Scale-Invariant Infinite Hierarchical Topic Model

Shusei Eshima
Department of Government, Harvard University
shuseieshima@g.harvard.edu

Daichi Mochihashi
The Institute of Statistical Mathematics
daichi@ism.ac.jp

Problem Statement

- Topic models: summarize, annotate, and categorize documents for a human reader
- Hierarchical topic models: learning hierarchical latent topic organization, especially for a large number of topics (say, > 1000) [1, 2, 3]

Existing models: fragmented tree structures

- Expected probabilities of topics decay exponentially along the depth of tree
- Heuristic rules to update topics
- Restricting the tree structure (e.g., truncating the depth to three levels)

Contributions

1. Adjusting the probability scale by considering the size of the parent topic → less fragmentation
2. Extending the tree-structured stick-breaking process (TSSB) [4] → variety of applications
3. Efficiently drawing an infinite topic tree for each document from a base infinite tree in a hierarchical Bayesian fashion

Scale-Invariant TSSB

TSSB is a crucial building block of recent neural hierarchical topic models [2, 3]. Let $\boldsymbol{\kappa} \prec \boldsymbol{\epsilon}$ indicate that $\boldsymbol{\kappa}$ is an ancestor of $\boldsymbol{\epsilon}$,

$$\pi_{\boldsymbol{\epsilon}} = \nu_{\boldsymbol{\epsilon}} \prod_{\boldsymbol{\kappa} \prec \boldsymbol{\epsilon}} (1-\nu_{\boldsymbol{\kappa}}) \cdot \prod_{\boldsymbol{\kappa} \preceq \boldsymbol{\epsilon}} \phi_{\boldsymbol{\kappa}}\,,\ \phi_{\boldsymbol{\epsilon}k} = \psi_{\boldsymbol{\epsilon}k} \prod_{j=1}^{k-1} (1 - \psi_{\boldsymbol{\epsilon}j})$$

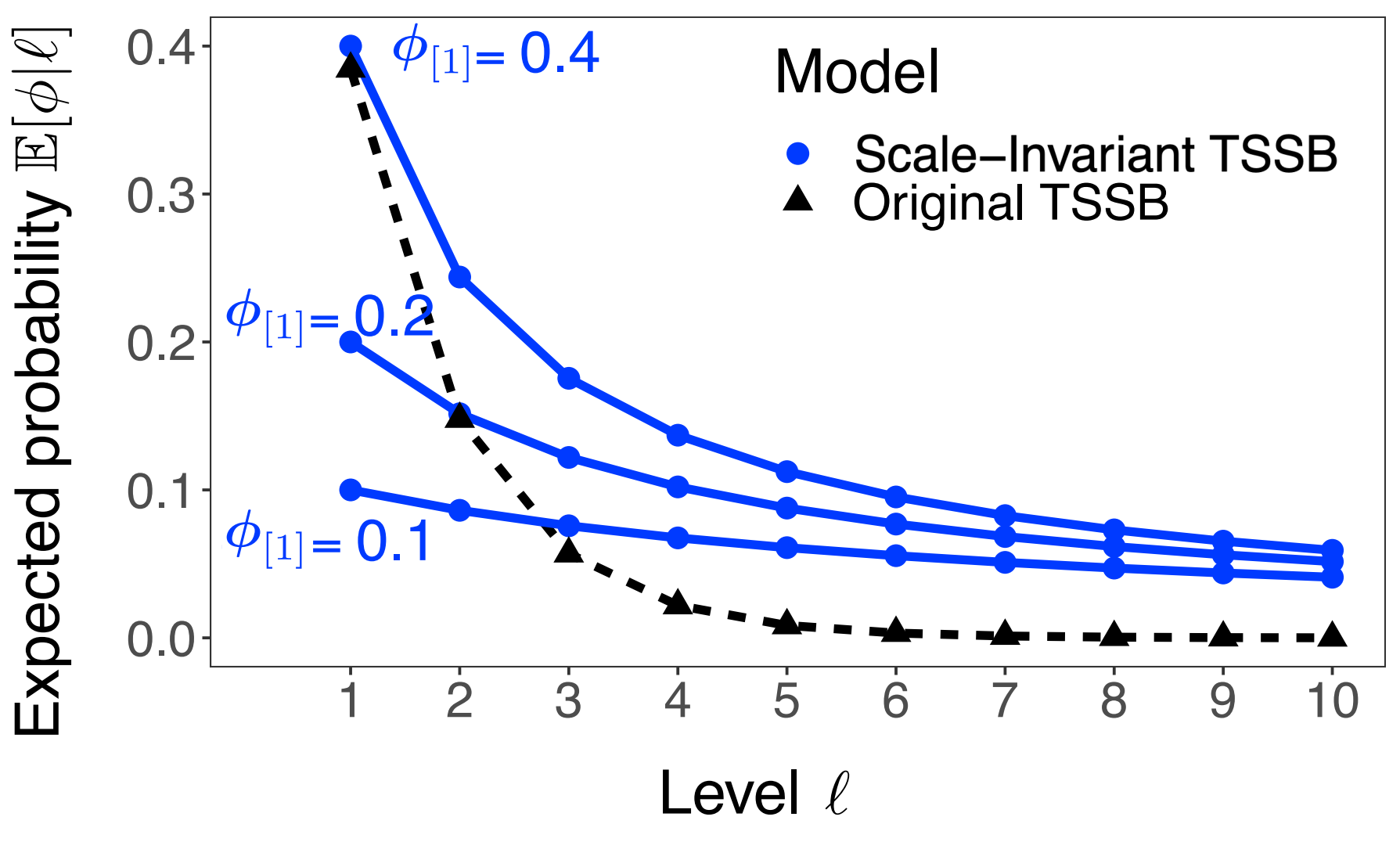
$$\nu_{\boldsymbol{\epsilon}} \sim \mathrm{Be}(1, \alpha_0)\,,\ \psi_{\boldsymbol{\epsilon}} \sim \mathrm{Be}(1, \gamma_0)\,.$$

- The first term: the probability of stopping at the topic $\boldsymbol{\epsilon}$ vertically.
- The next product terms: to passing ancestors of $\boldsymbol{\epsilon}$ while horizontally stopping at $\boldsymbol{\epsilon}$ and its ancestors.
- Vertical and horizontal probabilities of stopping follow Beta distribution.

The scale-invariant TSSB rescales γ_0,

$$\psi_{\boldsymbol{\epsilon}} \sim \mathrm{Be}(1, \phi_{\boldsymbol{\epsilon}'}\gamma_0).$$

Using the horizontal breaking proportion of a parent topic, $\phi_{\boldsymbol{\epsilon}'}$, to draw a *relative* stick length for its child topic → A larger break if the stick to break is shorter. The expected probabilty of horizontal break at ℓ is $\mathbb{E}[\phi\,|\,\ell] \approx 1/(2\gamma+1/\mathbb{E}[\phi\,|\,\ell-1])$ for $\ell \geq 2$. It was originally $\mathbb{E}[\phi\,|\,\ell] \approx 1/(2\gamma+1)^{\ell}$.
SI-TSSB has a slower decay.



References

[1] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum, "Hierarchical topic models and the nested Chinese restaurant process," in *NIPS*, 2003, pp. 17–24.

[2] M. Isonuma, J. Mori, D. Bollegala, and I. Sakata, "Tree-structured neural topic model," in *ACL*, 2020, pp. 800–806.

[3] Z. Chen, C. Ding, Z. Zhang, Y. Rao, and H. Xie, "Tree-structured topic modeling with nonparametric neural variational inference," in *ACL*, 2021, pp. 2343–2353.

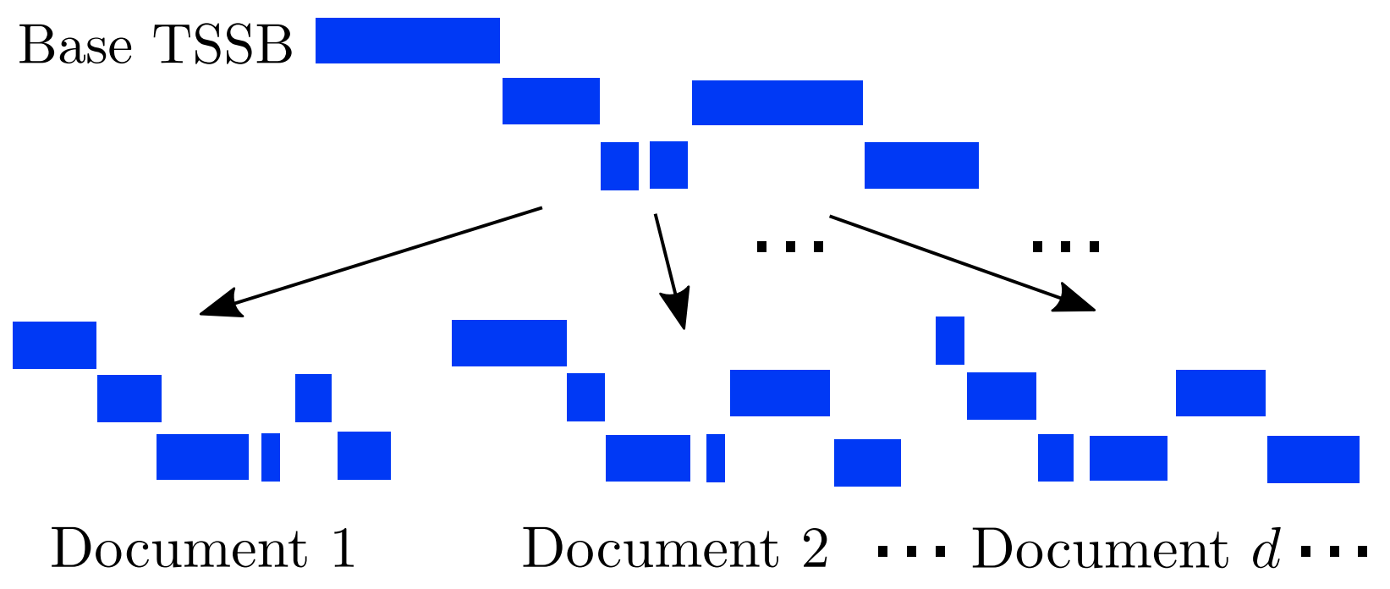
[4] R. P. Adams, Z. Ghahramani, and M. I. Jordan, "Tree-structured stick breaking for hierarchical data," in *NIPS*, 2010, pp. 19–27.

[5] Y. W. Teh, "A Bayesian interpretation of interpolated Kneser-Ney," NUS School of Computing, Tech. Rep., 2006.

Scale-Invariant Infinite Hierarchical Topic Model (ihLDA)

Document-Topic Distribution:

- We apply the hierarchical Dirichlet process separately to the vertical and horizontal probabilities.
$\nu_{\boldsymbol{\epsilon}} \sim \mathrm{Be}(a\tau_{\tilde{\boldsymbol{\epsilon}}}, a(1-\sum_{\boldsymbol{\kappa}\preceq\tilde{\boldsymbol{\epsilon}}} \tau_{\boldsymbol{\kappa}}))$ where $\tau_{\boldsymbol{\epsilon}} = \nu_{\boldsymbol{\epsilon}} \prod_{\boldsymbol{\kappa}\prec\boldsymbol{\epsilon}} (1-\nu_{\boldsymbol{\kappa}})$
$\psi_{\boldsymbol{\epsilon}k} \sim \mathrm{Be}(b\phi_{\tilde{\boldsymbol{\epsilon}}k}, b(1-\sum_{j=1}^{k} \phi_{\tilde{\boldsymbol{\epsilon}}j}))$



Topic-Word Distribution:

- The hierarchical Pitman-Yor process [5]
- The semantic similarity between a parent topic and its children
- Increasing the specificity as the tree deepens

Inference:

- Gibbs sampling, Retrospective sampling + Binary search, Slice sampling

Data Generation Process: $\boldsymbol{\pi}^{(d)}$:= a TSSB for a doc d.

1. Draw a base TSSB $\widetilde{\boldsymbol{\pi}}$.
2. Draw topic-word distributions $H_{\boldsymbol{\epsilon}}$ from the HPY for each topic in $\widetilde{\boldsymbol{\pi}}$.
3. Draw a document-topic distribution for each document d, $\boldsymbol{\pi}^{(d)} \sim \mathrm{HTSSB}(\widetilde{\boldsymbol{\pi}})$.
4. For each word position i in a document d, draw a topic, $z_{di} \sim \boldsymbol{\pi}^{(d)}$ and draw a word, $w_{di} \sim H_{z_{di}}$.

Comparing Top Words from Three Models (max level = 3 for comparison)

Proposed:	Probabilistic: nCRP [1]	Neural: TSNTM [2]
L1: said year would also people	L1: said year one time would	L1: said show year also would
L2: said people mobile technology phone	L2: said year also would company	L2: said year game world time
L3: said software site user mail	L3: film show magic would child	L3: england first game ireland win
L2: said would government people law	L3: film indian star india actor	L3: said labour blair party election
L3: tax said government would budget	L3: film dvd effect extra man	L3: said would people law gov.
L3: labour election said party blair	L3: film harry potter dvd warner	L3: said would gov. election tax
L2: film said best award year	L2: best award film actor actress	L3: said would tax gov. election
L3: music band song year album	L2: film star story life singer	L3: said would tax gov. election
L3: game dvd film year sony	L2: film star movie actress also	L3: said would tax gov. election

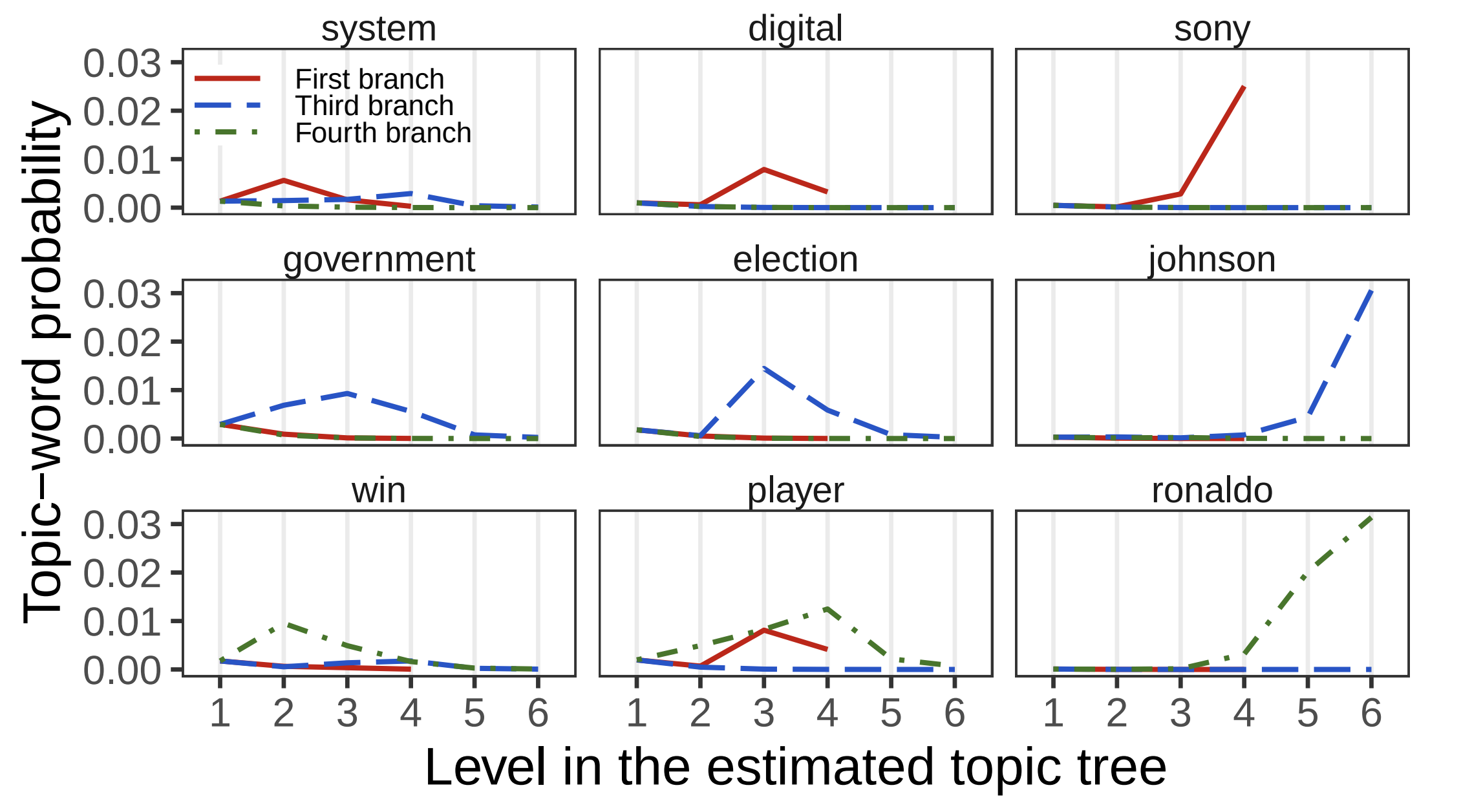
Experiments

Data:

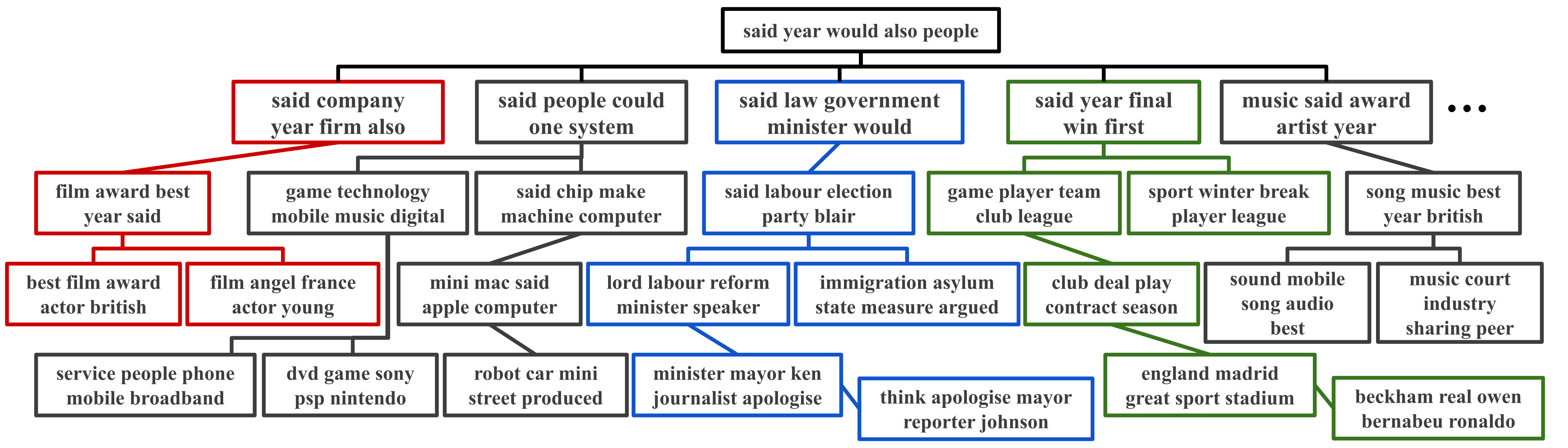
- BBC News: 2,225 documents in five topic areas from the BBC news website
- 20News: a collection of 18,828 posts from 20 USENET newsgroups
- Wikipedia: 50,153 English articles randomly sampled from ten main categories

Setup: Comparing topics with at least 100 assigned words (no truncation in parentheses)

Topic-Word Probabilities (BBC, L6)



Top Words of Selected Topics from the Estimated Topic Tree (BBC, L6)



said year would also people

said company year firm also
said people could one system
said law government minister would
said year final win first
music said award artist year
• • •

film award best year said
game technology mobile music digital
said chip make machine computer
said labour election party blair
game player team club league
sport winter break player league
song music best year british

best film award actor british
film angel france actor young
mini mac said apple computer
lord labour reform minister speaker
immigration asylum state measure argued
club deal play contract season
sound mobile song audio best
music court industry sharing peer

service people phone mobile broadband
dvd game sony psp nintendo
robot car mini street produced
minister mayor ken journalist apologise
think apologise mayor reporter johnson
england madrid great sport stadium
beckham real owen bernabeu ronaldo

Evaluation Metrics

Model	Max Lvl.	Tree Diversity (↑)			Topic Uniqueness (↑)			Average Overlap (↓)			# of Topics		
		BBC	20News	Wiki	BBC	20News	Wiki	BBC	20News	Wiki	BBC	20News	Wiki
ihLDA	3	2.24 (2.24)	**2.88** (2.86)	**2.63** (2.49)	**0.60** (0.60)	**0.82** (0.80)	**0.66** (0.63)	0.28 (0.28)	0.11 (0.14)	0.16 (0.19)	38 (38)	27 (31)	17 (18)
	≥ 4	**2.53** (2.54)	**2.88** (2.80)	2.50 (2.51)	0.55 (0.49)	0.76 (0.51)	0.65 (0.63)	0.26 (0.30)	0.12 (0.38)	0.15 (0.16)	85 (134)	67 (203)	73 (101)
nCRP	3	1.92	2.16	–	0.36	0.32	–	**0.03**	0.02	–	517	2108	–
rCRP		0.15	–	–	0.01	–	–	0.53	–	–	278	–	–
TSNTM		1.98	2.54	2.47	0.43	0.80	0.64	0.26	0.09	0.06	22	41	44
nTSNTM		2.11	2.57	2.34	0.46	0.68	0.60	0.09	**0.01**	**0.02**	68	81	111

- Tree Diversity: uniqueness of topics while considering the importance of the parent topics
- Topic Uniqueness [3]: uniqueness of all topics
- Average Overlap [3]: the average repetition rate of the top words (less overlap does not necessarily mean better interpretability)
- The results without truncation are shown in parentheses for the proposed model.