COGNITIVE COMPUTATION GROUP
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

A Discriminative Latent Variable Model for Online Clustering

Rajhans Samdani, **Kai-Wei Chang**, Dan Roth
Department of Computer Science
University of Illinois at Urbana-Champaign

Motivating Example: Coreference

- Coreference resolution: cluster denotative noun phrases (***mentions***) in a document based on underlying entities

> **[Bill Clinton]**, recently elected as the **[President of the USA]**, has been invited by the **[Russian President]**, **[Vladimir Putin]**, to visit **[Russia]**. **[President Clinton]** said that **[he]** looks forward to strengthening ties between **[USA]** and **[Russia]**.

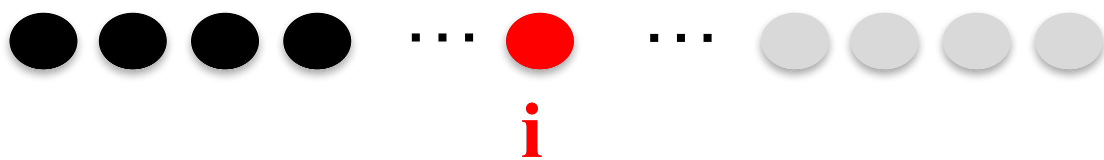
- The task: learning a clustering function from training data
  - Used expressive features between mention pairs (e.g. string similarity).
  - Learn a similarly metric between mentions.
  - Cluster mentions based on the metric.

Learning this metric using a joint distribution over clustering

- The mention arrives in a left-to-right order

Online Clustering

- Online clustering: items arrive in a given order



- **Motivating property:** cluster item **i** with no access to future items on the right, only the previous items to the left

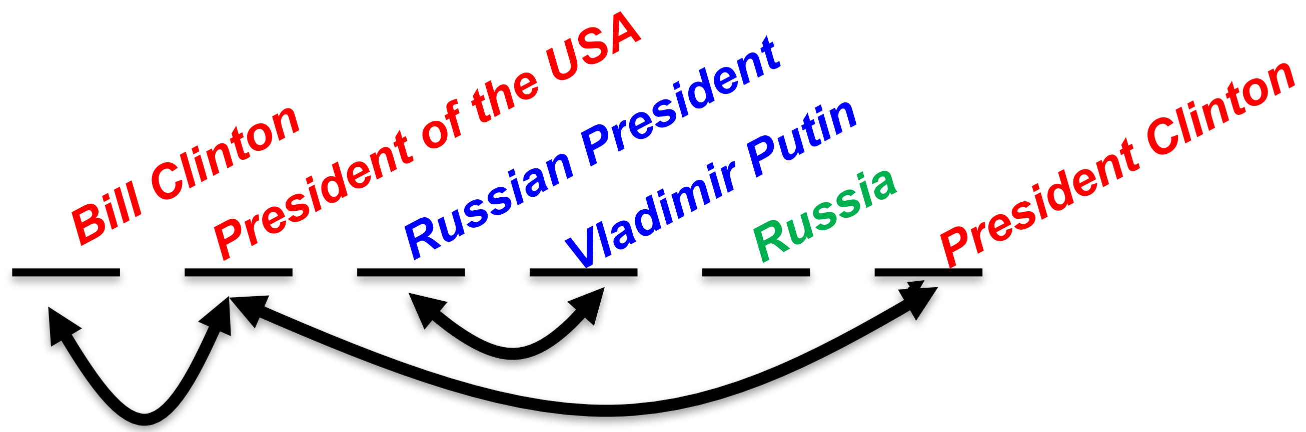
- This setting is general and is natural in many tasks.
  - E.g., cluster posts in a forum, cluster network attack

- An online clustering algorithm is likely to be more efficient than a batch algorithm under such setting.

Greedy Best-Left-Link Clustering

[Bill Clinton], recently elected as the **[President of the USA]**, has been invited by the **[Russian President]**, **[Vladimir Putin]**, to visit **[Russia]**. **[President Clinton]** said that **[he]** looks forward to strengthening ties between **[USA]** and **[Russia]**.

- Best-Left-Linking decoding: (Bengtson and Roth '08).



- A Naïve way to learn the model:
  - decouple (i) learning a similarity metric between pairs; (ii) hard clustering of mentions using this metric.

Our Contribution

- A novel discriminative latent variable model, **Latent Left-Linking Model (L^3M)**, for jointly learning metric and clustering, that outperforms existing models
- Training the pair-wise similarity metric for clustering using a latent variable structured prediction
- Relaxing the single best-link: consider a distribution over links
- Efficient learning algorithm that decomposes over individual items in the training stream

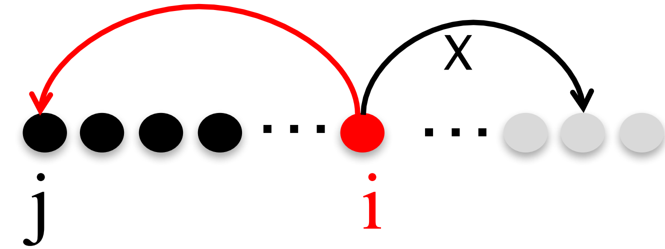
Outline

- Motivation, examples and problem description
- Latent Left-Linking Model (L^3M)
  - Likelihood computation
  - Inference
  - Role of temperature
  - Alternate latent variable perspective
- Learning
  - Discriminative structured prediction learning view
  - Stochastic gradient based decomposed learning
- Empirical study

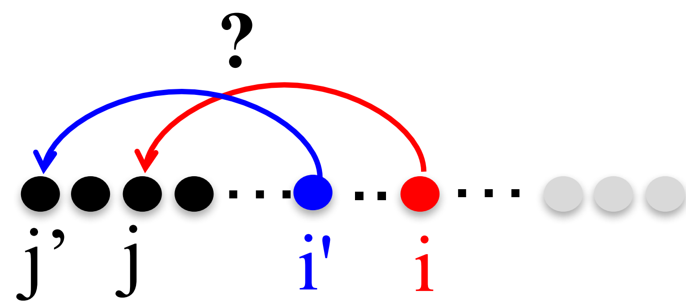
Latent Left-Linking Model (L^{3}M)

Modeling Axioms

- Each item can link only to some item on its left (creating a ***left-link***)



- Event **i** linking to **j** is **?** Of i' linking to **j'**

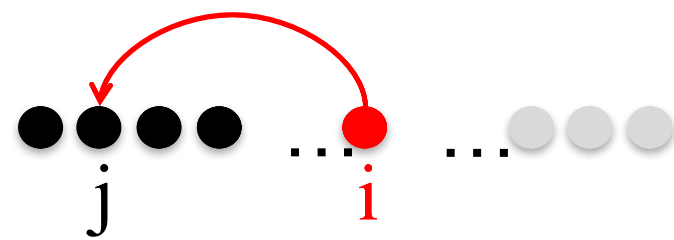


- Probability of **i** linking to **j**

Pr[j Ã i] / exp(w ¢ Á(i, j)/°)

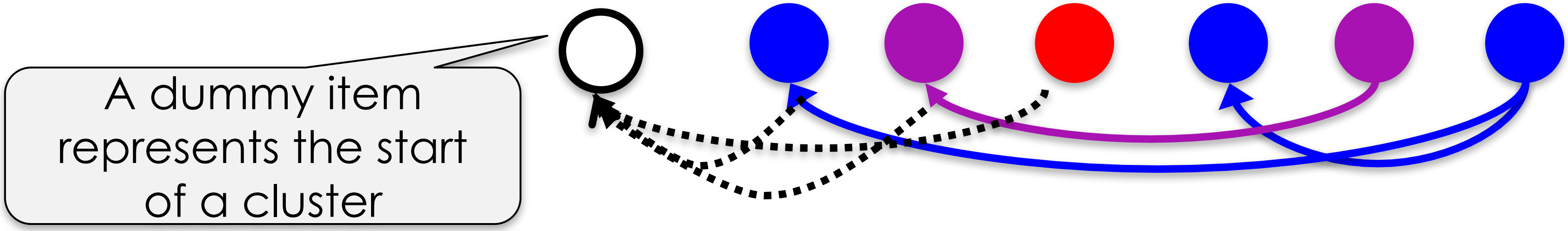
  - ° 2 [0,1] Is a temperature-like user-tuned parameter

exp (w ¢ Á(i, j)/°)



L^3M: Likelihood of Clustering

- **C** is a clustering of data stream **d**
  - $C(i, j) = 1$ *if* **i** and **j** co-clustered else 0



- Prob. of **C** : multiply prob. of items connecting as per **C**

$$\Pr[C; w] = \prod_i \Pr[i, C; w] = \prod_i \left(\sum_{j<i} \Pr[j\ \text{Ã}\ i]\ C(i, j)\right)$$

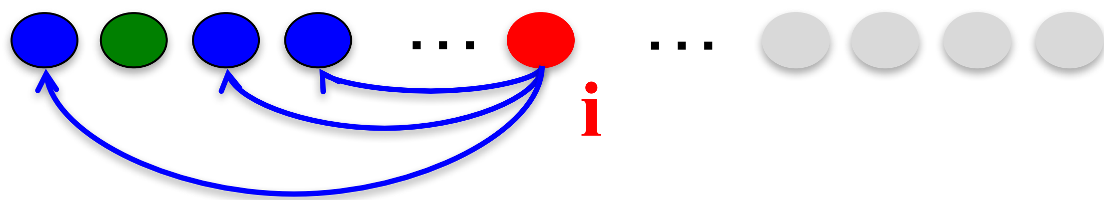
$$/ \prod_i \left(\sum_{j<i} \exp(w\ \text{¢}\ \text{Á}(i, j)\ /\text{°})\ C(i, j)\right)$$

- Partition/normalization function efficient to compute

$$Z_d(w) = \prod_i \left(\sum_{j<i} \exp(w\ \text{¢}\ \text{Á}(i, j)\ /\text{°})\right)$$

L³M: Greedy Inference/Clustering

- Sequential arrival of items:



Prob. of **i** connecting to previously formed cluster **c**

= sum of probs. of **i** connecting to items in **c**:

Pr[c ¯ i] = Σ

$$\mathbf{Pr[c\ ¯\ i] = \Sigma_{j\ 2\ c}\ Pr[j\ \tilde{A}\ i;\ w]\ /\ \Sigma_{j\ 2\ c}\ exp(w\ ¢\ \acute{A}(i,\ j)\ /°)}$$

- Greedy clustering:
  - Compute **c***= argmax$_c$ **Pr[c ¯ i]**
  - Connect **i** to **c*** if **Pr[c* ¯ i]** > **t** **(threshold)** otherwise **i** starts a new cluster
  - May not yield the most likely clustering

Inference: role of temperature °

- Prob. of **i** connecting to previous item **j**

Pr[j Ã i] / exp(w ¢ Á(i, j)/°)

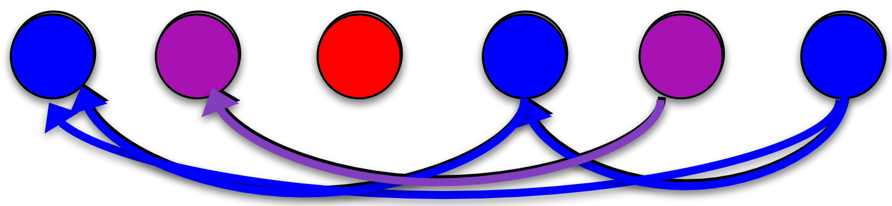
- ° tunes the importance of high-scoring links
  - As ° decreases from 1 to 0, high-scoring links become more important
  - For ° = **0**, **Pr[j Ã i]** is a Kronecker delta function centered on the argmax link (assuming no ties)

Pr[c ¯ i] / $\Sigma_{j\ 2\ c}$ exp(w ¢ Á(i, j) /°)

- For ° = **0**, clustering considers only the “best-left-link” and **greedy clustering is exact**

Latent Variables: Left-Linking Forests

- Left-linking forest, **f** : the <u>parent</u> (arrow directions reversed) of each item on its left

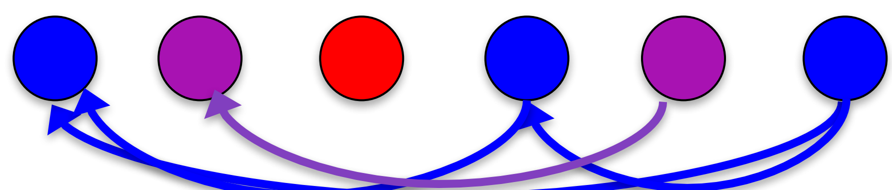


- Probability of forest **f** based on sum of edge weights in **f**

$$\mathbf{Pr[f;\ w]\ /\ exp(\Sigma_{(i,\ j)\ 2\ f}\ w\ ¢\ Á(i,\ j)\ /°)}$$

- L^{3}M: same as expressing the probability of **C** as the sum of probabilities of all consistent **(latent)** Left-linking forests

$$\mathbf{Pr[C;\ w]=\Sigma_{f2\ F(C)}\ Pr[f;\ w]}$$



COGNITIVE COMPUTATION GROUP
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Outline

- Motivation, examples and problem description
- Latent Left-Linking Model
  - Inference
  - Role of temperature
  - Likelihood computation
  - Alternate latent variable perspective
- Learning
  - Discriminative structured prediction learning view
  - Stochastic gradient based decomposed learning
- Empirical study

L^3M: Likelihood-based Learning

- Learn **w** from annotated clustering $\mathbf{C_d}$ for data d 2 D

- L^3M: Learn **w** via regularized neg. log-likelihood

$$\mathbf{LL(w) = ¯\ kwk^2} \quad \text{Regularization}$$

$$\mathbf{+ \textstyle\sum_d \ \log Z_d(w)} \qquad \text{Partition Function}$$

$$\mathbf{- \textstyle\sum_d \sum_i \log (\sum_{j<i} \exp(w\ ¢\ Á(i, j)\ /°)\ C_d\ (i, j))}$$

Un-normalized Probability

- Relation to other latent variable models:
  - Learn by marginalizing underlying latent left-linking forests
  - **°=1**: Hidden Variable CRFs (Quattoni et al, 07)
  - **°=0**: Latent Structural SVMs (Yu and Joachims, 09)

Training Algorithms: Discussion

- The objective function $\mathbf{LL(w)}$ is non-convex

- Can use **Concave-Convex Procedure (CCCP)** (Yuille and Rangarajan 03; Yu and Joachims, 09)
  - Pros: guaranteed to converge to a local minima (Sriperumbudur et al, 09)
  - Cons: requires entire data stream to compute single gradient update

- Online updates based on Stochastic (sub-)gradient descent (SGD)
  - Sub-gradient can be decomposed to a per-item basis
  - Cons: no theoretical guarantees for SGD with non-convex functions
  - Pros: can learn in an online fashion; Converge much faster than CCCP
  - Great empirical performance

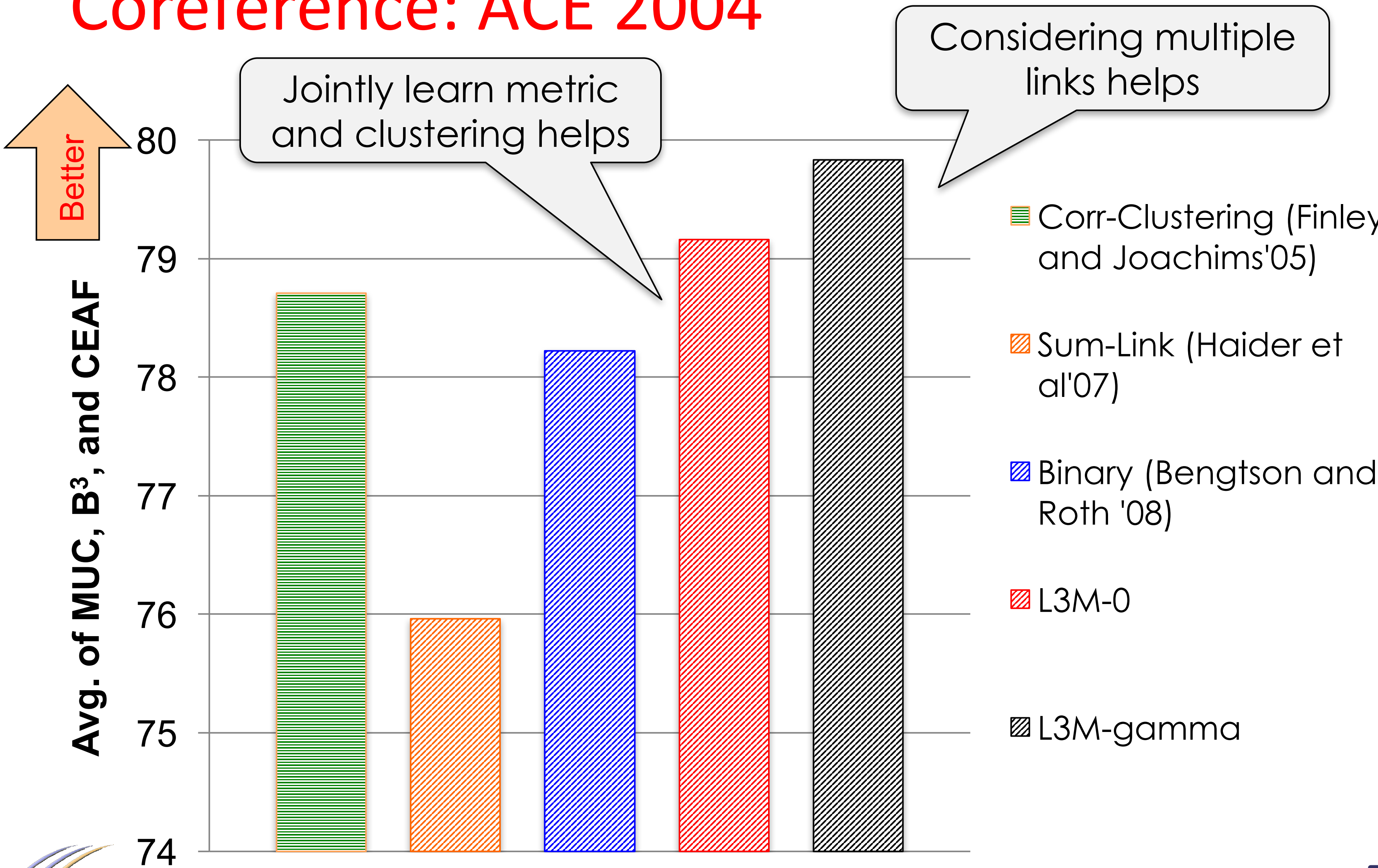
Outline

- Motivation, examples and problem description
- Latent Left-Linking Model
  - Inference
  - Role of temperature
  - Likelihood computation
  - Alternate latent variable perspective
- Learning
  - Discriminative structured prediction learning view
  - Stochastic gradient based decomposed learning
- Empirical study

Experiment: Coreference Resolution

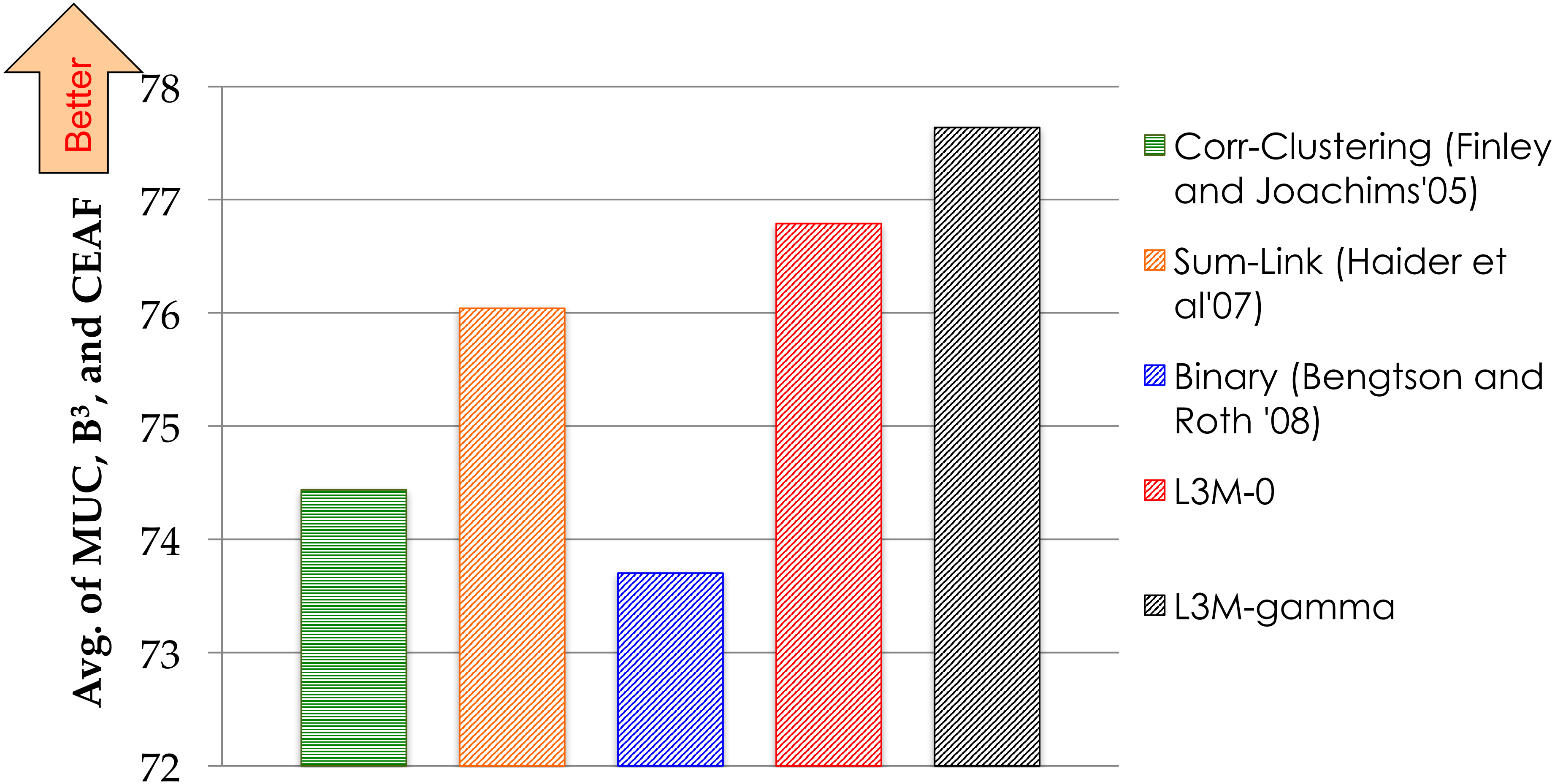
- Cluster denotative noun phrases called *mentions*
- Mentions follow a left-to-right order
- Features: mention distance, substring match, gender match, etc.
- Experiments on **ACE 2004** and **OntoNotes-5.0.**
- Report average of three popular coreference clustering evaluation metrics: MUC, B^3, and CEAF

Coreference: ACE 2004



Cognitive Computation Group
University of Illinois at Urbana-Champaign

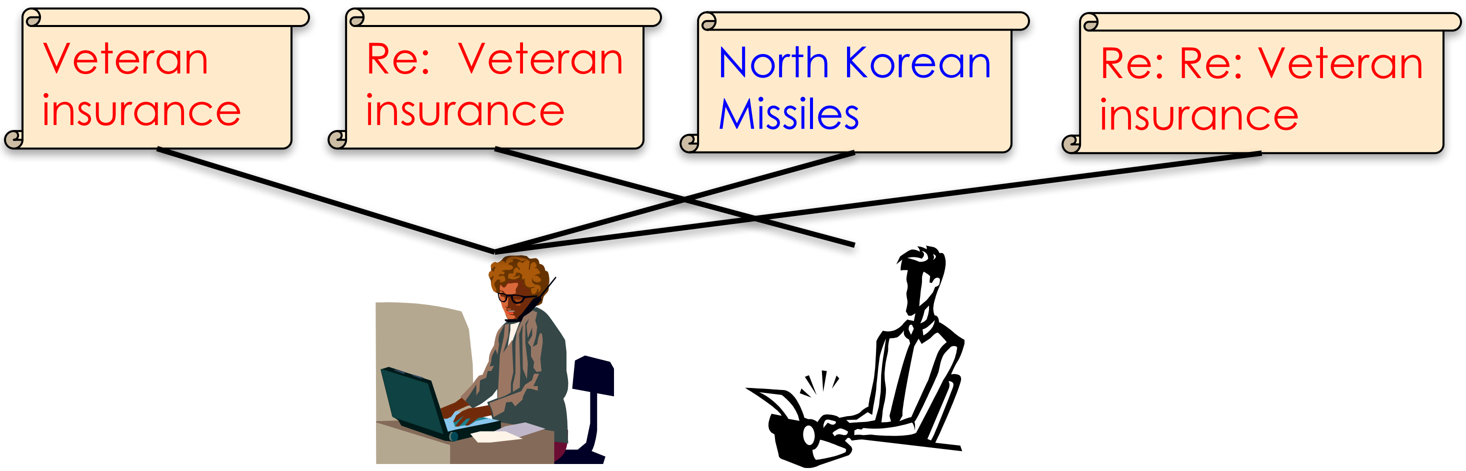
Coreference: OntoNotes-5.0



By incorporating with domain knowledge constraints, **L^{3}M** achieves the state of the art performance on OntoNotes-5.0 (Chang et al. 13)

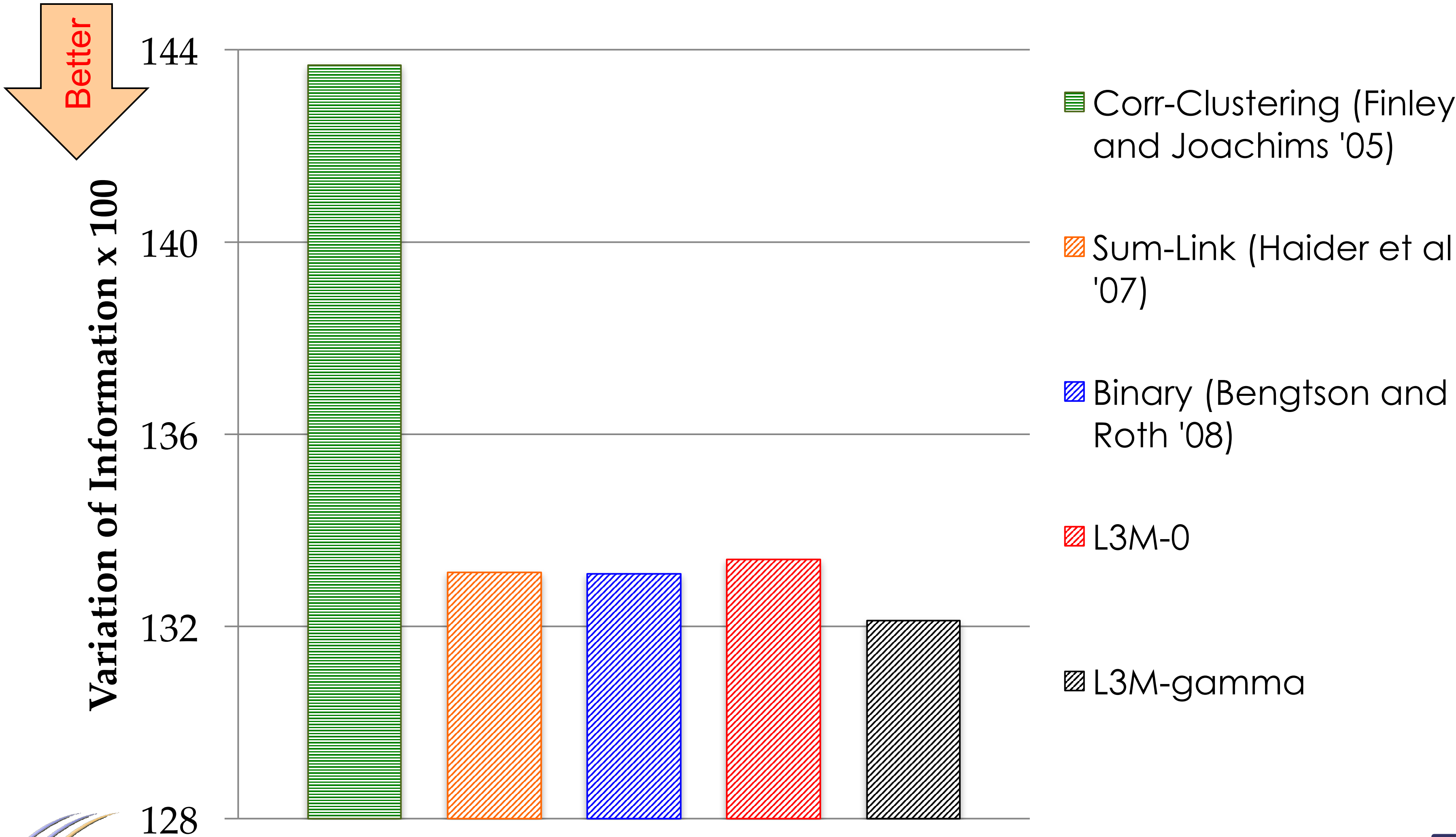
Experiments: Document Clustering

- Cluster the posts in a forum based on authors or topics.
- Dataset: discussions from www.militaryforum.com
- The posts in the forum arrive in a time order:

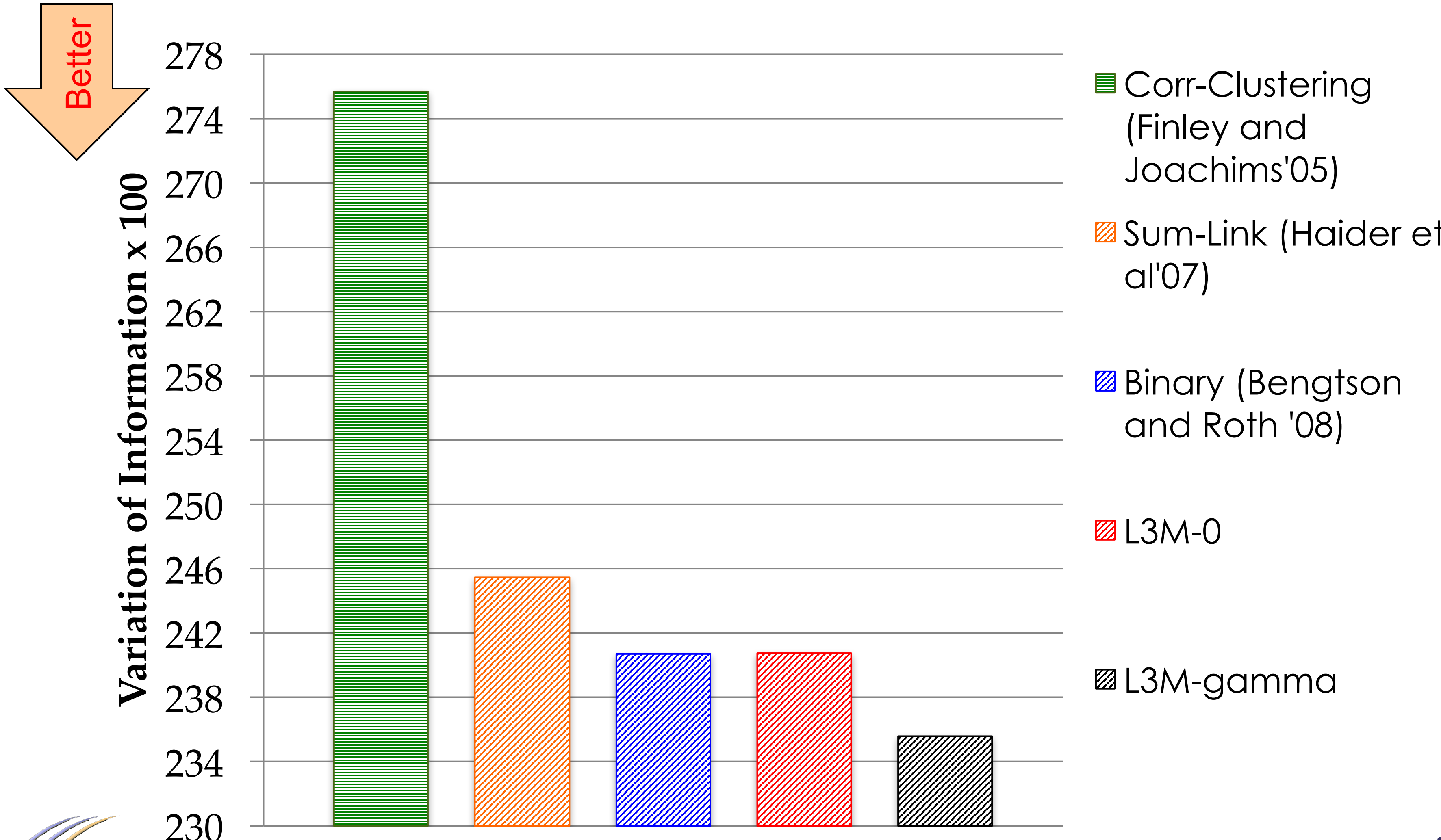


- Features: common words, tf-idf similarity, time between arrival
- Evaluate with Variation-of-Information (Meila, 07)

Author Based Clustering



Topic Based Clustering



Conclusions

- Latent Left-Linking Model
  - Principled probabilistic modeling for online clustering tasks
  - Marginalizes underlying latent link structures
  - Tuning ° helps – considering multiple links helps
  - Efficient greedy inference
- SGD-based learning
  - Decompose learning into smaller gradient updates over individual items
  - Rapid convergence and high accuracy
- Solid empirical performance on problems with a natural streaming order