PhraseCTM: Correlated Topic Modeling on Phrases within Markov Random Fields

Input: Text corpus with phrases extracted by AutoPhrase [1] **Output:** Phrase-level topics and the correlation among them

Semantically Coherent Links for MRF

- Motivations:
  - It's nontrivial to apply CTM directly on phrases: (1) phrases are much less than words; (2) CTM doesn't perform well on short documents.
- Some observations:
  - the topic of a phrase is highly related to the topics of other words and phrases in the same document.
  - some phrases' meaning can be implied from their component words.

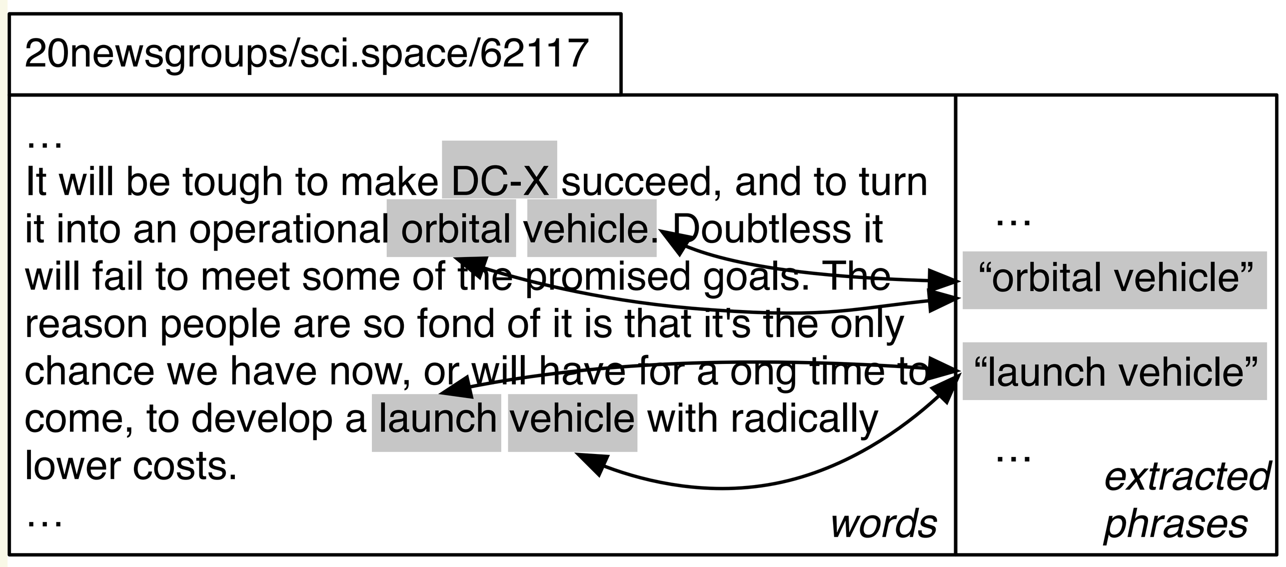


Figure: The arrows show semantically coherent links for MRF.

- Not all phrases can be implied by their component words.
  - *e.g.*, the newspaper *Boston Globe* [2].
- Semantically coherent links
  - Format a document as "words, phrases, semantically coherent links between phrases and component words".
  - determine the semantic coherent links between $w_i^{(\mathcal{P})}$ and $w_{l(i)}$ by utilizing NPMI, $s(w_i^{(\mathcal{P})}, w_{l(i)}) = \min_{j,k \in l(i)}\{\text{NPMI}(w_j, w_k)\} > \tau = 0.4$

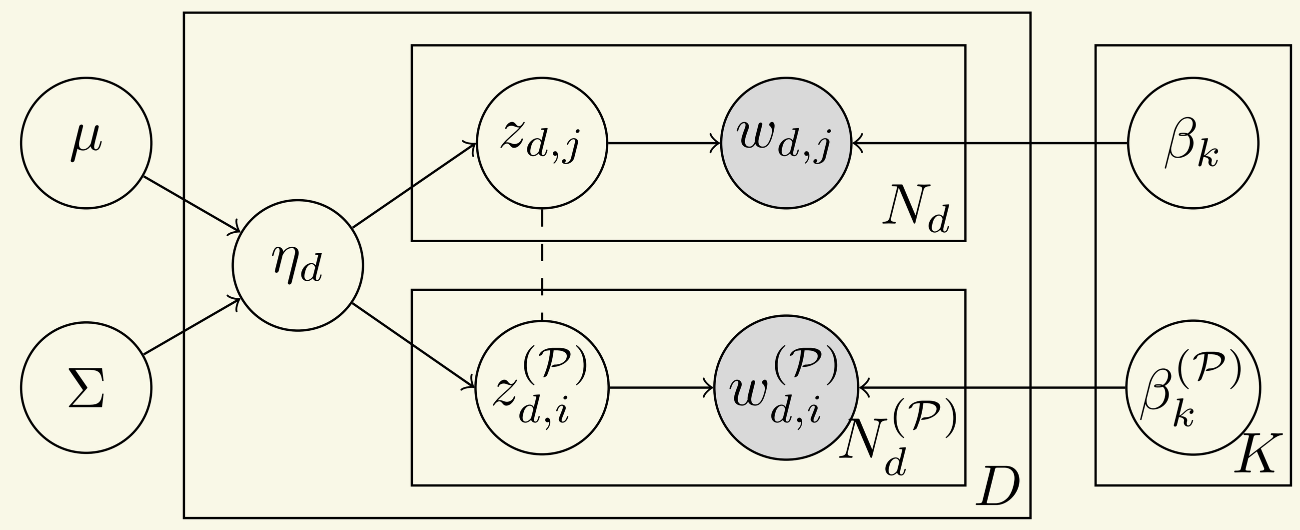
PhraseCTM

In a Markov Random Field of document d, we have

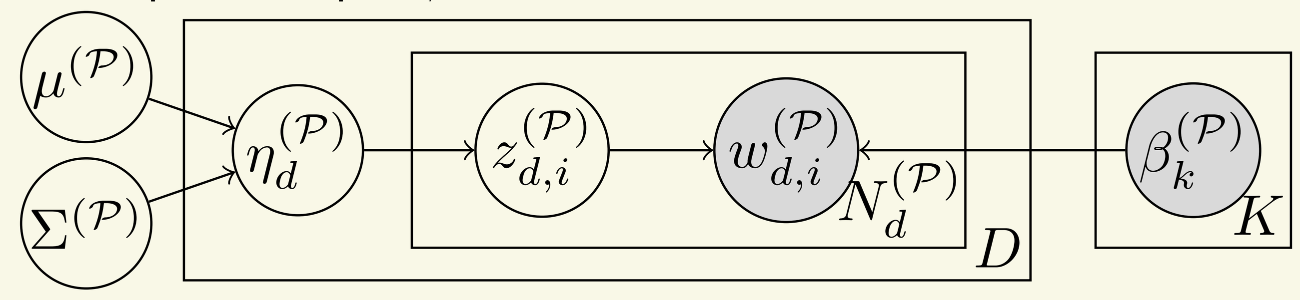
$$p(z_d, z_d^{(\mathcal{P})}|\eta_d) = \frac{1}{A_d(\eta_d)} \prod_{m=1}^{N_d} p(z_{d,m}|\eta_d) \cdot \prod_{i=1}^{N_d^{(\mathcal{P})}} p(z_{d,i}^{(\mathcal{P})}|\eta_d) \cdot \exp\{\sum_{i=1}^{N_{L_d}} (\frac{\kappa}{|l(d,i)|} \sum_{j \in l(d,i)} I(z_{d,i}^{(\mathcal{P})} = z_{d,j}))\}$$

, and capture the correlation between topics like CTM:

$$p(z_{d,j} = k|\eta_d) = \exp \eta_{d,k} / \sum_k \exp \eta_{d,k} \ , \quad \eta_d \sim \mathcal{N}(\mu, \Sigma)$$



(a) The first stage: training on our proposed model PhraseCTM. When observed words W and phrases $W^{(\mathcal{P})}$, we learn word topics β, and phrase topics $\beta^{(\mathcal{P})}$.



(b) The second stage: inferring the phrase topics' correlation. When given the phrases $W^{(\mathcal{P})}$, and the phrase topics $\beta^{(\mathcal{P})}$ learned from the first stage, we infer $\Sigma^{(\mathcal{P})}$ as the correlation result.

Figure: Illustration of two stages of our method

We solve PhraseCTM by variational inference, and get the correlation $corr^{(\mathcal{P})}(i,j) = \frac{\Sigma_{i,j}^{(\mathcal{P})}}{\sqrt{\Sigma_{i,i}^{(\mathcal{P})}\Sigma_{j,j}^{(\mathcal{P})}}}$.

Dataset

	$\|V\|$	$\|V^{(\mathcal{P})}\|$	$\|W\|$	$\|W^{(\mathcal{P})}\|$	$\|D\|$	$\|W\|/\|D\|$	$\|W^{(\mathcal{P})}\|/\|D\|$
20 Newsgroup	22,787	4,245	1,361,843	51,024	18,828	72.3	2.7
Argentina@Wiki	20,847	5,505	1,052,674	98,502	8,617	122.2	11.4
Mathematics@Wiki	43,779	27,371	6,062,815	594,704	27,947	216.9	21.3
Chemistry@Wiki	76,265	67,979	11,346,781	1,546,088	60,375	187.9	25.6
PubMed Abstracts	34,125	24,233	11,274,350	968,928	99,214	113.6	9.8

Table: The statistics of the datasets. In average, phrases appear more sparse than words. Phrases are extracted by AutoPhrase [1].

An Example

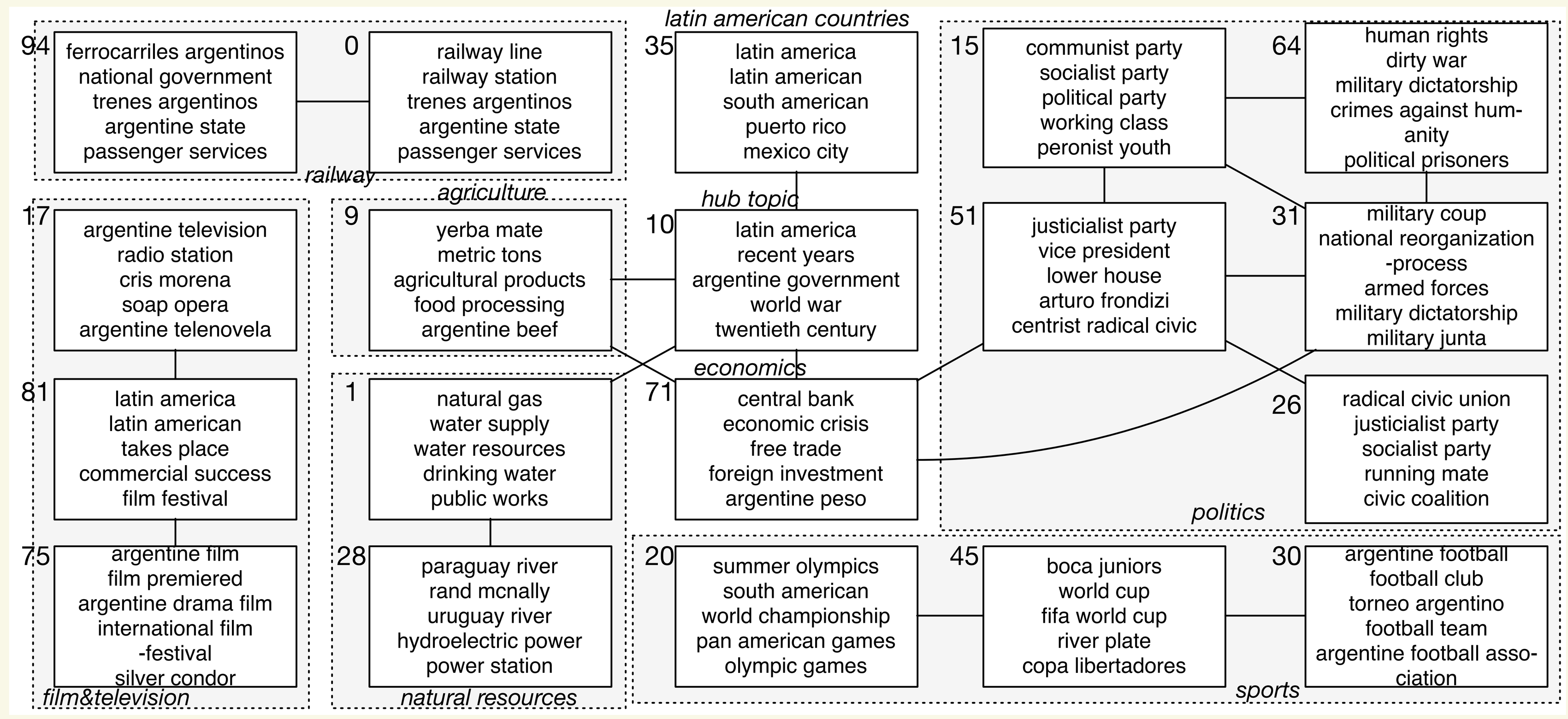


Figure: A part of the topic graph (K=100) generated by our method on the Argentina-related Wikipedia pages.

Human Study

	CTM		PhraseCTM	
	Maths	Argentina	Maths	Argentina
Group A	12.4	-	-	7.5
Group B	-	14.0	6.7	-
In Average	13.2		7.1	

Table: Human time consumption on topic labeling for correlated topics generated by CTM and PhraseCTM, measured in minutes.

Quantitative Result

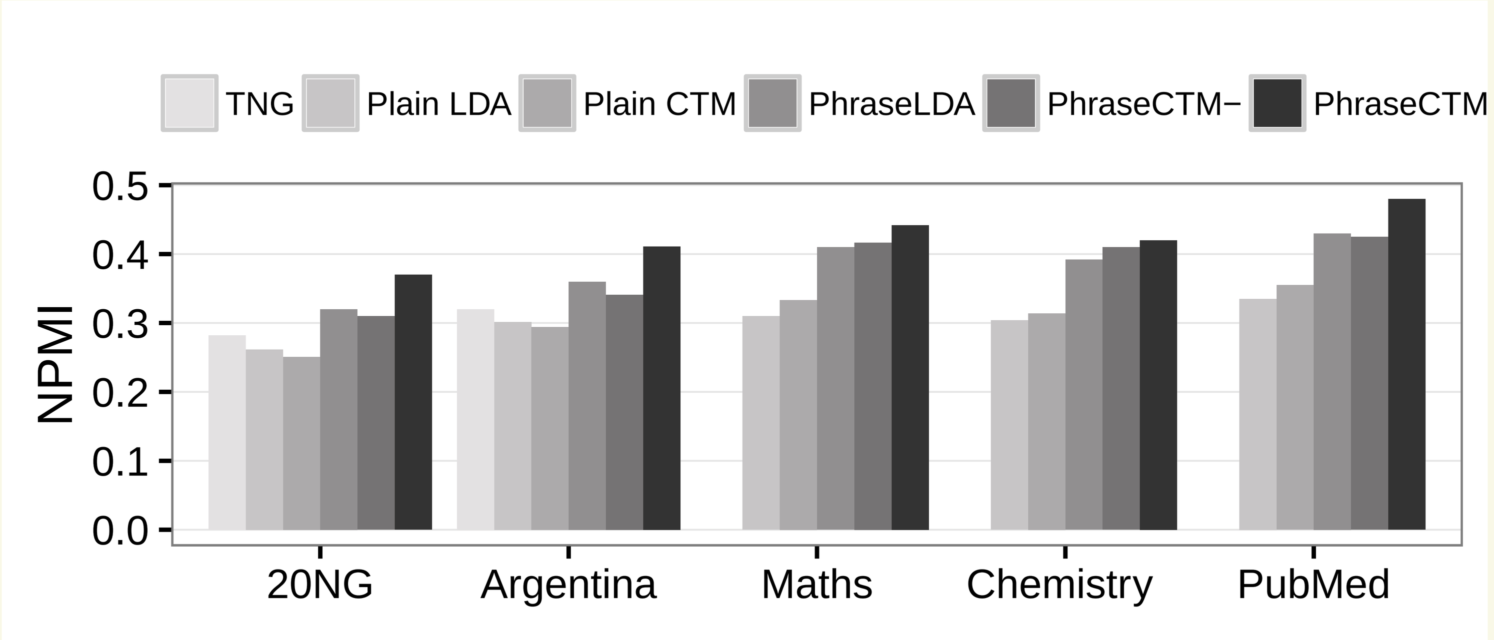


Figure: The quality of the learned topics.

References

[1] Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. Automated phrase mining from massive text corpora. In: *TKDE 2018*.

[2] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In : *NIPS 2013*.

Weijing Huang, Tengjiao Wang, Wei Chen, Siyuan Jiang, Kam-Fai Wong

huangwaleking@gmail.com, tjwang,pekingchenwei@pku.edu.cn, sjiang1@nd.edu, kfwong@se.cuhk.edu.hk

Peking University, China; University of Notre Dame, USA; The Chinese University of Hong Kong, Hong Kong