

Incorporating **Social Context** and **Domain Knowledge** for **Entity Recognition**

Jie Tang, Zhanpeng Fang

Department of Computer Science, Tsinghua University

Jimeng Sun

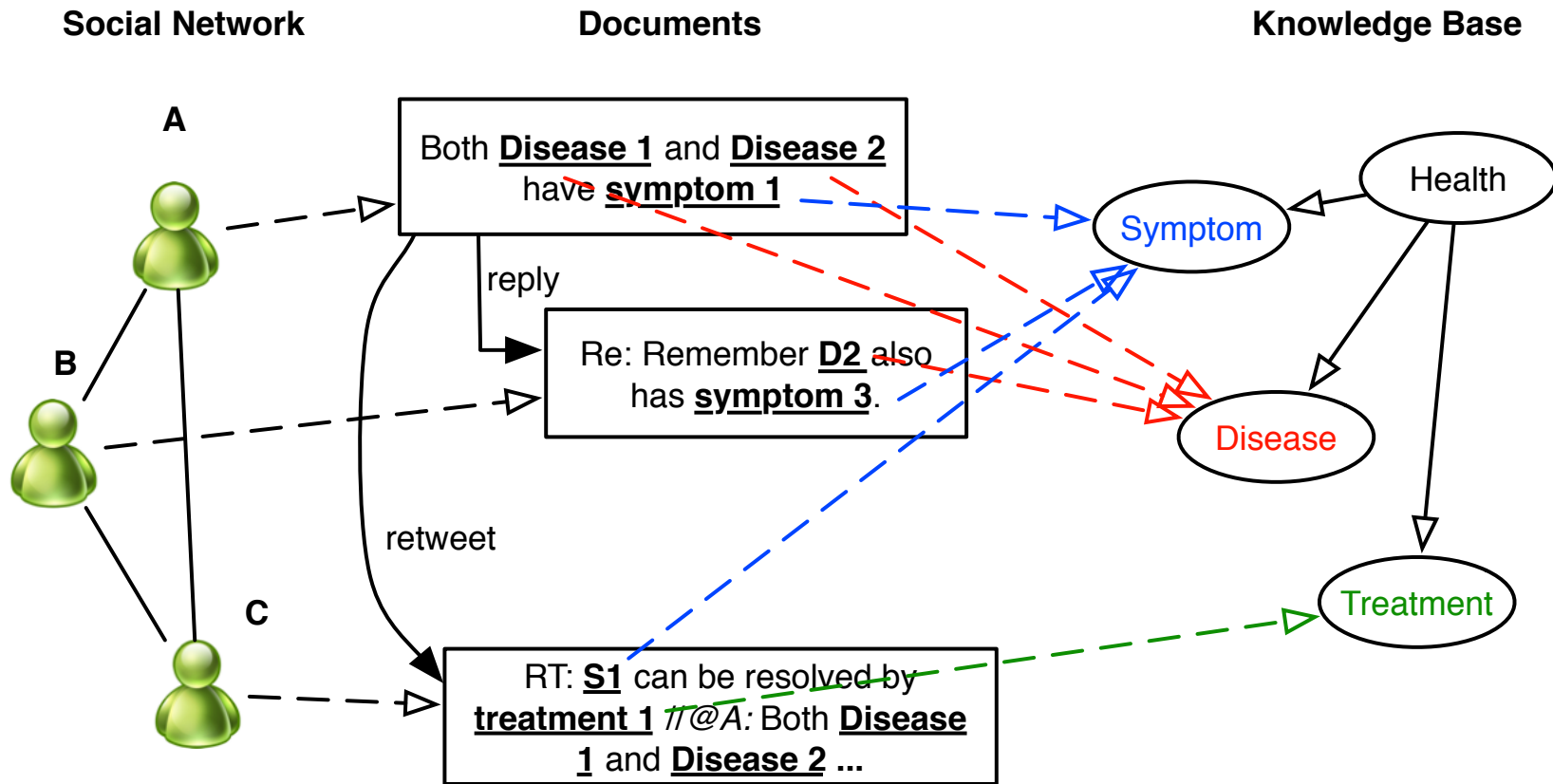
College of Computing, Georgia Institute of Technology

Entity Recognition in Social Media



- People use **blogs**, **forums**, and **review sites** to share opinions on politicians or products.
- One fundamental analytic issue is to recognize **entity instances** from the **UGC short documents**. However, the problem is very challenging
 - “**S4**” vs. “**Samsung Galaxy S4**”
 - “**Fruit company**” vs. “**Apple Inc.**”
 - “**Peace West King**” vs. “**Xilai Bo**” (a sensitive Chinese politician)
 - ...

A Concrete Example



Challenges: short text + social networks + domain knowledge = ?



Related Work

- **Entity recognition**

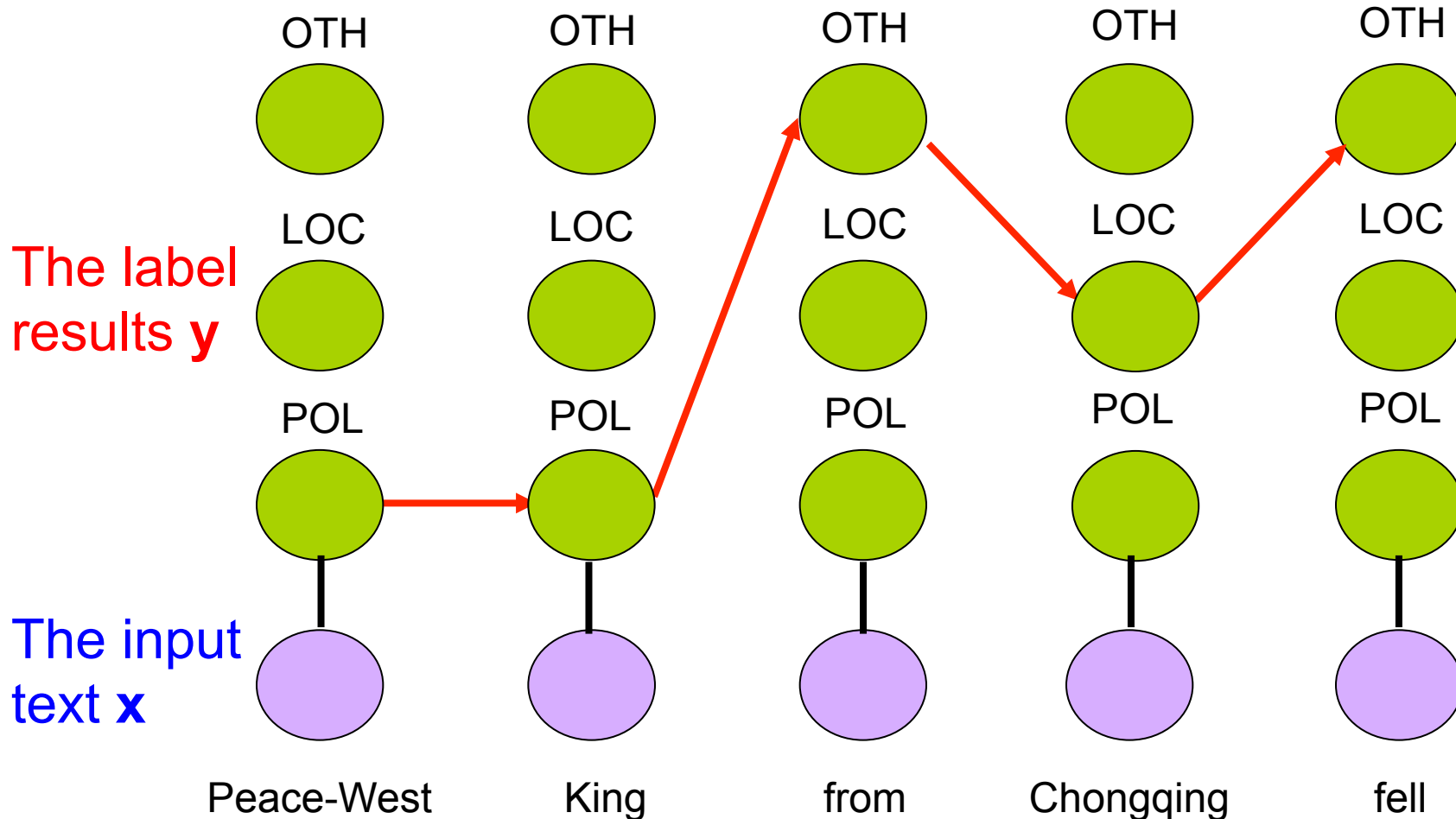
- Modeling as a ranking problem based on boosting and voted perceptron (Collins [9])
- Incorporating long-distance dependency (Finkel et al. [13])
- Use Labeled LDA [26] to exploit Freebase to help extraction (Ritter et al. [27])
- Entity morph (Huang et al. [17])

- **Entity resolution**

- A collective method for entity resolution in relational data (Bhattacharya and Getoor [4])
- A hierarchical topic model for resolving name ambiguity (Kataria et al. [18])
- Name disambiguation in digital libraries (Tang et al. [32])

Approach Framework —SOCINST

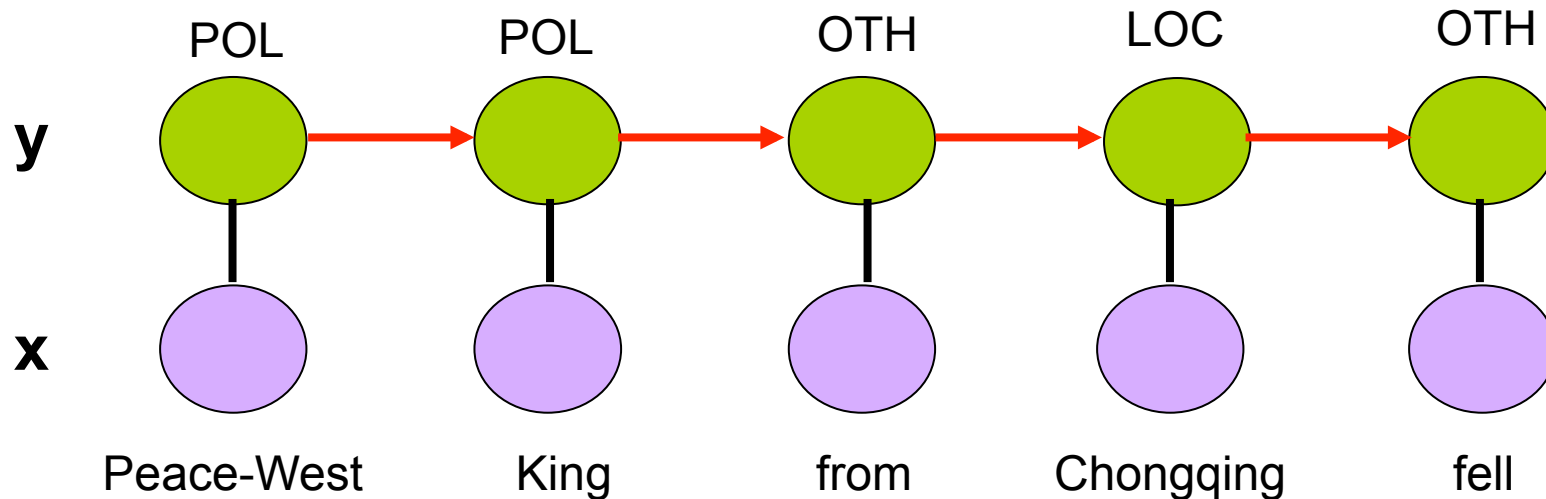
Preliminary: Sequential Labeling



$$\mathbf{y}^* = \max_{\mathbf{y}} p(\mathbf{y} | \mathbf{x}; f, \Theta)$$

where f represents features and Θ are model parameters.

Sequential Labeling with CRFs



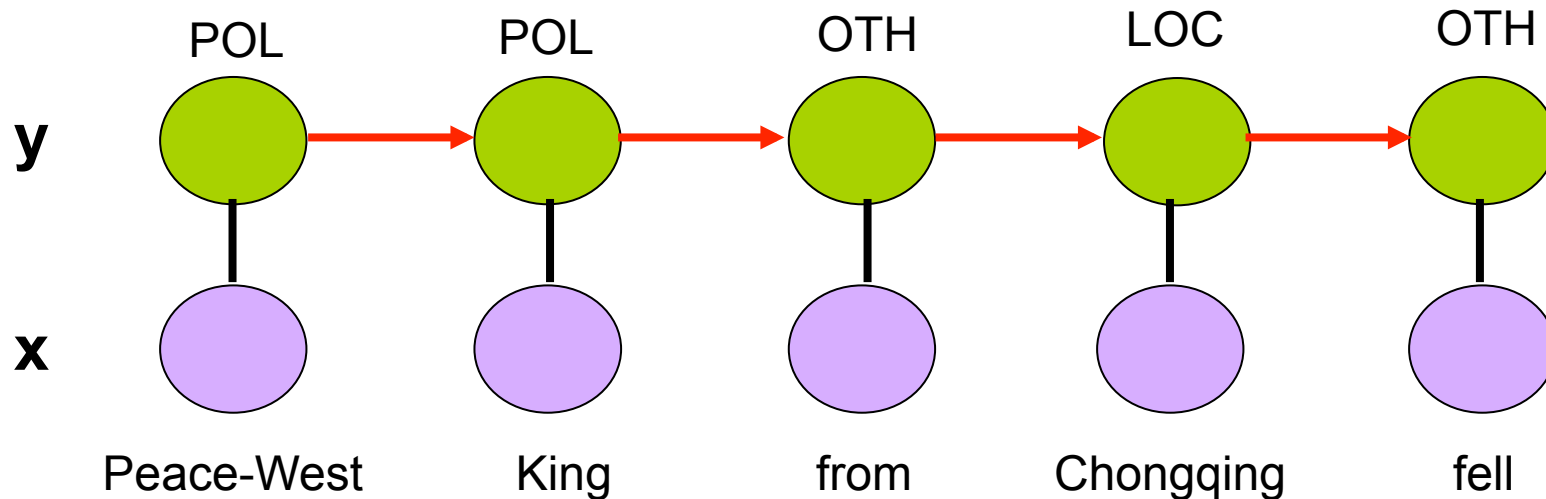
μ and λ are parameters to be learned from the training data.

$$p(\mathbf{y} | \mathbf{x}, \lambda, \mu) = \frac{1}{Z} \exp\left(\sum_i \sum_k \lambda_k f_k(x_i, y_i) + \sum_i \sum_j \mu_j f_j(\mathbf{x}, y_i, y_{i+1})\right)$$

f_k denotes the k -th feature defined for token x_i

f_j denotes the j -th feature defined for two consecutive tokens x_i and x_{i+1}

Sequential Labeling with CRFs

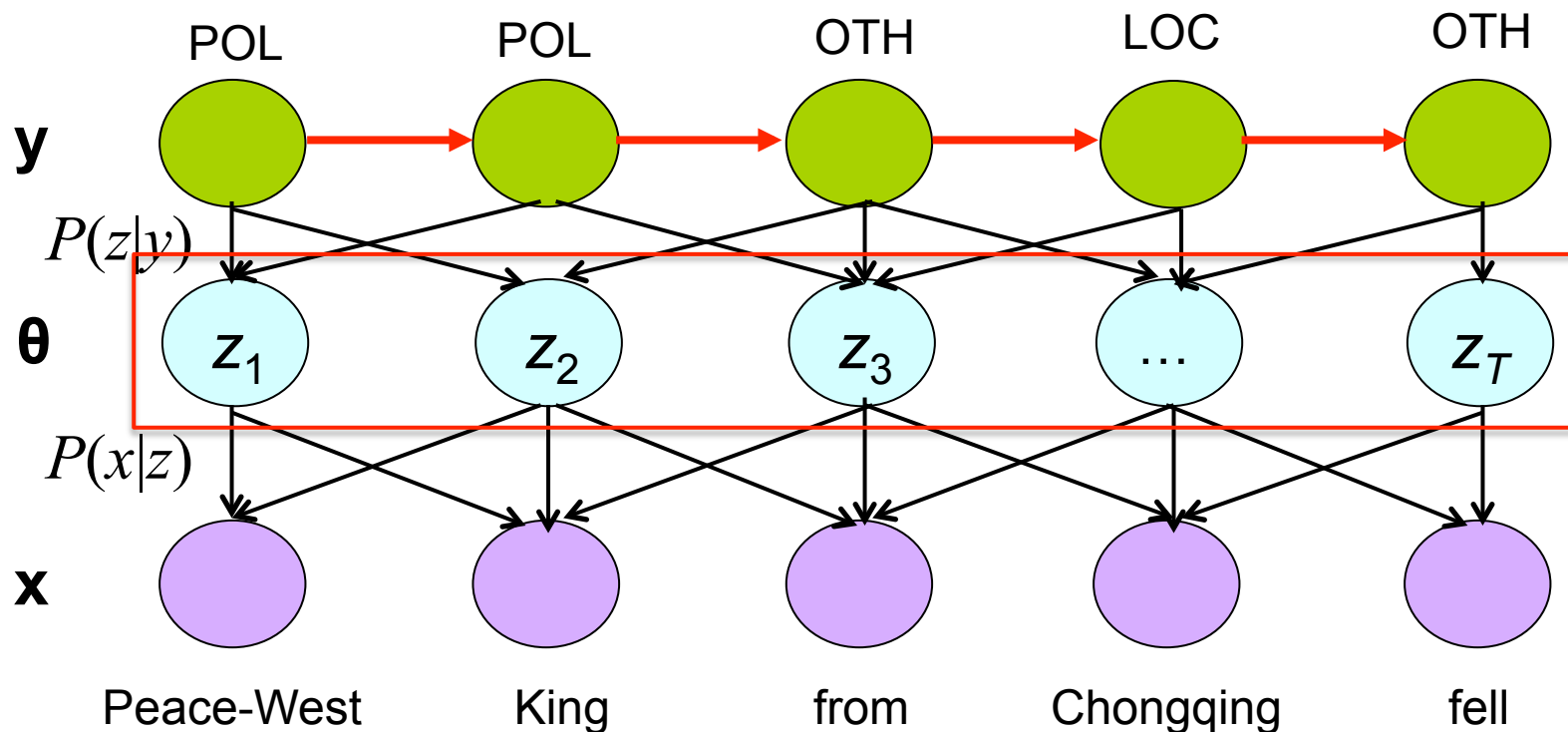


Performance of the model will be bad when dealing with short-text due to sparsity

f_k denotes the k -th feature defined for token x_i

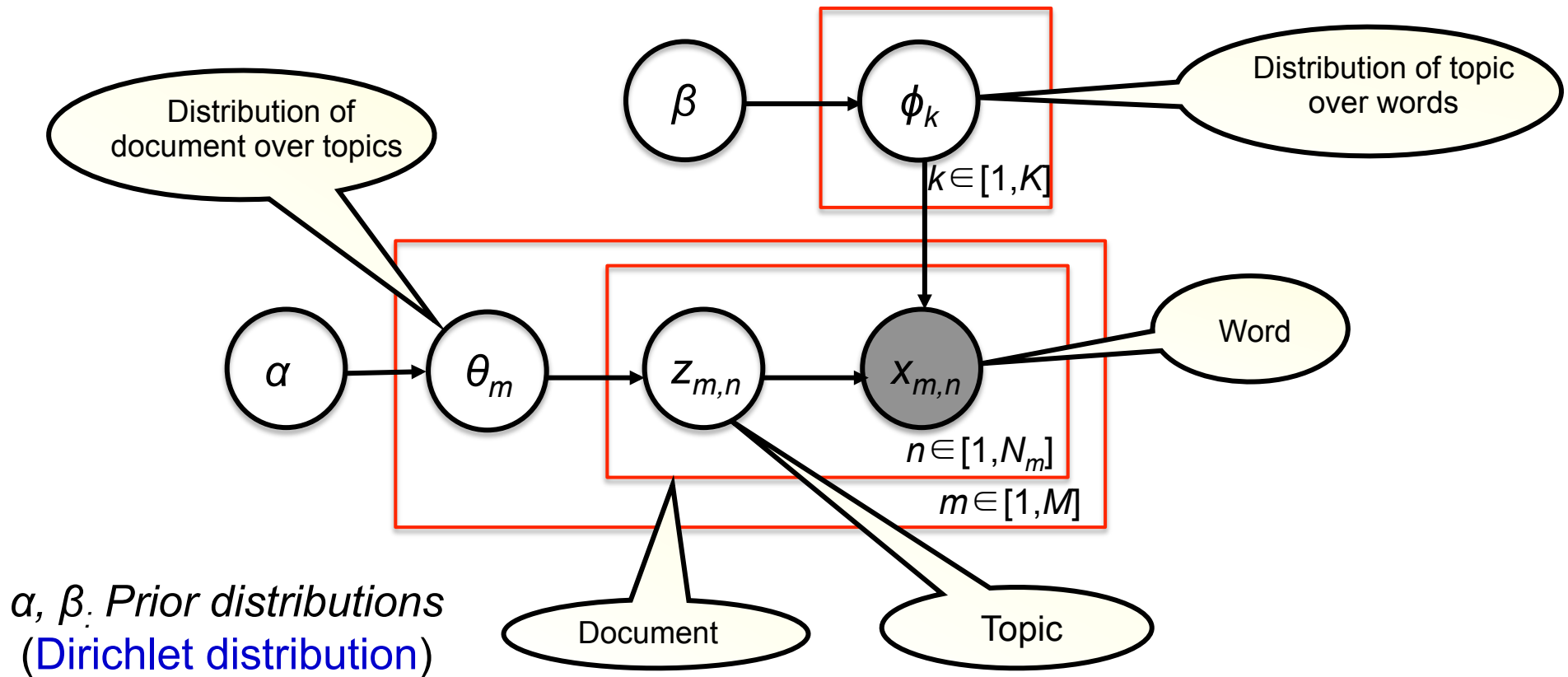
f_j denotes the j -th feature defined for two consecutive tokens x_i and x_{i+1}

Sequential Labeling Incorporating Topics



$$p(\mathbf{y} | \mathbf{x}, \theta, \lambda, \mu) = \frac{1}{Z} \exp\left(\sum_i \sum_k \lambda_k f_k(x_i, \theta_i, y_i) + \sum_i \sum_j \mu_j f_j(\mathbf{x}, \theta, y_i, y_{i+1})\right)$$

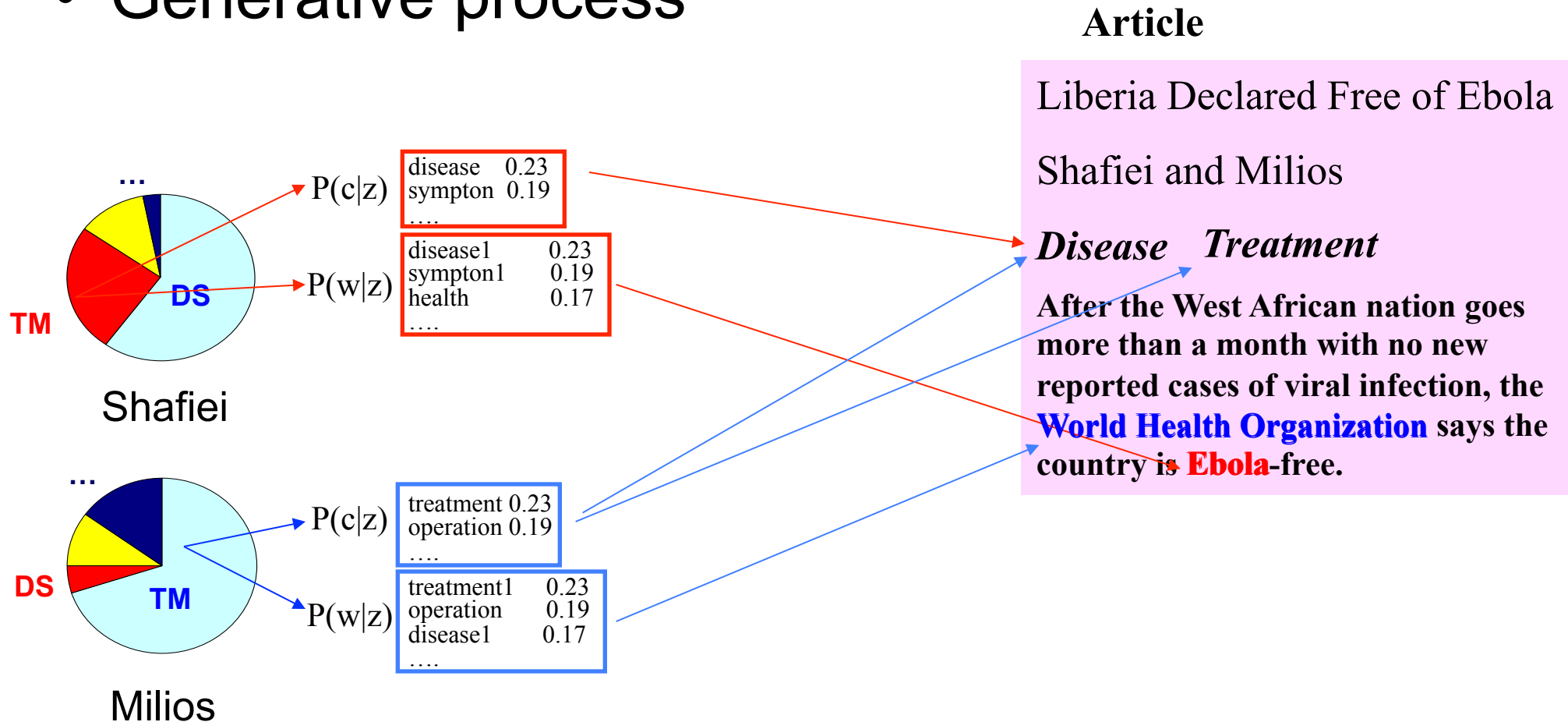
Latent Dirichlet Allocation



$$p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi} | \alpha, \beta) = \prod_{z=1}^K p(\phi_z | \beta) \prod_{d=1}^M p(\theta_d | \alpha) \prod_{i=1}^{N_d} p(x_i | \phi_z) p(z | \theta_d)$$

Extend to Model Authorship and Categories

- Generative process



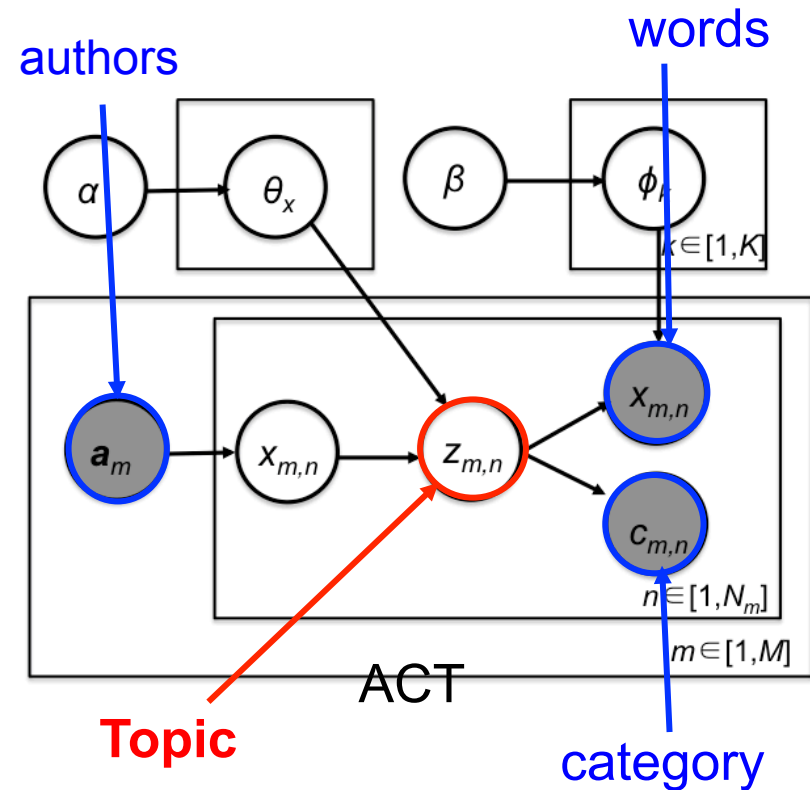
ACT Model

Generative process:

1. For each topic z , draw ϕ_z and ψ_z respectively from Dirichlet priors β_z and μ_z ;
2. For each word w_{di} in document d :

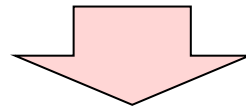
- draw an author x_{di} from \mathbf{a}_d uniformly;
- draw a topic z_{di} from a multinomial distribution $\theta_{x_{di}}$ specific to author x_{di} , where θ is generated from a Dirichlet prior α ;
- draw a word w_{di} from multinomial $\phi_{z_{di}}$;
- draw a category tag c_{di} from multinomial $\psi_{z_{di}}$.

$$P(z_{di}, x_{di} | \mathbf{z}_{-di}, \mathbf{x}_{-di}, \mathbf{w}, \mathbf{c}, \alpha, \beta, \mu) \propto \frac{m_{x_{di}z_{di}}^{-di} + \alpha_{z_{di}}}{\sum_z (m_{x_{di}z}^{-di} + \alpha_z)} \frac{n_{z_{di}w_{di}}^{-di} + \beta_{w_{di}}}{\sum_v (n_{z_{di}v}^{-di} + \beta_v)} \frac{n_{z_{di}c_d}^{-d} + \mu_{c_d}}{\sum_c (n_{z_{di}c}^{-d} + \mu_c)}$$



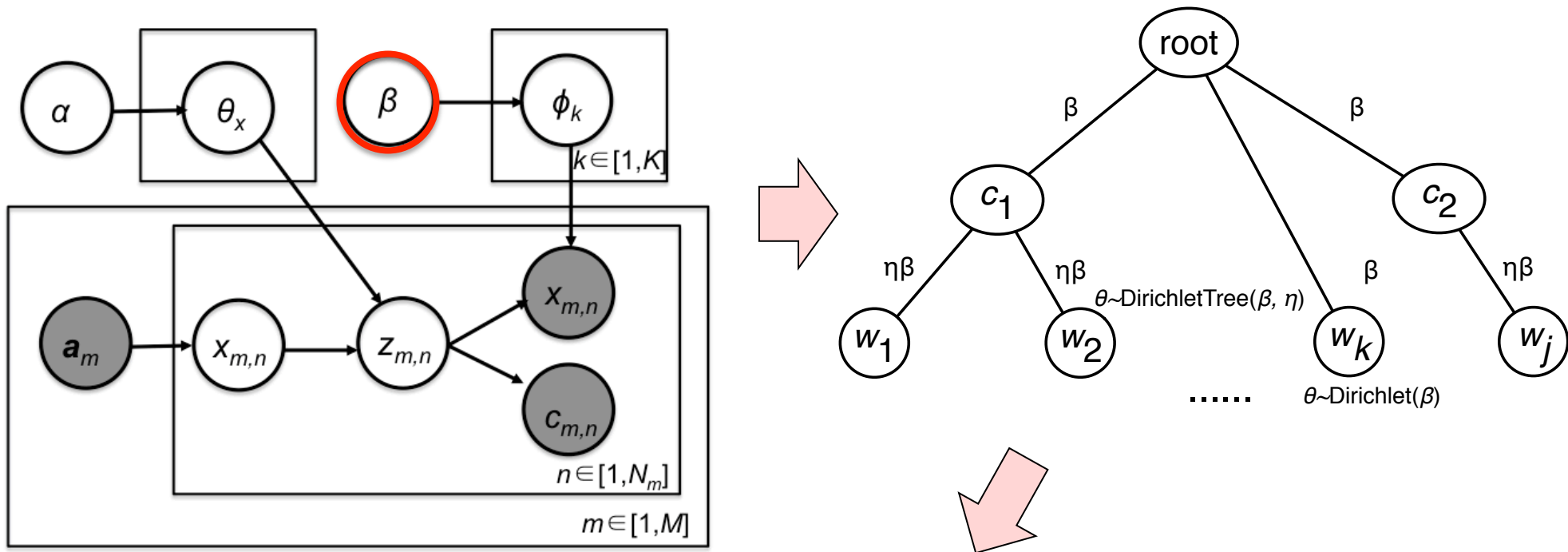
Still challenges

However, we still cannot model domain knowledge and social context!



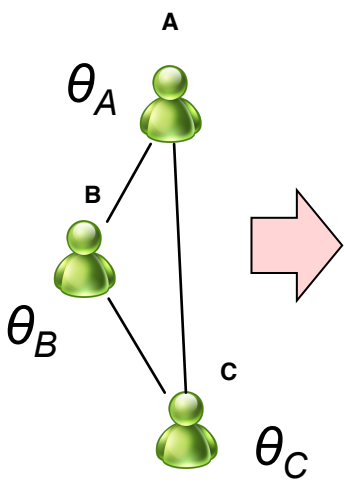
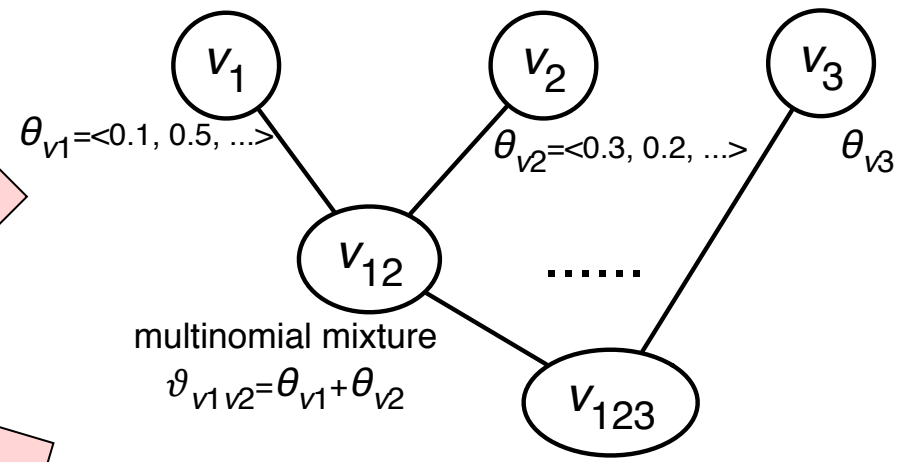
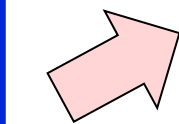
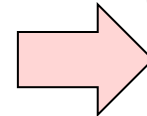
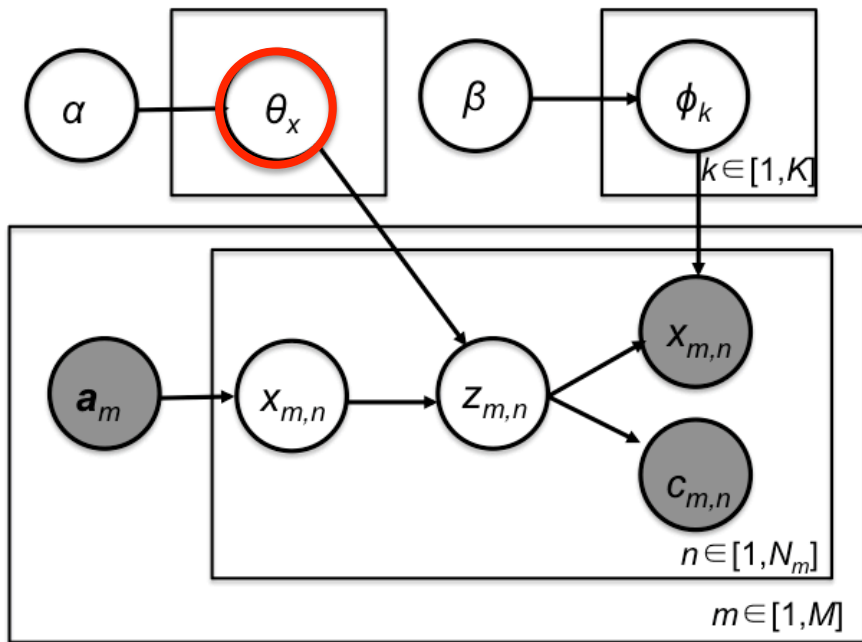
SOCINST: Modeling Domain Knowledge and Social Context Simultaneously

Modeling Domain Knowledge



$$\text{DirichletTree}(\beta, \eta) = \left(\prod_{i=1}^W \phi_{zw_i}^{\eta_{w_i}} \right) \times \left(\prod_{j, c_j \in C} \frac{\Gamma(\sum_{k, w_k \in W(c)} \eta_{w_k})}{\prod_{k, w_k \in W(c)} \Gamma(\eta_{w_k})} \left(\sum_{k, w_k \in W(c)} \phi_{zw_i}^{\eta_{w_i}} \right)^{\Delta(s)} \right)$$

Modeling Social Context



User A's Social context is defined as a mixture of topic distributions of neighbors, i.e.

$$\sum_{j \in NB(v_i)} \gamma_j \theta_j$$

Theoretical Basis

- **Aggregation property** of Dirichlet distribution

If

$$(\theta_1, \dots, \theta_i, \theta_{i+1}, \dots, \theta_K) \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_i, \alpha_{i+1}, \dots, \alpha_K)$$

then

$$(\theta_1, \dots, \theta_i + \theta_{i+1}, \dots, \theta_K) \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_i + \alpha_{i+1}, \dots, \alpha_K)$$

- **Inverse of the aggregation property**

If

$$(\theta_1, \dots, \theta_K) \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_K)$$

then

$$(\theta_1, \dots, \tau\theta_i, (1-\tau)\theta_i, \dots, \theta_K) \sim \text{Dirichlet}(\alpha_1, \dots, \tau\alpha_i, (1-\tau)\alpha_i, \dots, \alpha_K)$$

Model Learning

Input: a social network G , a document set D , a knowledge base KB ;

Output: estimated parameters θ, ϕ

For each author v , draw θ_v from Dirichlet prior α ;

For each topic z , draw ϕ_z from Dirichlet prior β ;

foreach document d **do**

if v_d does not have relationship with others **then**

foreach word $w_{di} \in w_d$ **do**

 Draw a topic $z_{di} \sim \text{multi}(\theta_v)$ from the topic model of user v ;

 Call `SamplingWord`(z_{di}, w_{di});

end

end

else if v_d have relationship with v' **then**

 Construct a multinomial mixture $\vartheta_{v_d v'}$ by combining topics distributions specific to users v_d and v' ;

foreach word $w_{di} \in w_d$ **do**

 Draw a topic $z_{di} \sim \text{multi}(\vartheta)$ from the distribution specific to the pair;

 Call `SamplingWord`(z_{di}, w_{di});

end

end

end

`SamplingWord`(z_{di}, w_{di})

if w_{di} is an instance of a concept $c \in KB$ **then**

 Draw a concept path $\{c_k\}_k \sim \text{multi}(\pi)$ from a topic-specific concept path distribution;

 Draw word $w_{di} \sim \text{multi}(\psi_c)$ from a concept-specific multinomial distribution;

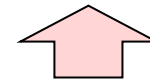
end

else

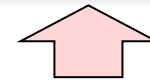
 Draw word $w_{di} \sim \text{multi}(\phi_{z_{di}})$ directly from a topic-specific multinomial distribution;

end

$$P(z_{di} | \mathbf{z}_{-di}, \mathbf{w}, \cdot) = \frac{n_{vz_{di}}^{-di} + \gamma n_{v'z_{di}} + \alpha}{\sum_z (n_{vz}^{-di} + \gamma n_{v'z}) + W\alpha} \times \prod_{k=1}^T \frac{m_{z_{di}c_{di}^k}^{-di} + W_{c_{di}^k}\beta}{\sum_{c_s} (m_{z_{di}c_s}^{-di} + W_{c_s^k}\beta)} \times \frac{m_{c_{di}w_{di}}^{-di} + \eta}{\sum_w m_{c_{di}w}^{-di} + W_c\eta}$$

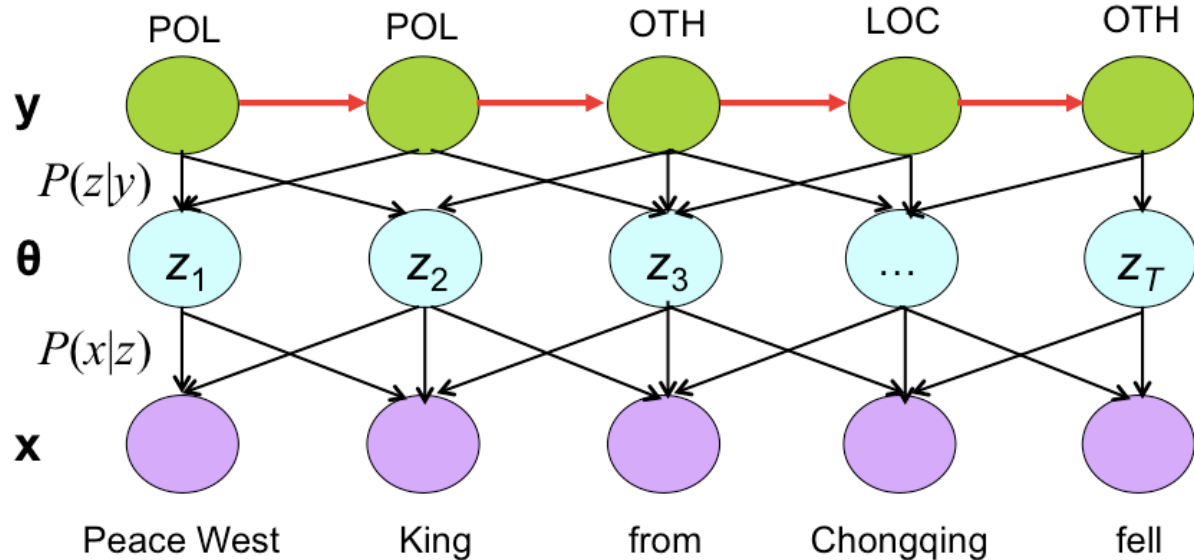
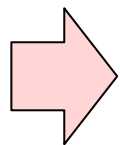
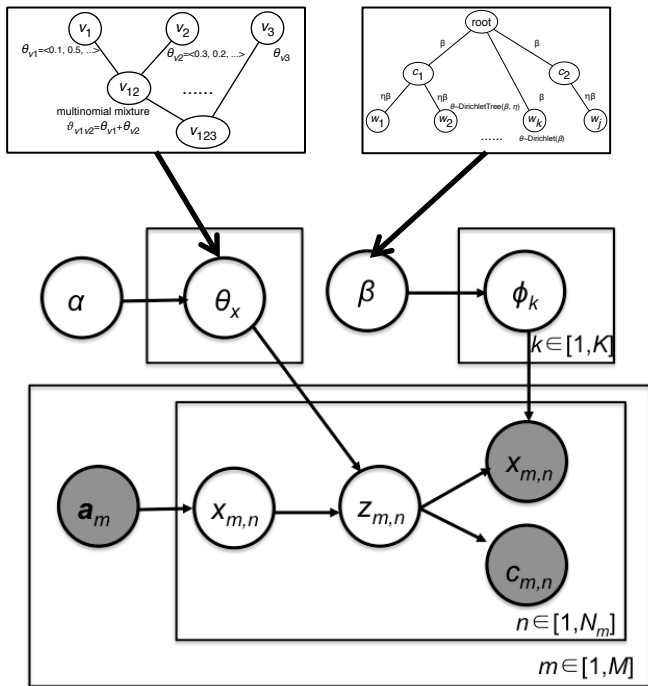


$$P(z_{di} | \mathbf{z}_{-di}, \mathbf{w}, \cdot) = \frac{n_{vz_{di}}^{-di} + \alpha}{\sum_z n_{vz}^{-di} + W\alpha} \times \prod_{k=1}^T \frac{m_{z_{di}c_{di}^k}^{-di} + W_{c_{di}^k}\beta}{\sum_{c_s} (m_{z_{di}c_s}^{-di} + W_{c_s^k}\beta)} \times \frac{m_{c_{di}w_{di}}^{-di} + \eta}{\sum_w m_{c_{di}w}^{-di} + W_c\eta}$$



$$P(z_{di} | \mathbf{z}_{-di}, \mathbf{w}, \cdot) = \frac{n_{vz_{di}}^{-di} + \alpha}{\sum_z n_{vz}^{-di} + K\alpha} \times \frac{m_{z_{di}w_{di}}^{-di} + \beta}{\sum_w m_{z_{di}w}^{-di} + W\beta}$$

Sequential Labeling Incorporating Topics



$$p(\mathbf{y} | \mathbf{x}, \theta, \lambda, \mu) = \frac{1}{Z} \exp\left(\sum_i \sum_k \lambda_k f_k(x_i, \theta_i, y_i) + \sum_i \sum_j \mu_j f_j(\mathbf{x}, \theta, y_i, y_{i+1})\right)$$

Experiments

Data Sets

- **All codes and datasets** can be **downloaded** here <http://aminer.org/socinst/>

- Dataset

Domain	#documents	#instances	#relationships
Weibo	1,800	545	10,763
I2B2	899	2,400	27,175
ICDM'12 Contest	2,110	565	NA

- Goal:
 - **Weibo**: Our goal is to extract real morph instances in the dataset.
 - **I2B2**: Our goal here is to extract private health information instances in the dataset.
 - **ICDM'12 Contest**: Our goal is to recognize product mentions in the dataset.

I2B2

HISTORY OF PRESENT ILLNESS :

Mr. **Blind** is a 79-year-old white male with a history of diabetes mellitus , inferior myocardial infarction , who underwent open repair of his increased diverticulum **November 13th** at **Sephsandpot Center**. The patient developed hematemesis **November 15th** and was intubated for respiratory distress. He was transferred to the **Valtawnprinceel Community Memorial Hospital** for endoscopy and esophagoscopy on the **16th of November** which showed a 2 cm linear tear of the esophagus at 30 to 32 cm .

Patient

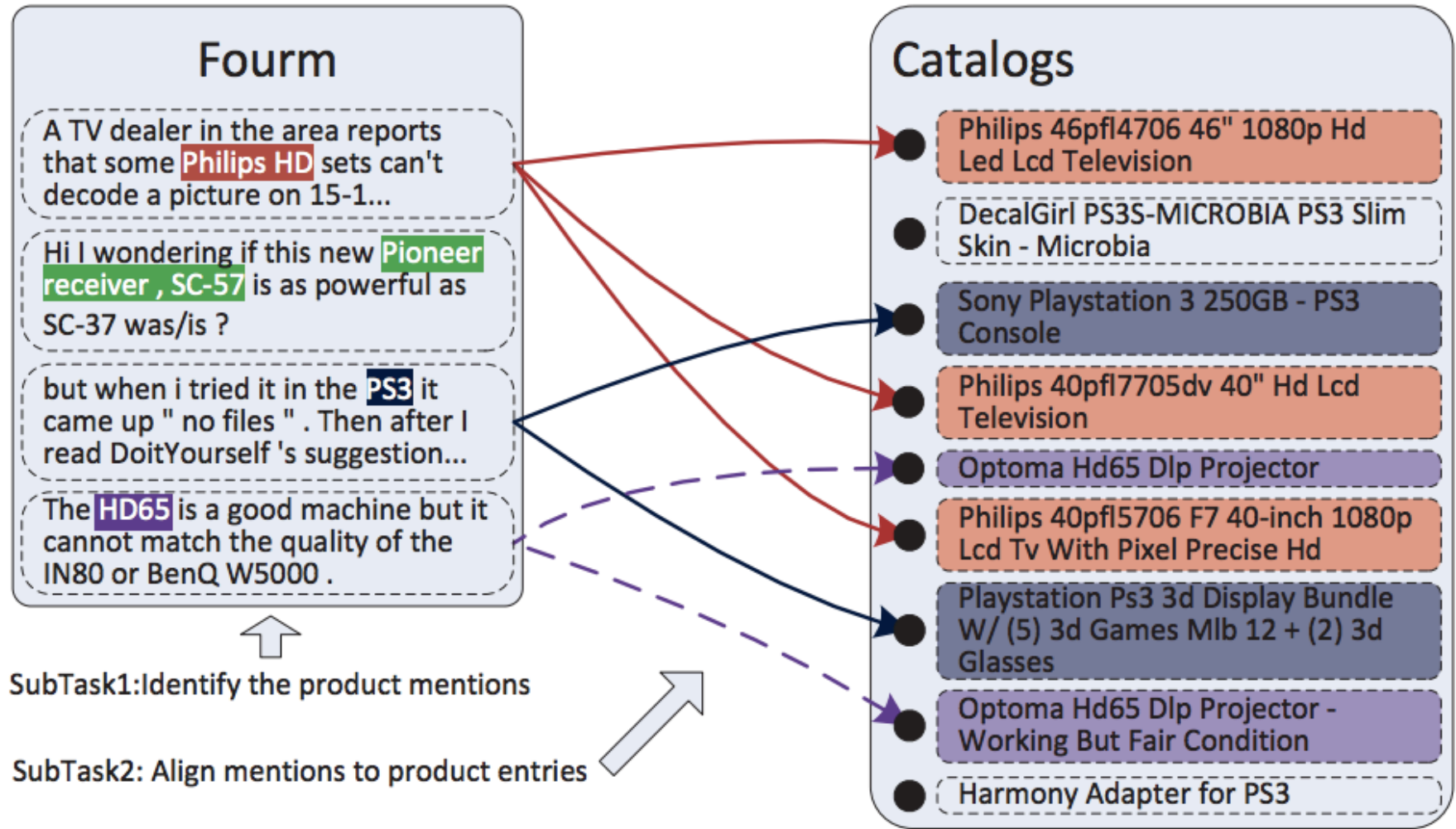
Doctor

Date

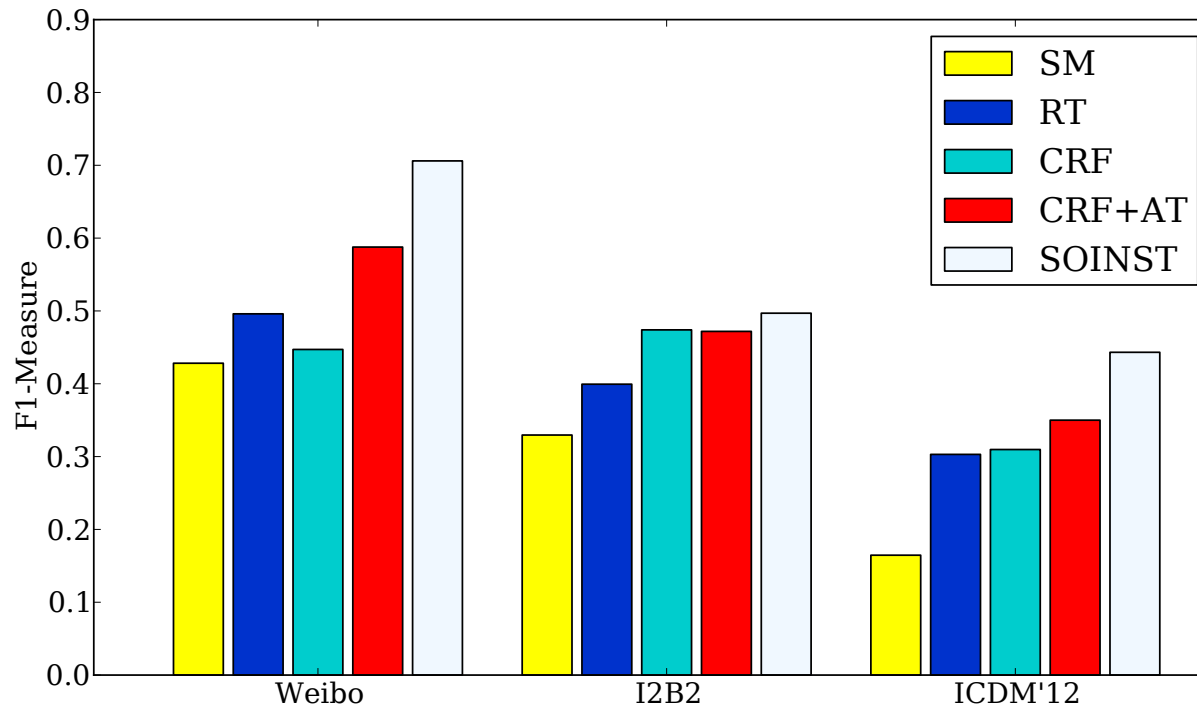
Location

Hospital

ICDM'12 Contest



Results



- **SM**: Simply extracts all the terms/symbols that are annotated
- **RT**: Recognizes target instances from the test data by a set of rule templates
- **CRF**: Trains a CRF model using features associated with each token
- **CRF+AT**: Uses Author-Topic (AT) [30] to train a model and then it use the learned topics as features for CRF for instance recognition
- **SOCINST**: Our proposed model

Results

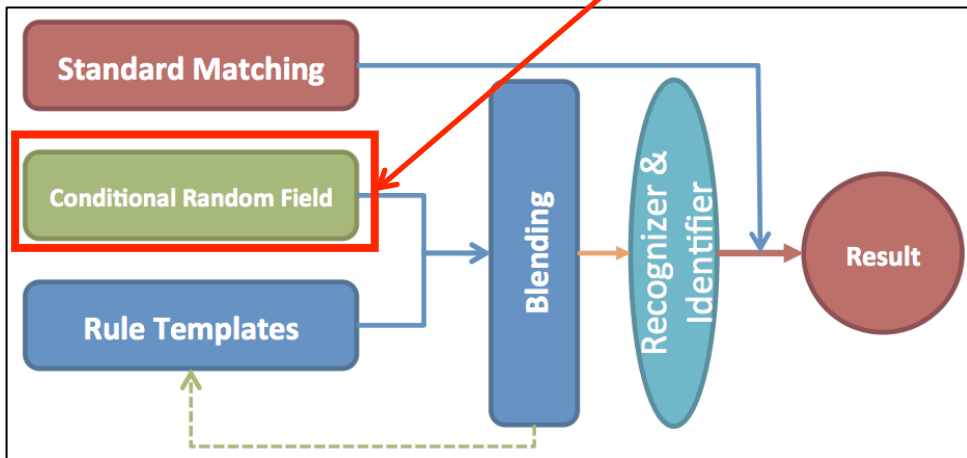
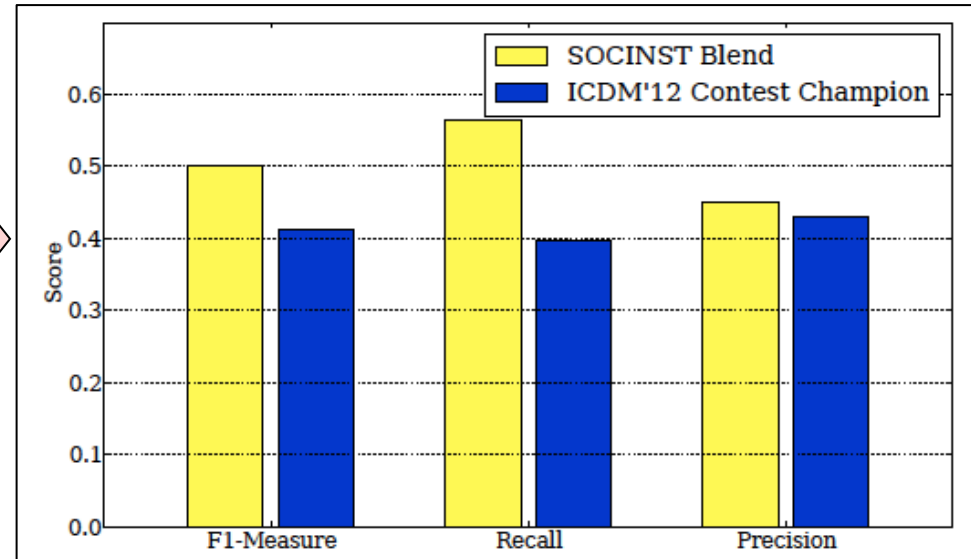
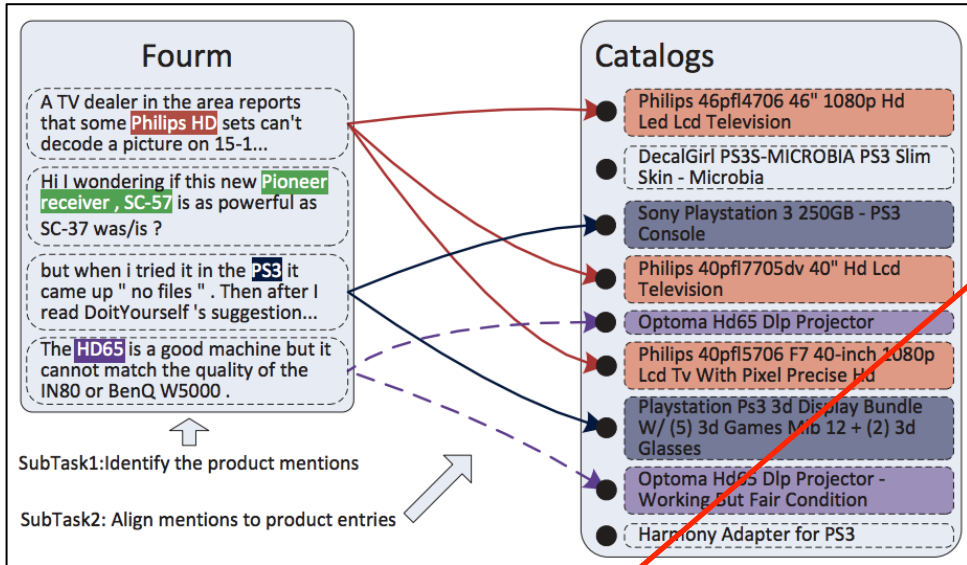
- **SM:** Simply extracts all the terms/symbols that are annotated
- **RT:** Recognizes target instances from the test data by a set of rule templates.
- **CRF:** Trains a CRF model using features associated with each token
- **CRF+AT:** Uses Author-Topic (AT) [30] to train a model and then it use the learned topics as features for CRF for instance recognition
- **SOCINST:** Our proposed model

Data	Method	Recall	Precision	F1-Measure
Weibo	SM	55.34	34.92	42.82
	RT	39.62	66.31	49.60
	CRF	29.24	94.89	44.71
	CRF+AT	43.71	89.67	58.77
	SOCINST	65.72	76.27	70.60
I2B2	SM	39.58	28.24	32.96
	RT	39.60	40.29	39.94
	CRF	40.99	56.19	47.40
	CRF+AT	41.37	54.92	47.19
	SOCINST	43.94	57.18	49.69
ICDM'12 Contest	SM	9.47	62.50	16.46
	RT	23.69	42.01	30.30
	CRF	21.80	53.48	30.97
	CRF+AT	26.54	51.37	35.00
	SOCINST	37.91	53.33	44.32

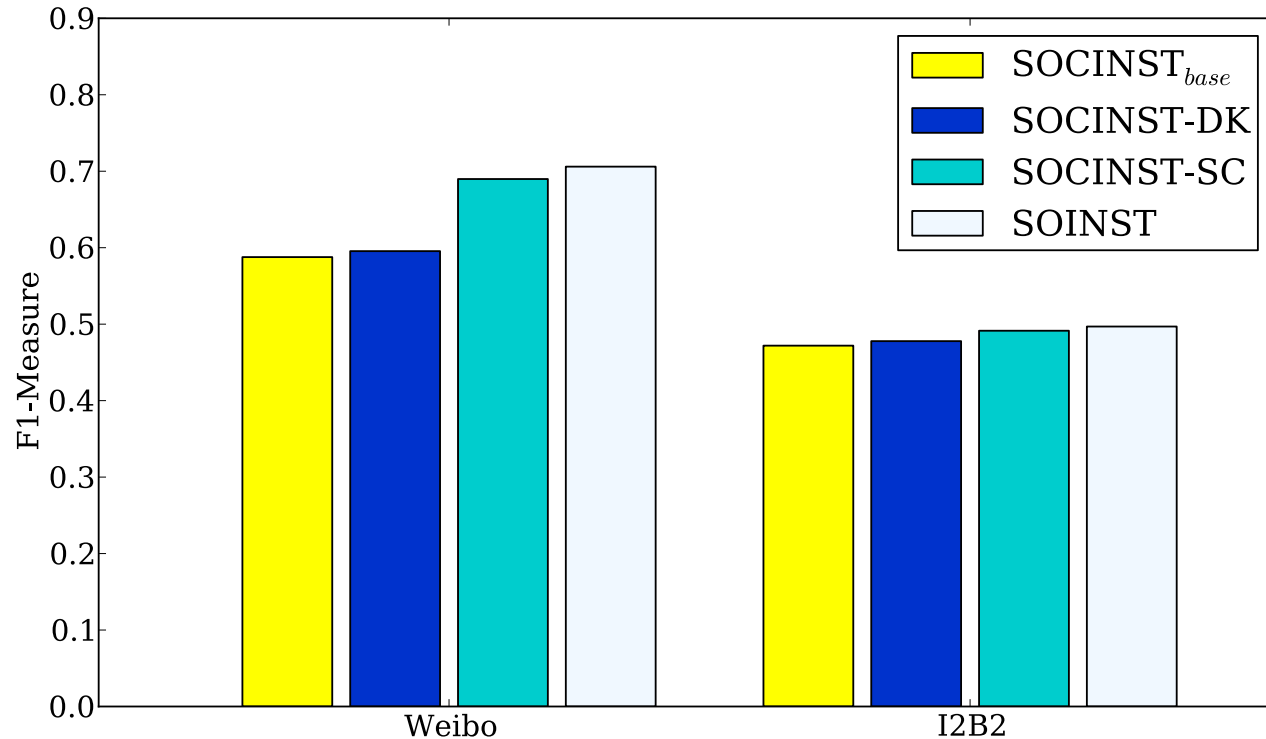
More Results—ICDM'12 Contest

Performance comparison of SOCINST and the first place [38] in ICDM'12 Contest.

By incorporating the modeling results into the CRF model



Effects of Social Context and Domain Knowledge

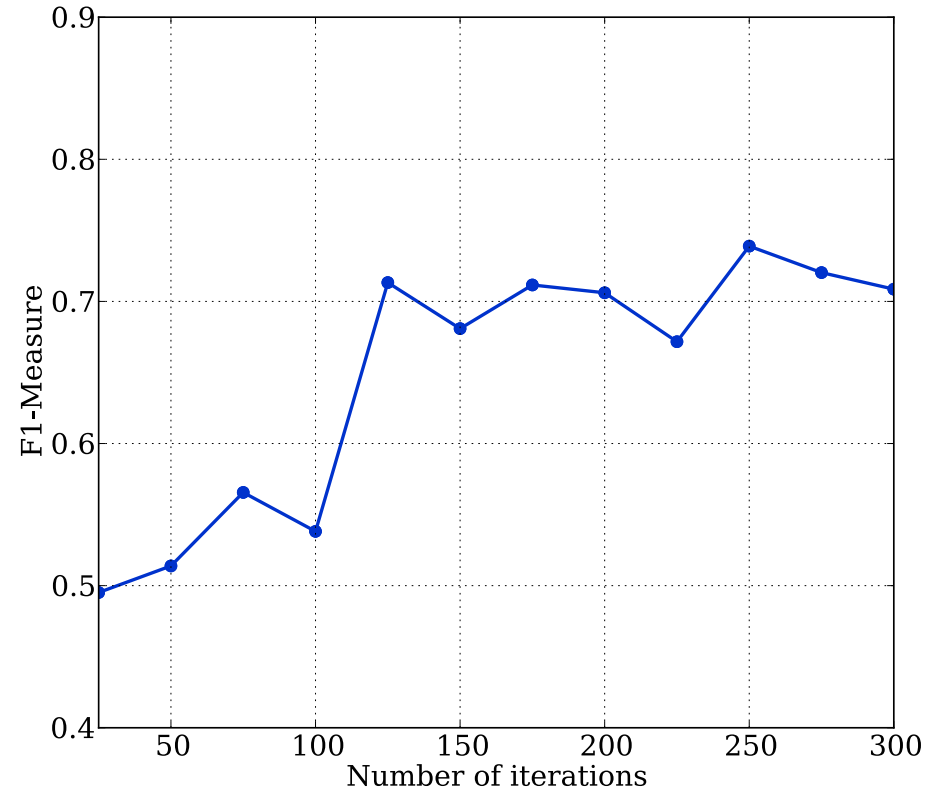
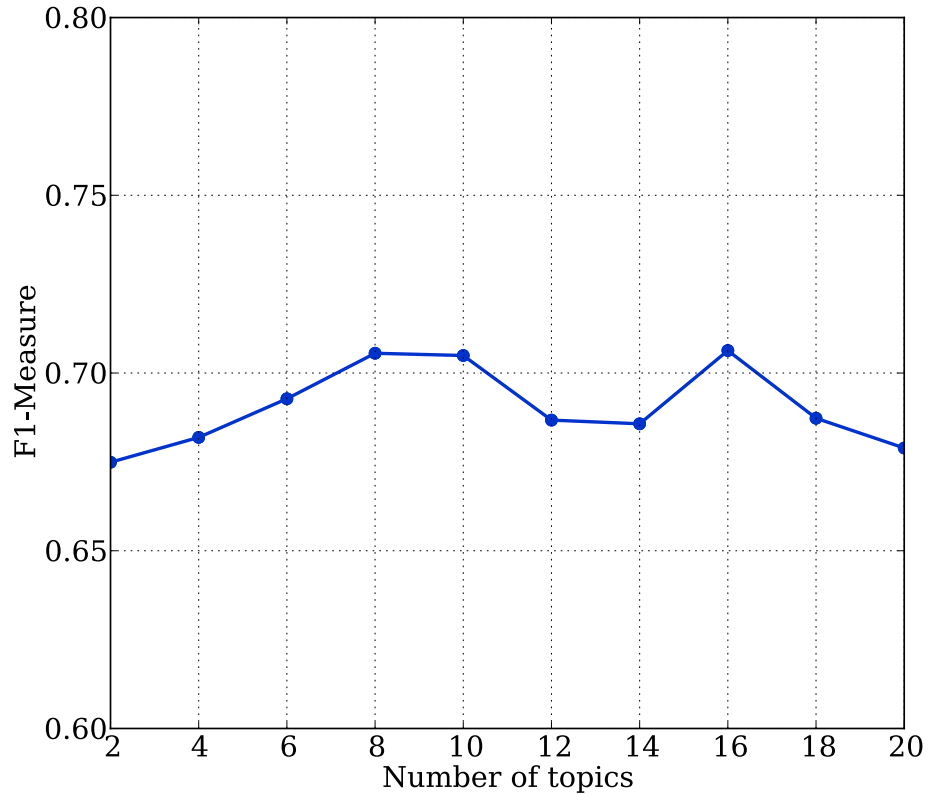


SOCINST_{base}— we removed both social context and domain knowledge from our method;

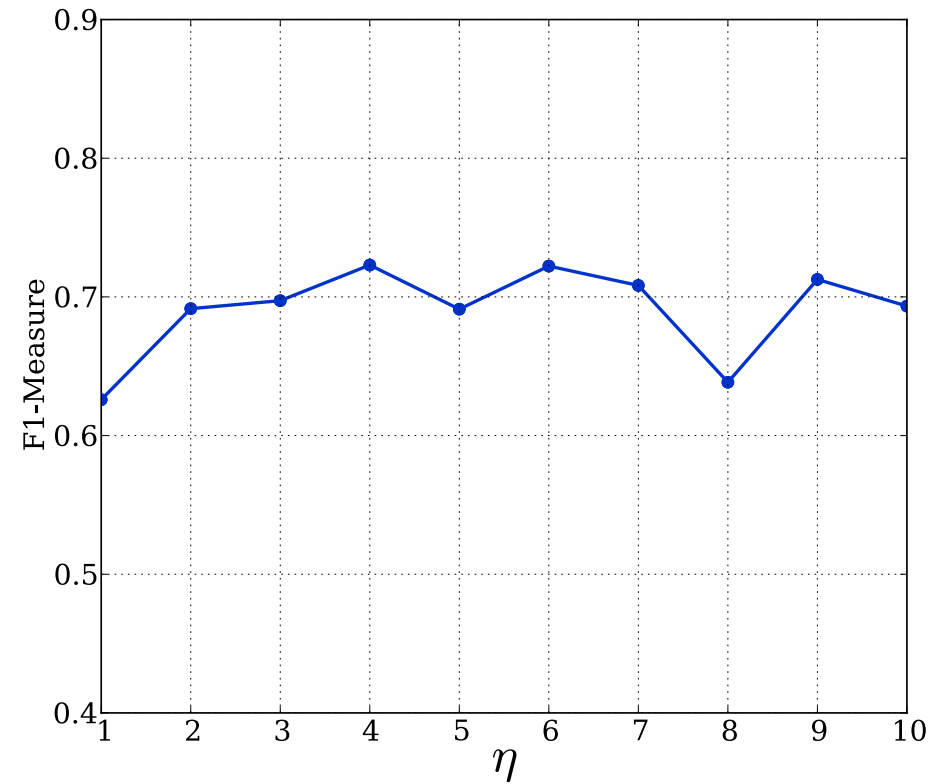
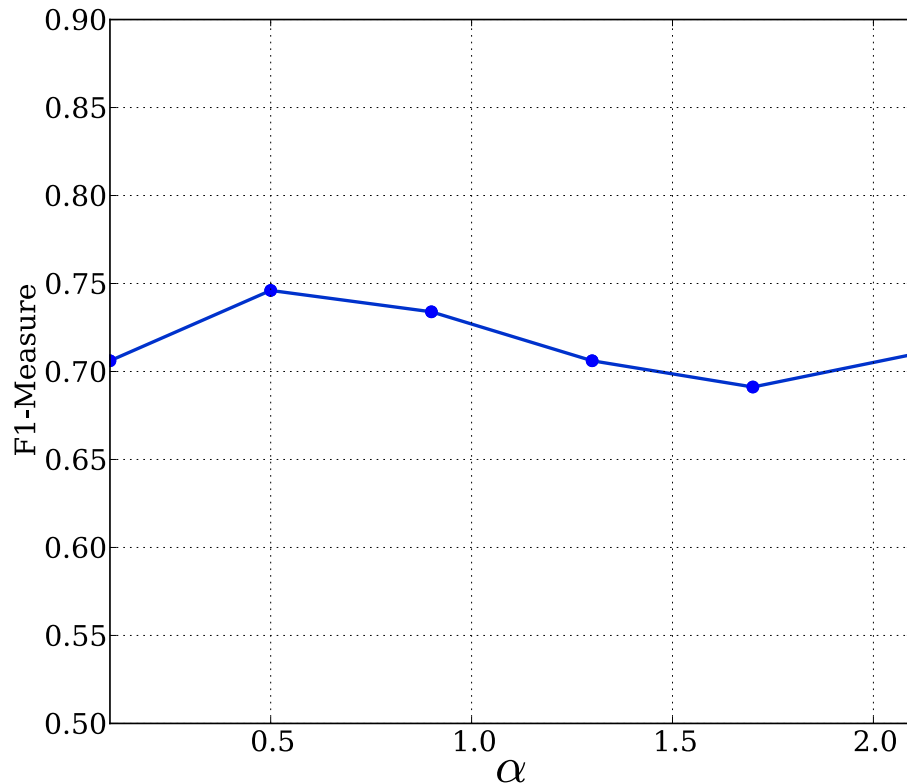
SOCINST-SC— we removed social context from our method;

SOCINST-DK— we removed domain knowledge from our method;

Parameter Analysis



Parameter Analysis (cont.)



*** All the other hyperparameters fixed**
The number of topics is set to $K = 15$

AMiner

(<http://aminer.org>)



The screenshot displays the AMiner web interface. At the top, there is a search bar with the text "Whatever comes to your mind" and a home button. The main content area shows a document page from "CMPUT690 Principles of Knowledge Discovery in Databases" by Osmar R. Zaiane, 1999. The document title is "Chapter I: Introduction to Data Mining". The text discusses the information age and the challenges of data storage and retrieval. Several phrases are highlighted in red boxes: "overwhelming", "This initial chaos has led", "database management systems (DBMS)", "text reports and military intelligence", and "decision-making". The right sidebar contains an "About" section with a "5" star rating, "Recommended tags" (data mining), and a "Your tags" section. At the bottom of the sidebar is a comment box with the text "What are you thinking?" and a "Post" button.

© Osmar R. Zaiane, 1999

CMPUT690 Principles of Knowledge Discovery in Databases

Chapter I: Introduction to Data Mining

We are in an age often referred to as the information age. In this information age, because we believe that information leads to power and success, and thanks to sophisticated technologies such as computers, satellites, etc., we have been collecting tremendous amounts of information. Initially, with the advent of computers and means for mass digital storage, we started collecting and storing all sorts of data, counting on the power of computers to help sort through this amalgam of information. Unfortunately, these massive collections of data stored on disparate structures very rapidly became **overwhelming**. **This initial chaos has led** to the creation of structured databases and **database management systems (DBMS)**. The efficient database management systems **have been very important assets** for management of a large corpus of data and especially for effective and efficient retrieval of particular information from a large collection whenever needed. The proliferation of database management systems has also contributed to recent massive gathering of all sorts of information. Today, we have far more information than we can handle: from business transactions and scientific data, to satellite pictures, **text reports and military intelligence**. Information retrieval is simply not enough anymore for **decision-making**. Confronted with huge collections of data, we have now created new needs to help us make better managerial choices. These needs are automatic summarization of data, extraction of the "essence" of information stored, and the discovery of patterns in raw data.

What kind of information are we collecting?

We have been collecting a myriad of data, from simple numerical measurements and text documents, to more complex information such as spatial data, multimedia channels, and hypertext documents. Here is a non-exclusive list of a variety of information collected in digital form in databases and in flat files.

- **Business transactions:** Every transaction in the business industry is (often) "memorized" for perpetuity. Such transactions are usually time related and can be inter-business deals such as purchases, exchanges, banking, stock, etc., or intra-business operations such as management of in-house wares and assets. Large department stores, for example, thanks to the widespread use of bar codes, store millions of transactions daily representing often terabytes of data. Storage space is not the major problem, as the price of hard disks is continuously dropping, but the effective use of the data in a reasonable time frame for competitive decision-making is definitely the most important problem to solve for businesses that struggle to survive in a highly competitive world.
- **Scientific data:** Whether in a Swiss nuclear accelerator laboratory counting particles, in the Canadian forest studying readings from a grizzly bear radio collar,



Conclusion

- Study the problem of instance recognition by incorporating social context and domain knowledge
- Propose a topic modeling approach to learn topics by considering social relationships between users and context information from a domain knowledge base
- Experimental results on three different datasets validate the effectiveness and the efficiency of the proposed method.



Future work

- The general idea of incorporating social context and domain knowledge for entity recognition represents a new research direction
- Combining the sequential labeling model and the proposed SOCINST into a unified model should be beneficial
- Further incorporating other social interactions, such as social influence, to help instance recognition is an intriguing direction

Thank you !

Collaborators:

Jimeng Sun (**Georgia Tech**)

Zhanpeng Fang (**THU**)

Jie Tang, KEG, Tsinghua U,
Download all data & Codes,

<http://keg.cs.tsinghua.edu.cn/jietang>
<http://aminer.org/socinst>

Modeling Short Text with Topics

$$p_d(x) = \lambda_B p(x|\theta_B) + (1-\lambda) \sum_{k=1}^K \pi_{d,k} p(x|\theta_k)$$

$$\log p(d) = \sum_{x \in V} n(x,d) \log [\lambda_B p(x|\theta_B) + (1-\lambda) \sum_{k=1}^K \pi_{d,k} p(x|\theta_k)]$$

