HyperMiner: Topic Taxonomy Mining with Hyperbolic Embedding

Yishi Xu[1] Dongsheng Wang[1] Bo Chen[1] Ruiying Lu[1] Zhibin Duan[1] Mingyuan Zhou[2]

1 Xidian University

2 The University of Texas at Austin

Source code: https://github.com/NoviceStone/HyperMiner

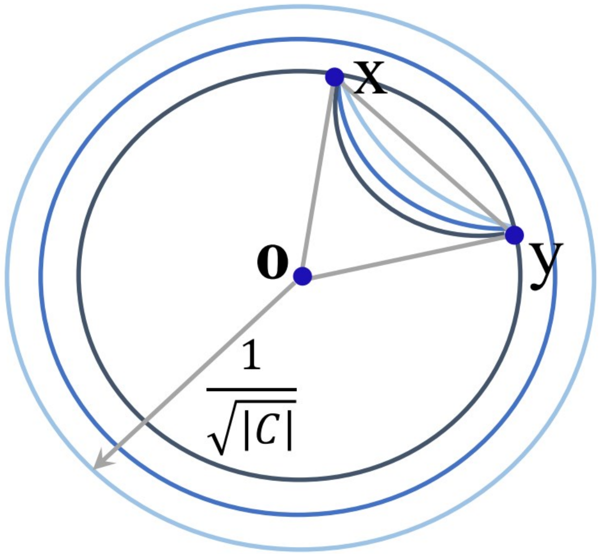
Motivation

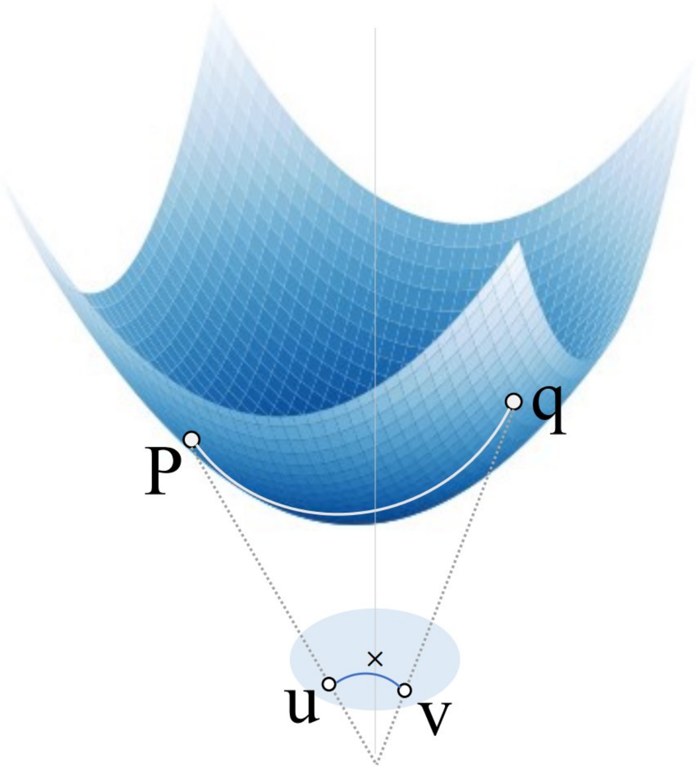
Existing embedded topic models generally hold the Euclidean embedding space assumption, leading to a fundamental limitation in capturing hierarchical relationships given that

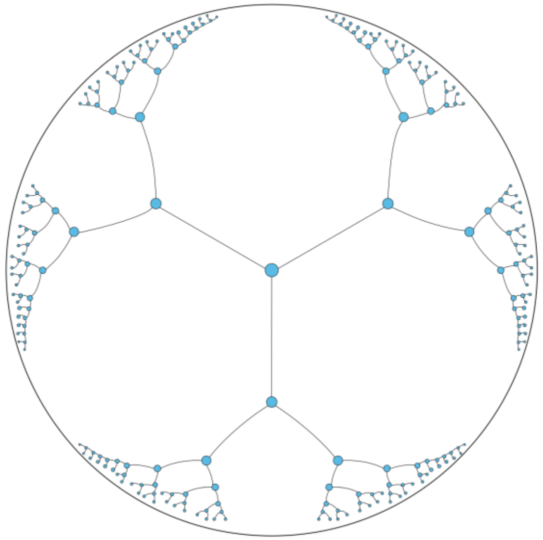
- The lexical hierarchy naturally exists for the words of vocabulary
- A semantic hierarchy is also expected between topics and words

Motivation

Hyperbolic geometry has shown superior performance in modeling hierarchical data, with the tree-likeness property of its distance metric.







Approach

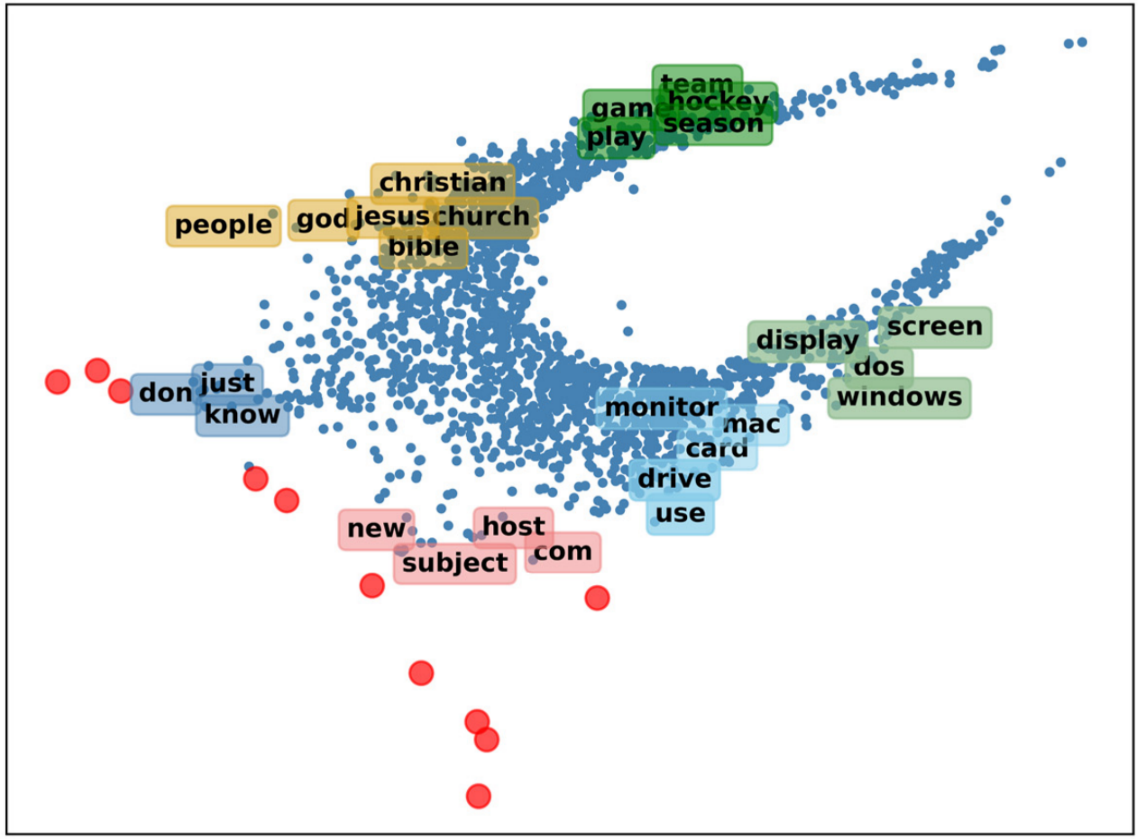
Embed both words and topics into a shared hyperbolic space instead of Euclidean space, so that the hyperbolic distance metric can be used to measure the semantic similarity between topic and words. Note that

- $\boldsymbol{\beta}_k \in \mathbb{R}^V$: distribution of words for the k-th topic
- $\boldsymbol{\alpha}_k \in \mathbb{R}^D$: vector representation of the k-th topic
- $\boldsymbol{\rho} \in \mathbb{R}^{D \times V}$: word embeddings of the vocabulary

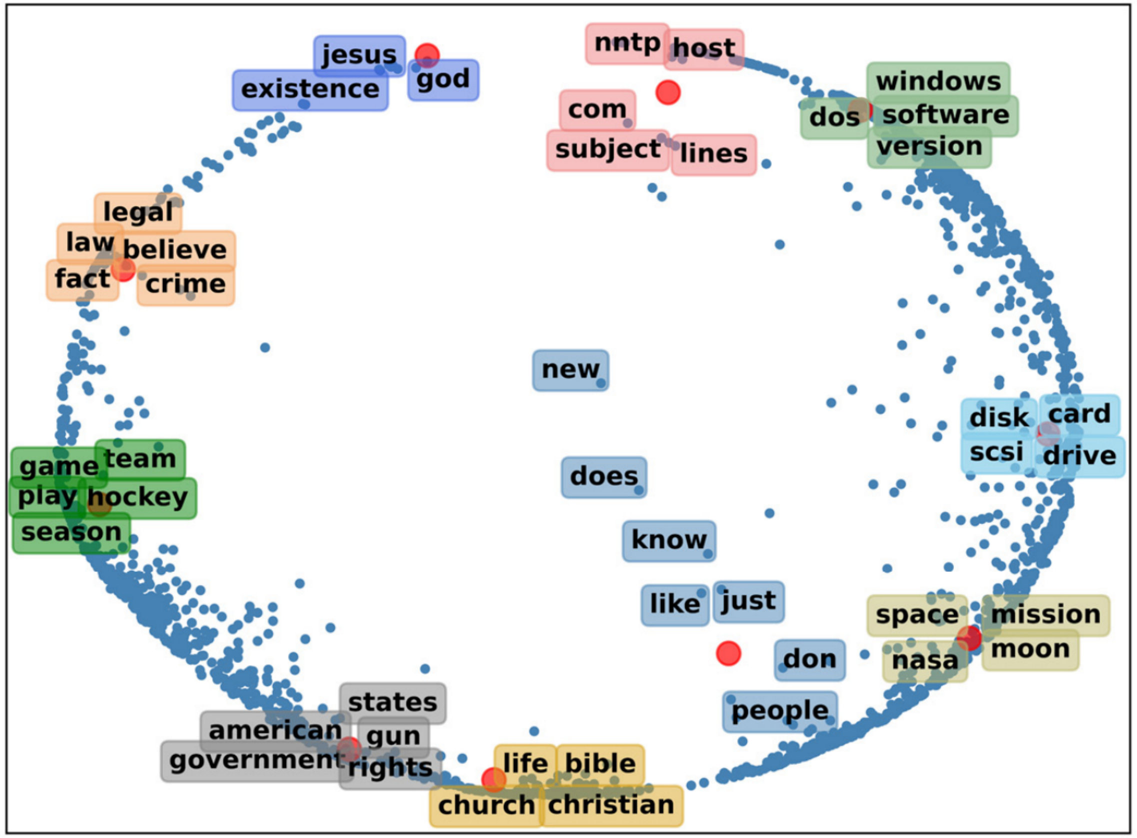
$$\boldsymbol{\beta}_k = \mathrm{Softmax}\big(\mathrm{dist}(\boldsymbol{\rho}, \boldsymbol{\alpha}_k)\big)$$

Inner product for Euclidean embeddings, replace it with distance metric for hyperbolic embeddings.

Implicit semantic hierarchy mining

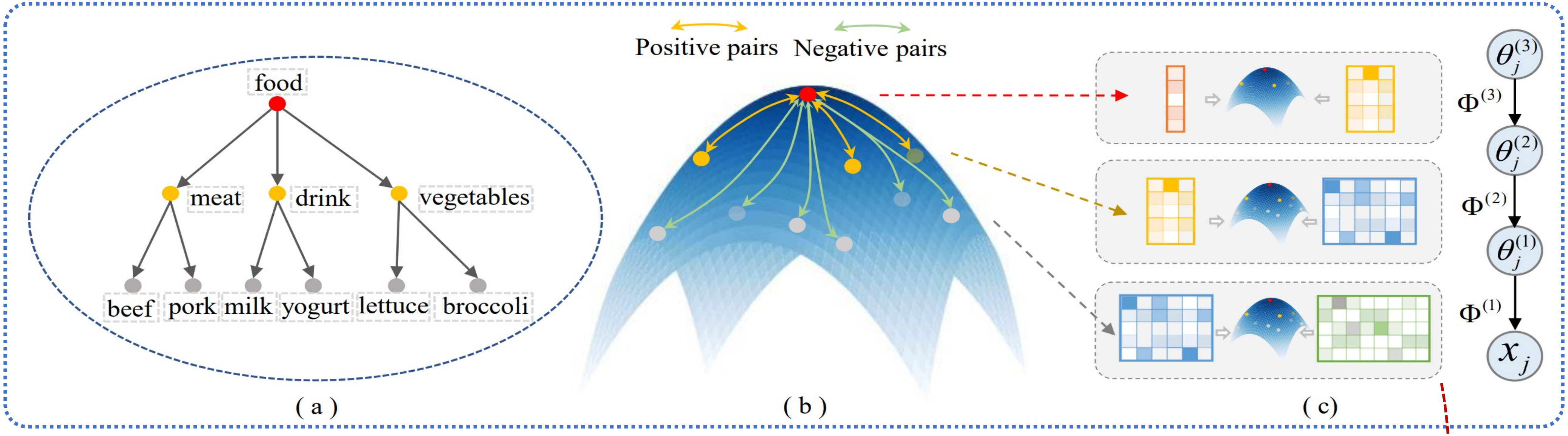


2D Euclidean embedding space learned by ETM



2D hyperbolic embedding space learned by HyperETM

Knowledge-guided topic modeling



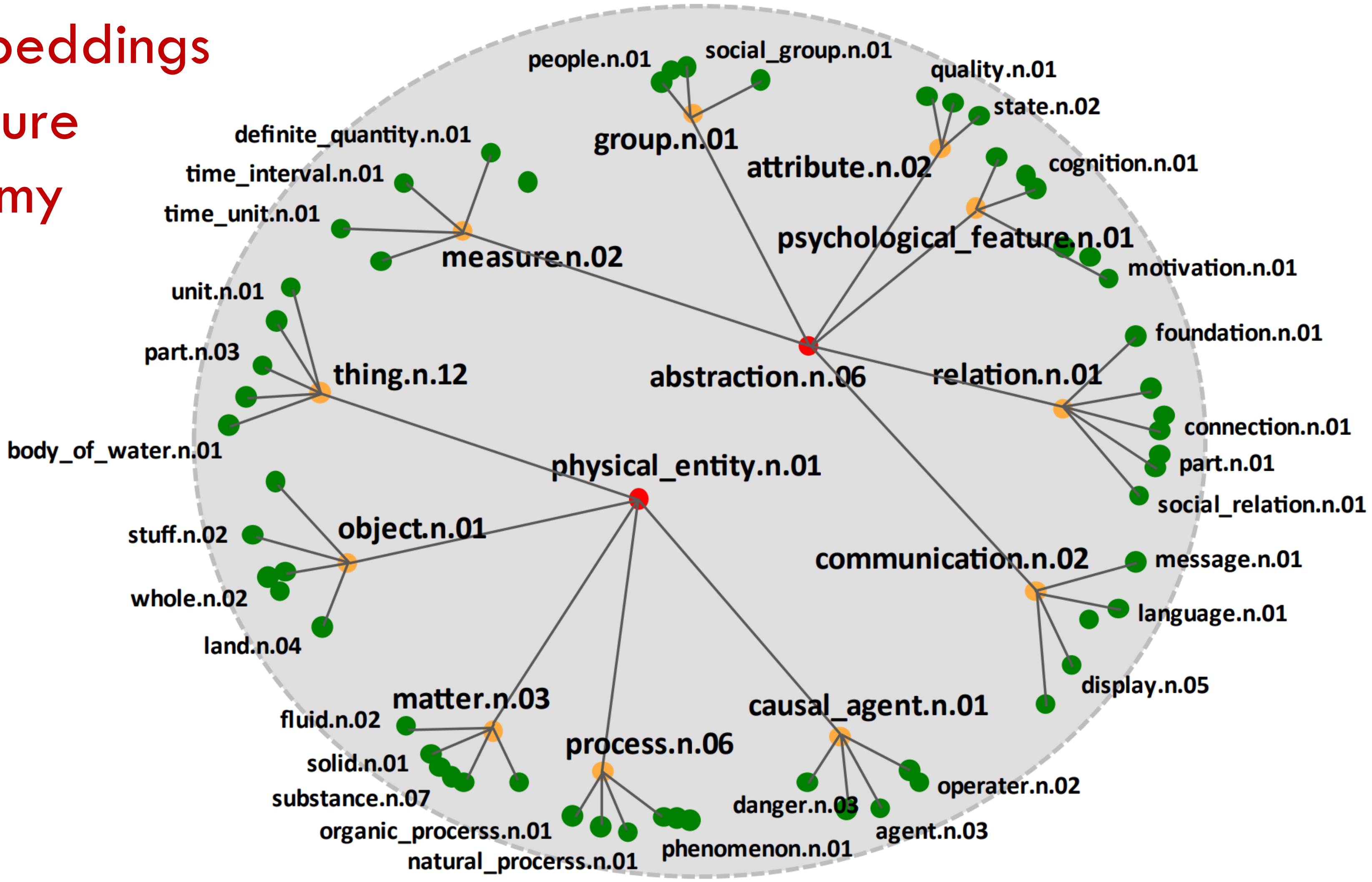
Injecting tree-structure knowledge via contrastive loss

$$\mathcal{L}_{\text{Contra}} = \mathbb{E}_{\boldsymbol{\alpha}_i \in \mathcal{T}}\left[-\log \frac{\exp\left(\text{dist}\left(\boldsymbol{\alpha}_i, \boldsymbol{\alpha}_i^{+}\right)/\tau\right)}{\exp\left(\text{dist}\left(\boldsymbol{\alpha}_i, \boldsymbol{\alpha}_i^{+}\right)/\tau\right) + \sum_{\boldsymbol{\alpha}_i^{-} \in \mathcal{Q}(\boldsymbol{\alpha}_i)} \exp\left(\text{dist}\left(\boldsymbol{\alpha}_i, \boldsymbol{\alpha}_i^{-}\right)/\tau\right)}\right]$$

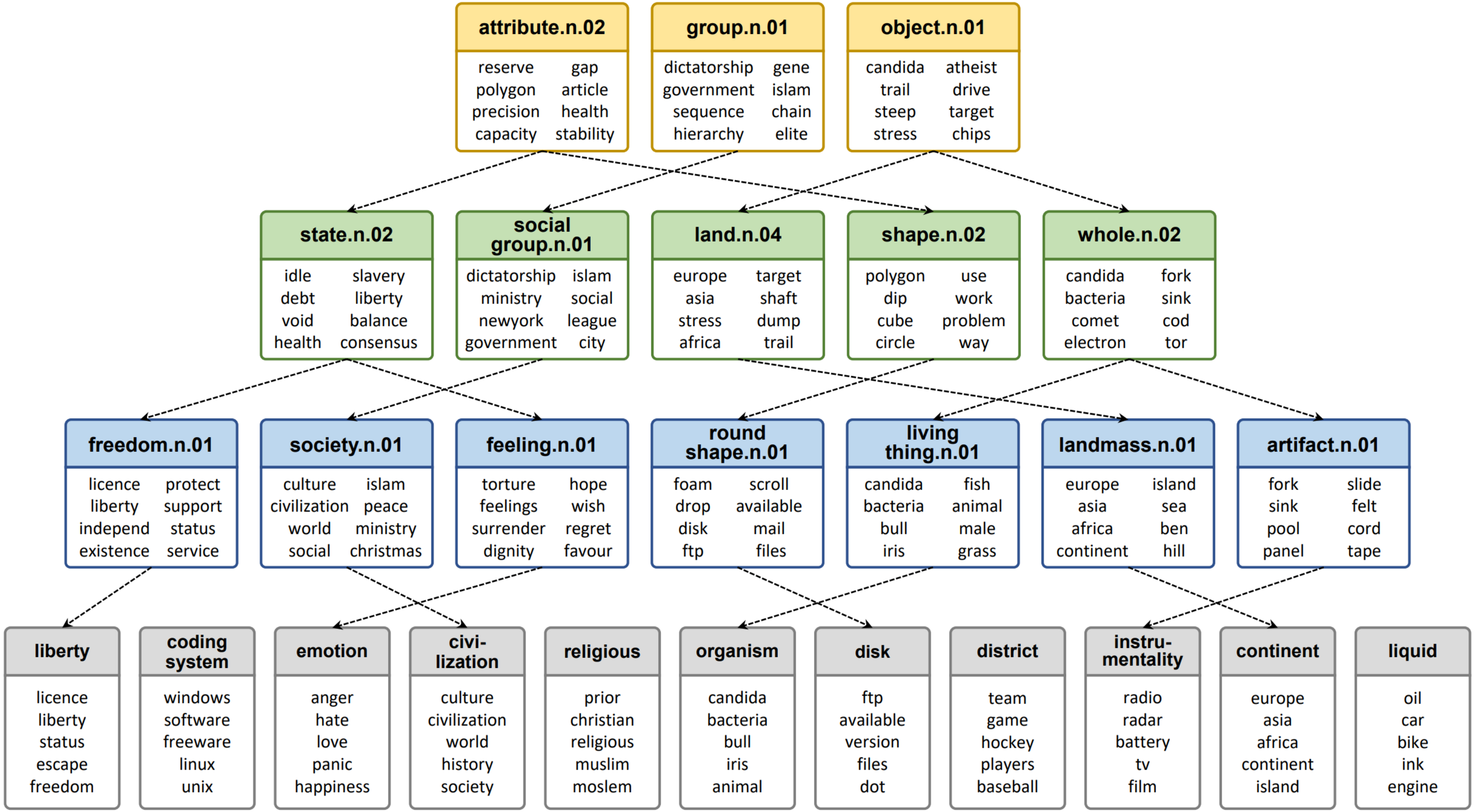
$\boldsymbol{\Phi}^{(l)} = \text{Softmax}\left(\text{dist}\left(\boldsymbol{\alpha}^{(l-1)}, \boldsymbol{\alpha}^{(l)}\right)\right)$ $\boldsymbol{\alpha}^{(l)}$: topic embeddings at layer l

Topic hierarchy visualization

Distribution of topic embeddings well preserves the structure of prior concept taxonomy



Mined topic taxonomy



Clustering performance

- Our approach yields better document representations

Method	20NG		TMN	
	km-Purity	km-NMI	km-Purity	km-NMI
LDA [7]	38.43 ± 0.52	35.98 ± 0.39	48.17 ± 0.69	30.96 ± 0.78
ProdLDA [16]	39.21 ± 0.63	36.52 ± 0.51	55.28 ± 0.67	35.57 ± 0.72
ETM [27]	42.68 ± 0.71	37.72 ± 0.64	59.35 ± 0.74	38.75 ± 0.86
WHAI [17]	40.89 ± 0.35	38.90 ± 0.27	58.06 ± 0.45	37.34 ± 0.48
SawETM [28]	43.36 ± 0.48	41.59 ± 0.62	61.13 ± 0.56	40.78 ± 0.63
TopicNet [31]	42.94 ± 0.41	40.76 ± 0.53	60.52 ± 0.50	40.09 ± 0.54
HyperETM	43.63 ± 0.51	39.06 ± 0.64	61.22 ± 0.62	40.52 ± 0.71
HyperMiner	$\underline{44.37} \pm 0.38$	$\underline{42.83} \pm 0.45$	$\underline{62.96} \pm 0.48$	$\underline{41.93} \pm 0.52$
HyperMiner-KG	$\mathbf{45.16} \pm 0.35$	$\mathbf{43.65} \pm 0.39$	$\mathbf{63.84} \pm 0.43$	$\mathbf{42.81} \pm 0.47$

Thank you.

Please feel free to contact us by e-mail

xuyishi@stu.xidian.edu.cn bchen@mail.xidian.edu.cn

mingyuan.zhou@mccombs.utexas.edu

Paper can be downloaded from