after he grievously wounded Herold's friend Charla Nash.

»

Grand Forks Herald · 7 hours ago

Daily Mail

Owner of chimp that went on Conn. rampage changes story, says she never gave animal Xanax

In humans, Xanax can lead to aggression in people who are unstable to begin with, said Dr. Emil Coccaro, chief of psychiatry at the University of Chicago Medical Center.

»

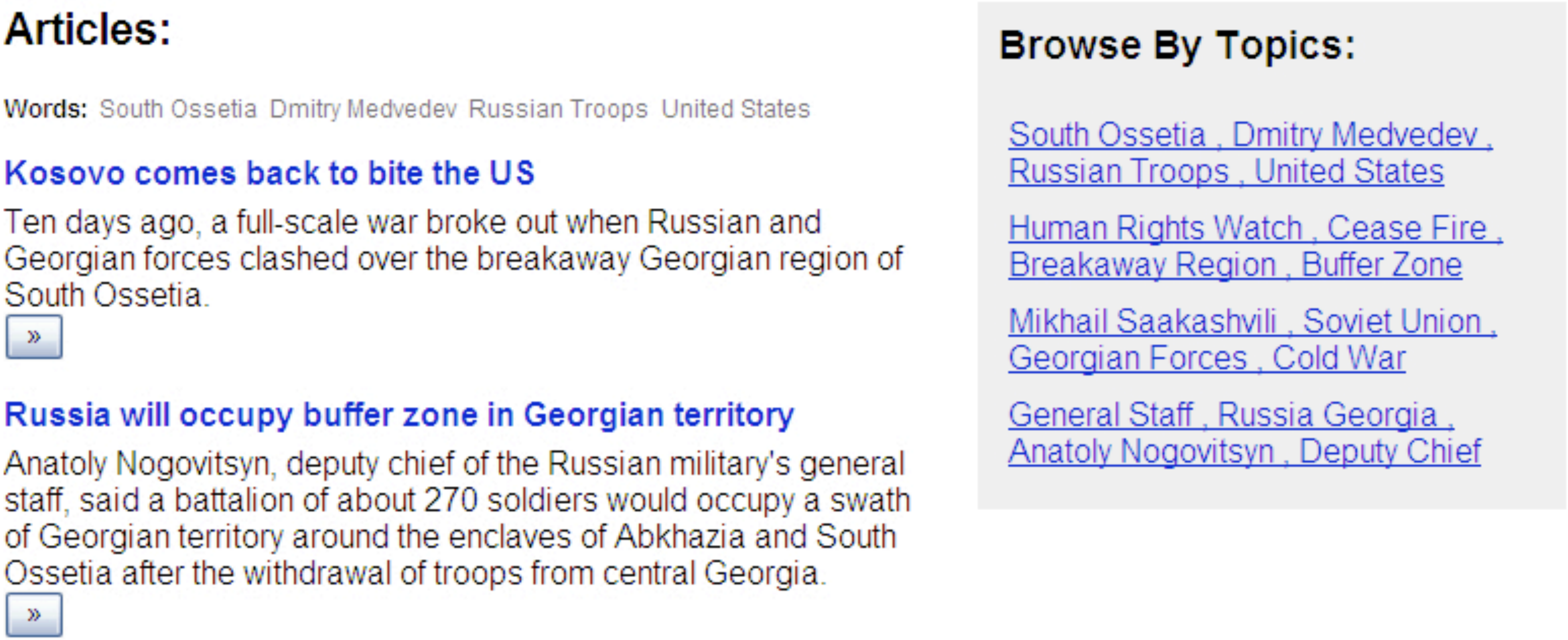
Grand Forks Herald · 4 hours ago

Daily Mail

Refine by topic

- Sandra Herold, Charla Nash, Lyme Disease, Drug Xanax
- Don Mecca, Ice Cream, Years Ago, Started Roaming
- Old Navy, Stamford Police, Frantic Owner, Hand Specialists
- Critical Condition, Frantically Stabbed, Monday Afternoon, Officer Inside

Topical Summarization



Articles:

Words: South Ossetia Dmitry Medvedev Russian Troops United States

Kosovo comes back to bite the US

Ten days ago, a full-scale war broke out when Russian and Georgian forces clashed over the breakaway Georgian region of South Ossetia.

Russia will occupy buffer zone in Georgian territory

Anatoly Nogovitsyn, deputy chief of the Russian military's general staff, said a battalion of about 270 soldiers would occupy a swath of Georgian territory around the enclaves of Abkhazia and South Ossetia after the withdrawal of troops from central Georgia.

Browse By Topics:

South Ossetia , Dmitry Medvedev , Russian Troops , United States

Human Rights Watch , Cease Fire , Breakaway Region , Buffer Zone

Mikhail Saakashvili , Soviet Union , Georgian Forces , Cold War

General Staff , Russia Georgia , Anatoly Nogovitsyn , Deputy Chief