# Large-scale Cross-Document Coreference Using Distributed Inference and Hierarchical Models

**Sameer Singh**[1] Amarnag Subramanya[2]
Fernando Pereira[2] Andrew McCallum[1]

[1]University of Massachusetts, Amherst MA

[2]Google Research, Mountain View CA

**Association for Computational Linguistics: Human Language Technologies**
*June 21, 2011*

Contributions:

- Cross-doc coreference on large datasets in a scalable way
- Perform distributed inference using MapReduce

Contributions:

- Cross-doc coreference on large datasets in a scalable way
- Perform distributed inference using MapReduce

1.5 **million mentions,** 500 **machines,** 38% **error reduction**

# Outline

# Coreference Problem

...60's and early 70's, **Kevin Smith** worked with...

...hip-hop is attributed to **Lovebug Starski**. What does it...

..filmmaker **Kevin Smith** returns to the role of Silent Bob...

...more irrelevant to **Kevin Smith**'s audacious "Dogma" than...

...the Lions drafted **Kevin Smith**, even though Smith was badly...

...backfield in the wake of **Kevin Smith**'s knee injury, and the addition...

...were coming," said Dallas cornerback **Kevin Smith**. "We just...

# Coreference Problem

# Undirected Graphical Model

The random variables are entities ($E$) and mentions ($M$)

# Undirected Graphical Model

The random variables are entities ($E$) and mentions ($M$)
For any assignment to entities ($E = \mathbf{e}$), we define the model score:

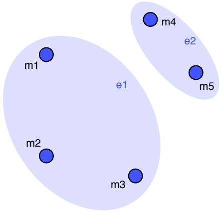
$$p(\mathbf{e}) \propto \exp \sum_{e \in \mathbf{e}} \left\{ \underbrace{\boxed{\sum_{m,n \in e} \psi_a^{mn}}}_{\text{affinity}} + \underbrace{\boxed{\sum_{m \in e, n \notin e} \psi_r^{mn}}}_{\text{repulsion}} \right\}$$

# Undirected Graphical Model

The random variables are entities ($E$) and mentions ($M$)
For any assignment to entities ($E = \mathbf{e}$), we define the model score:

$$p(\mathbf{e}) \propto \exp \sum_{e \in \mathbf{e}} \left\{ \underbrace{\boxed{\sum_{m,n \in e} \psi_a^{mn}}}_{\text{affinity}} + \underbrace{\boxed{\sum_{m \in e, n \notin e} \psi_r^{mn}}}_{\text{repulsion}} \right\}$$

For the following configuration,

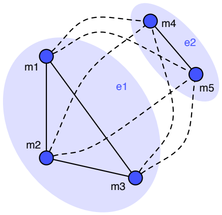


# Undirected Graphical Model

The random variables are entities ($E$) and mentions ($M$)
For any assignment to entities ($E = \mathbf{e}$), we define the model score:

$$p(\mathbf{e}) \propto \exp \sum_{e \in \mathbf{e}} \left\{ \underbrace{\boxed{\sum_{m,n \in e} \psi_a^{mn}}}_{\text{affinity}} + \underbrace{\boxed{\sum_{m \in e, n \notin e} \psi_r^{mn}}}_{\text{repulsion}} \right\}$$

For the following configuration,




$$\begin{aligned} p(e_1, e_2) \propto \exp \quad \{ & \quad \psi_a^{12} + \psi_a^{13} + \psi_a^{23} + \psi_a^{45} \\ + & \quad \psi_r^{15} + \psi_r^{25} + \psi_r^{35} \\ + & \quad \psi_r^{14} + \psi_r^{24} + \psi_r^{34} \} \end{aligned}$$

# Maximum a posteriori (MAP) Inference

We want to find the best configuration according to the model,

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} p(\mathbf{e})$$

# Maximum a posteriori (MAP) Inference

We want to find the best configuration according to the model,

$$\begin{aligned}\hat{\mathbf{e}} &= \arg\max_{\mathbf{e}} \; p(\mathbf{e}) \\ &= \arg\max_{\mathbf{e}} \; \sum_{e \in \mathbf{e}} \Big\{ \sum_{m,n \in e} \psi_a^{mn} \\ &\qquad\qquad + \sum_{m \in e, n \notin e} \psi_r^{mn} \Big\}\end{aligned}$$

# Maximum a posteriori (MAP) Inference

We want to find the best configuration according to the model,

$$
\begin{aligned}
\hat{\mathbf{e}} &= \arg\max_{\mathbf{e}} \; p(\mathbf{e}) \\
&= \arg\max_{\mathbf{e}} \; \sum_{e \in \mathbf{e}} \Big\{ \sum_{m,n \in e} \psi_a^{mn} \\
&\qquad\qquad + \sum_{m \in e, n \notin e} \psi_r^{mn} \Big\}
\end{aligned}
$$

Computational bottlenecks:

1. Space over all **e** is Bell Number($n$) in number of mentions
2. Evaluating model score for each $E = \mathbf{e}$ is $O(n^2)$

# MCMC for MAP Inference

Use MCMC sampling to perform MAP Inference

# MCMC for MAP Inference

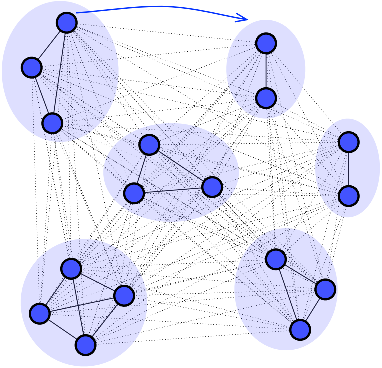
Use MCMC sampling to perform MAP Inference

1. `Initial configuration:` $\mathbf{e} \leftarrow \mathbf{e}_0$

# MCMC for MAP Inference

Use MCMC sampling to perform MAP Inference

1. Initial configuration: $\mathbf{e} \leftarrow \mathbf{e}_0$
2. Proposal Function: propose change to $\mathbf{e}$ to get $\mathbf{e}'$ (*e.g.* move mention $l$ from $e_s$ to $e_t$)



# MCMC for MAP Inference

Use MCMC sampling to perform MAP Inference

1. `Initial configuration:` $\mathbf{e} \leftarrow \mathbf{e}_0$
2. `Proposal Function:` `propose change to` $\mathbf{e}$ `to get` $\mathbf{e}'$
   (*e.g.* move mention $l$ from $e_s$ to $e_t$)
3. `Acceptance probability:` $\alpha(\mathbf{e}, \mathbf{e}') = \min\left(1, \frac{p(\mathbf{e}')}{p(\mathbf{e})}\right)$

# MCMC for MAP Inference

Use MCMC sampling to perform MAP Inference

1. Initial configuration: $\mathbf{e} \leftarrow \mathbf{e}_0$
2. Proposal Function: propose change to $\mathbf{e}$ to get $\mathbf{e}'$ (*e.g.* move mention $l$ from $e_s$ to $e_t$)
3. Acceptance probability: $\alpha(\mathbf{e}, \mathbf{e}') = \min\left(1, \frac{p(\mathbf{e}')}{p(\mathbf{e})}\right)$

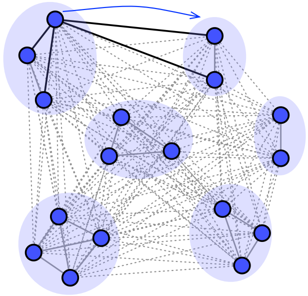
$$
\begin{aligned}
\log \frac{p(\mathbf{e}')}{p(\mathbf{e})} = & \sum_{m \in e_t} \psi_a^{lm} + \sum_{n \in e_s} \psi_r^{ln} \\
& - \sum_{n \in e_s} \psi_a^{ln} - \sum_{m \in e_t} \psi_r^{lm}
\end{aligned}
$$

# MCMC for MAP Inference

Use MCMC sampling to perform MAP Inference

1. Initial configuration: $\mathbf{e} \leftarrow \mathbf{e}_0$
2. Proposal Function: propose change to $\mathbf{e}$ to get $\mathbf{e}'$ (*e.g.* move mention $l$ from $e_s$ to $e_t$)
3. Acceptance probability: $\alpha(\mathbf{e}, \mathbf{e}') = \min\left(1, \frac{p(\mathbf{e}')}{p(\mathbf{e})}\right)$

$$\log \frac{p(\mathbf{e}')}{p(\mathbf{e})} = \sum_{m \in e_t} \psi_a^{lm} + \sum_{n \in e_s} \psi_r^{ln} - \sum_{n \in e_s} \psi_a^{ln} - \sum_{m \in e_t} \psi_r^{lm}$$



# MCMC for MAP Inference

### Advantages

- Only a small part of the model is examined for each sample
- Efficient, and scales well with model complexity

# MCMC for MAP Inference

### Advantages

- Only a small part of the model is examined for each sample
- Efficient, and scales well with model complexity

### Disadvantages

- Proportion of *good* proposals is small

# MCMC for MAP Inference

### Advantages

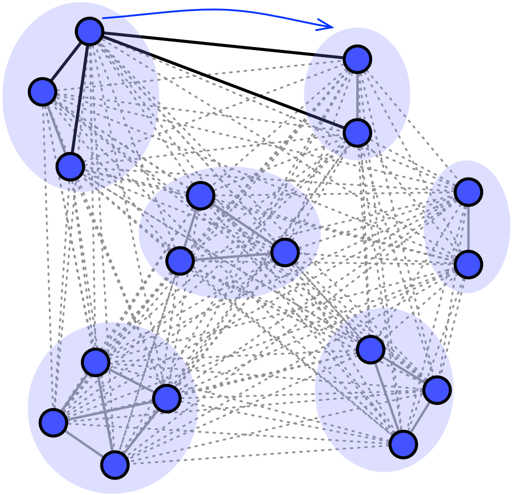
- Only a small part of the model is examined for each sample
- Efficient, and scales well with model complexity

### Disadvantages

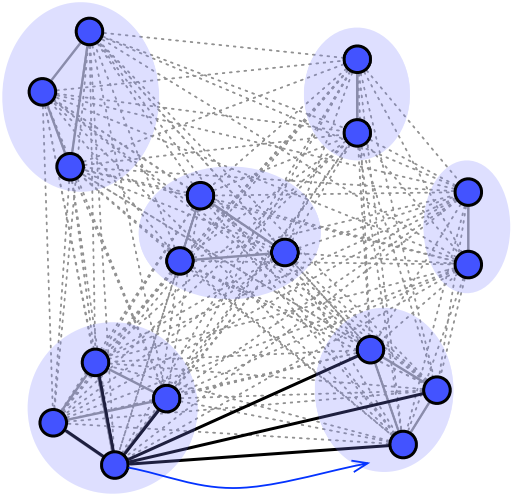
- Proportion of *good* proposals is small
- Can take a very large number of samples to converge

# Outline

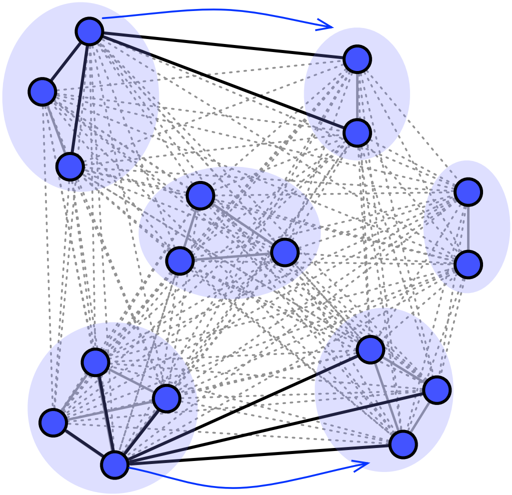
# Mutually Exclusive Proposals



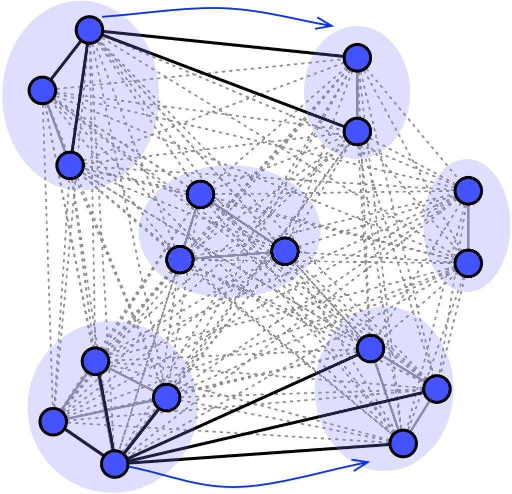
# Mutually Exclusive Proposals



# Mutually Exclusive Proposals

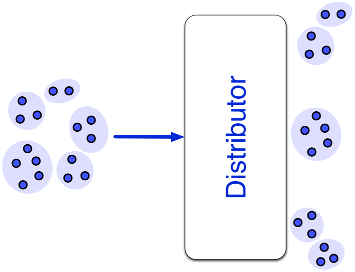


# Mutually Exclusive Proposals

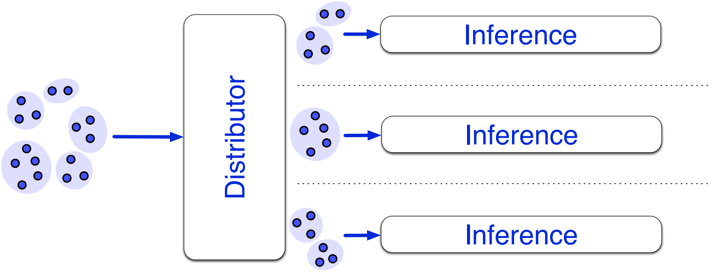


**These two proposals can be evaluated (and accepted) in parallel.**

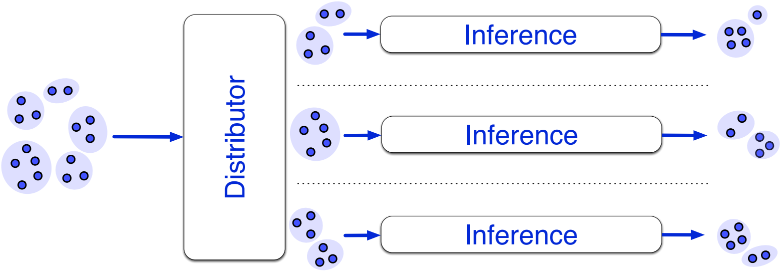
# Distributed Inference



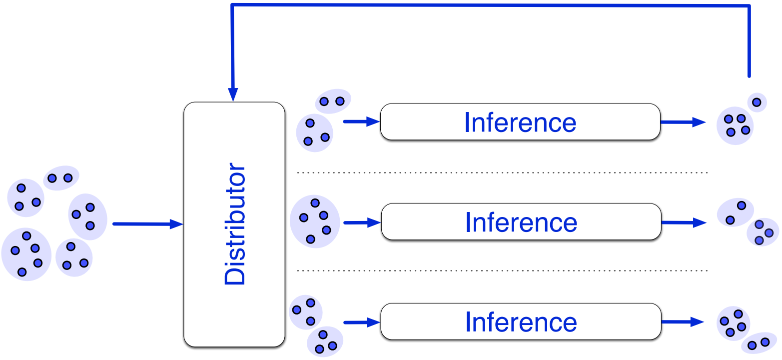
# Distributed Inference



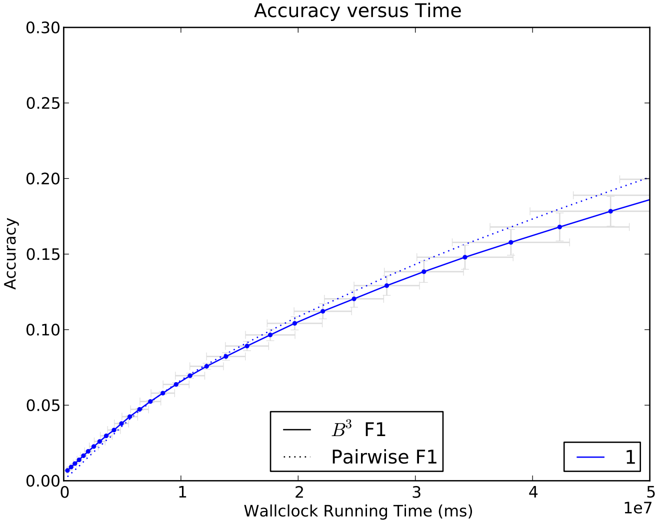
# Distributed Inference



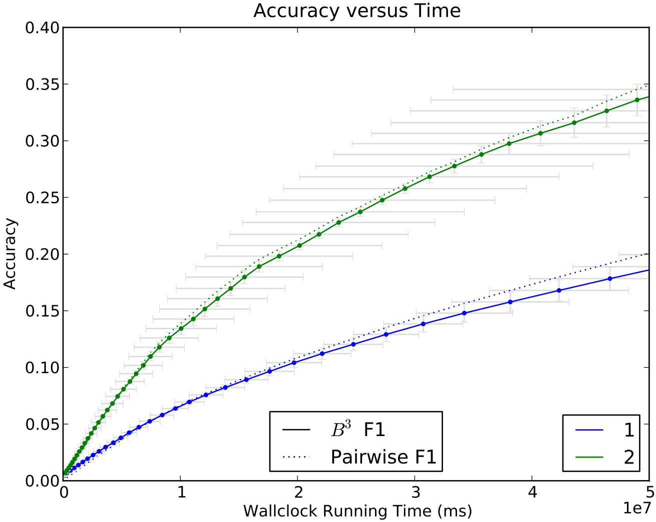
# Distributed Inference



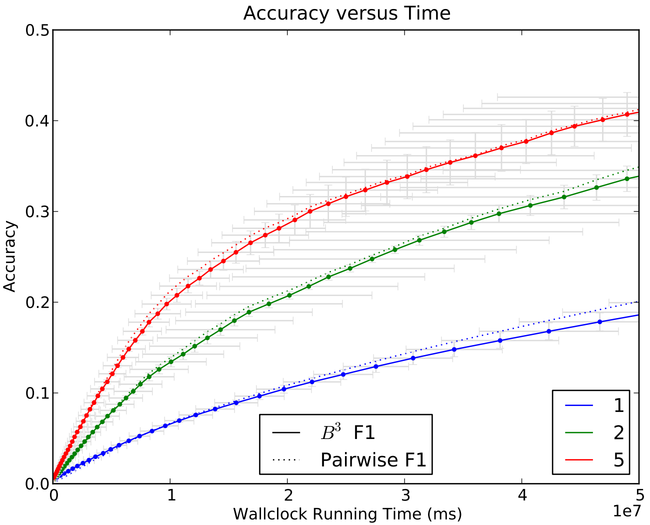
# Experiments



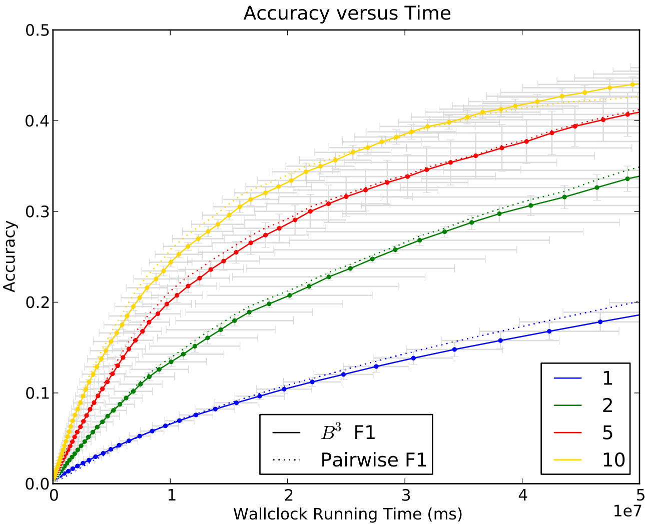
# Experiments



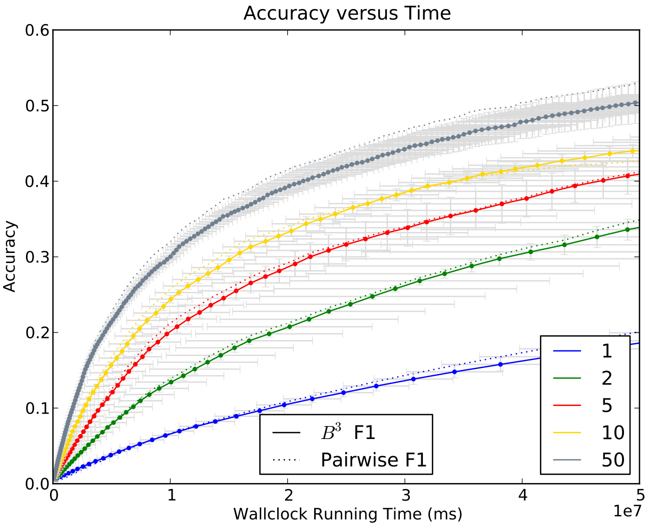
# Experiments



# Experiments

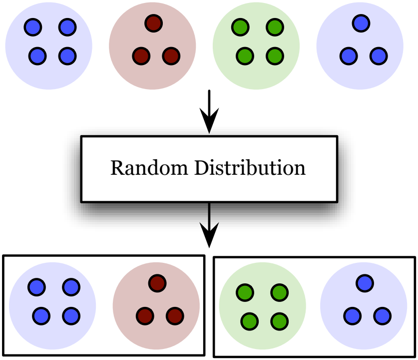


# Experiments



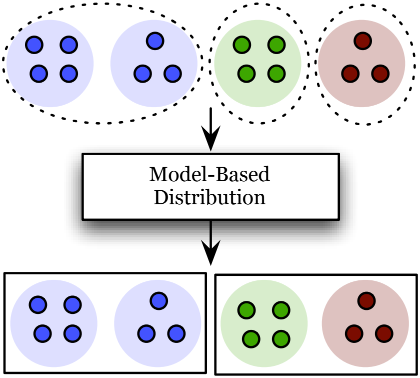
# Outline

# Improving the Distribution



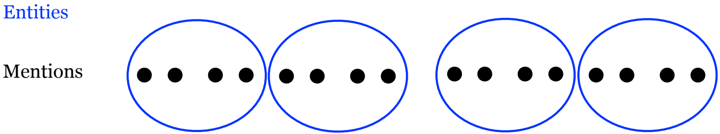

- Random distribution may not assign *similar* entities to the same machine
- Probability that similar entities will be assigned to the same machine is small

# Improving the Distribution

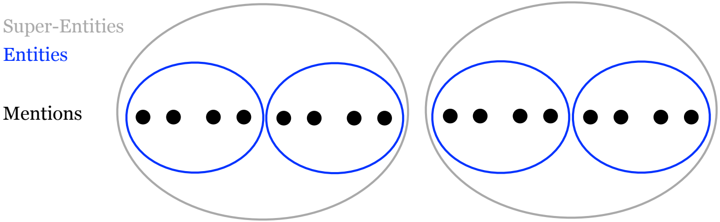



- Include Super-Entities
- Entities in the same super-entity are assigned the same machine

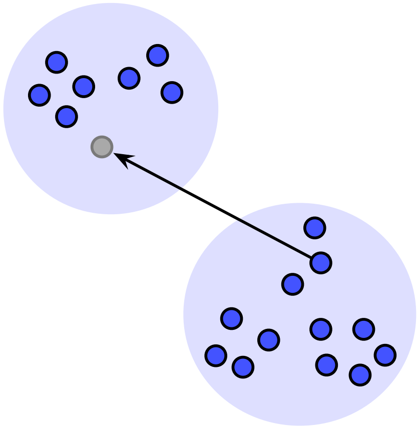
# Super-Entities



# Super-Entities

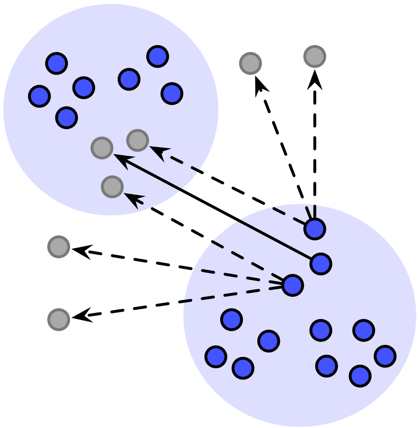


# Within each Worker



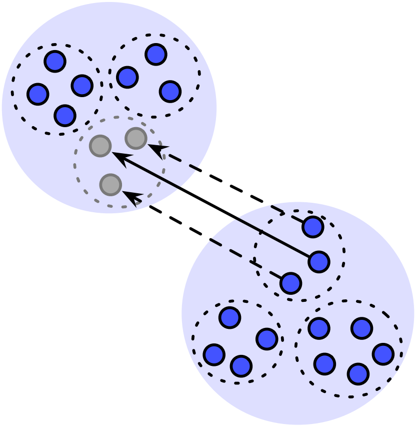
- Consider an accepted move for a mention

# Within each Worker



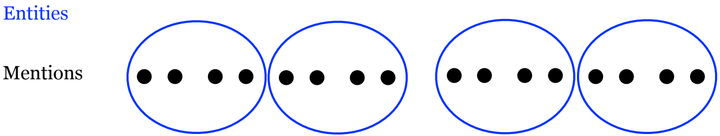
- Ideally, *similar* mentions should also move to the same entity
- Default proposal function does not utilize this
- *Good* proposals become more rare with larger datasets

# Within each Worker

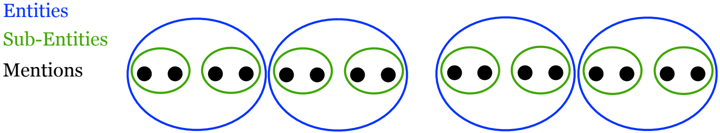


- Include Sub-Entities
- Propose moves of mentions in a sub-entity simultaneously

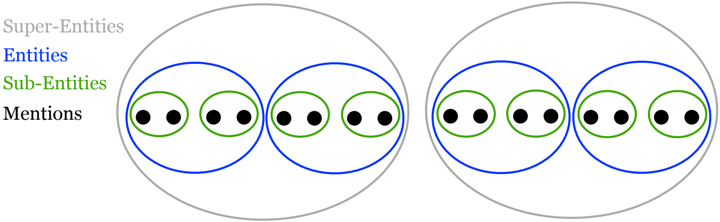
# Sub-Entities



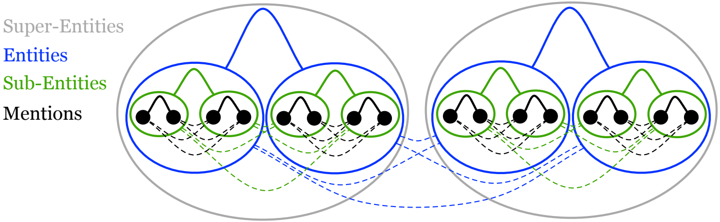
# Sub-Entities



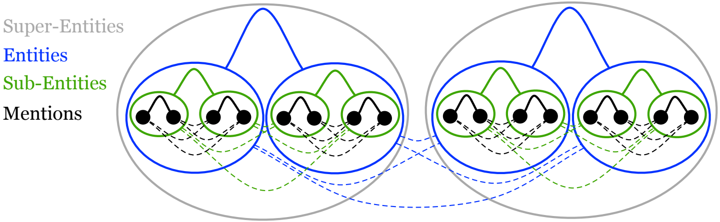
# Hierarchical Representation



# Hierarchical Representation

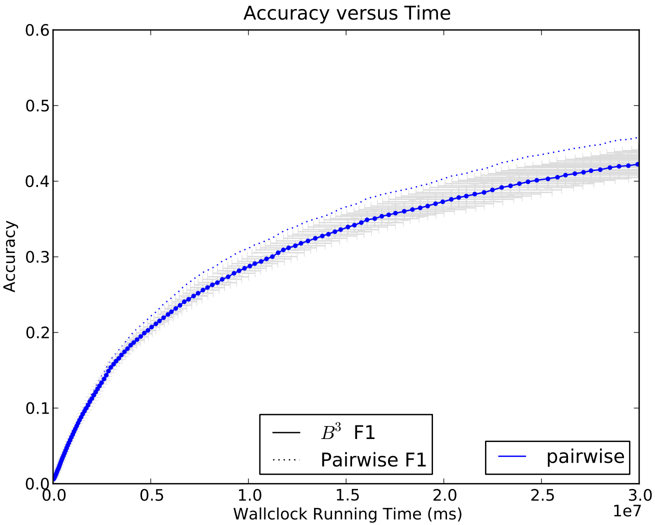


# Hierarchical Representation

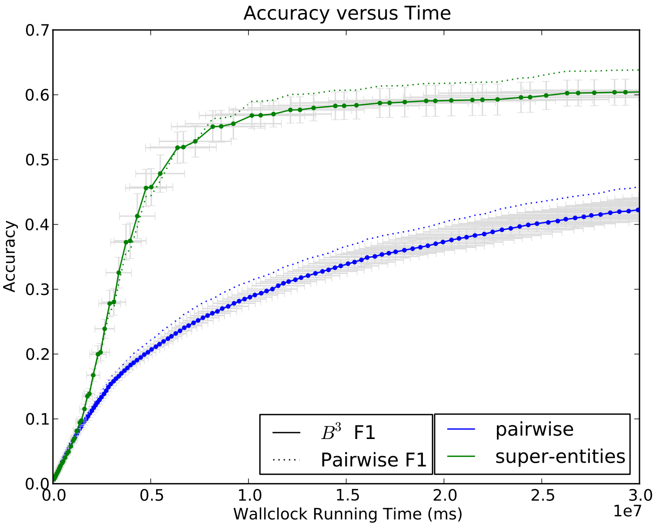



**Sampling:** Fix variables of two levels, sample the remaining level

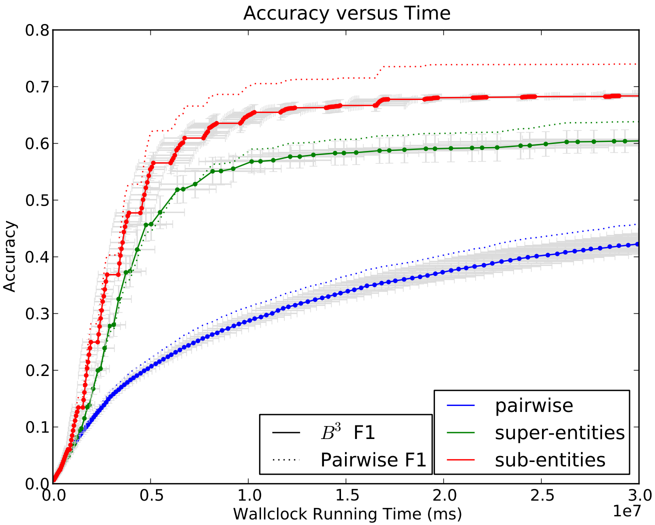
# Experiments



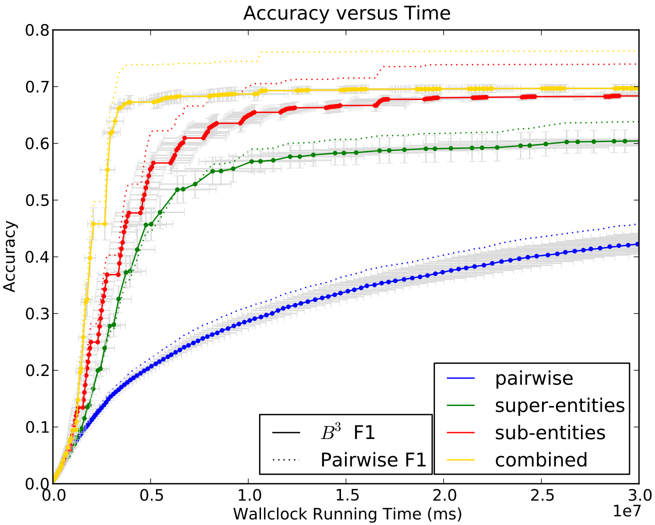
# Experiments



# Experiments



# Experiments



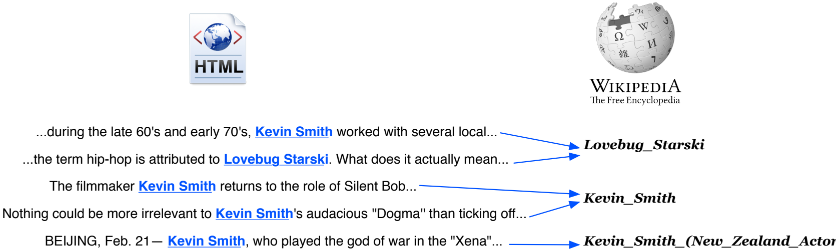
# Outline

# Wikipedia Link Data

- Automatically annotated dataset
  *without compromising on label quality*

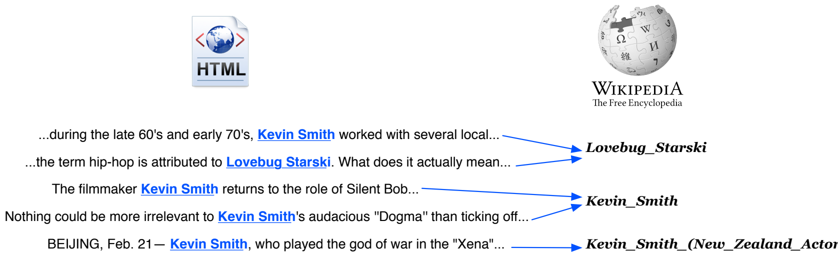
# Wikipedia Link Data

- Automatically annotated dataset
  *without compromising on label quality*
- extract links that point to pages on Wikipedia



# Wikipedia Link Data

- Automatically annotated dataset
  *without compromising on label quality*
- extract links that point to pages on Wikipedia




- treat links (and context) as mentions and target as entity label
- $\sim$1.5 million mentions

# Large-Scale Experiments

**Baselines**

# Large-Scale Experiments

**Baselines**

1. **Unique Strings**
   - Mention with identical mention strings are considered coreferent
   - Often used as approximate cross-document coreference

# Large-Scale Experiments

**Baselines**

1. **Unique Strings**
   - Mention with identical mention strings are considered coreferent
   - Often used as approximate cross-document coreference
2. **Distributed Clustering**
   - Related work performs clustering on the mentions
   - Distributed clustering with same *distance* as ours
   - **Subsquare** is a graph-based approach [Bshouty & Long, ICML 2010]

# Large-Scale Experiments

**Baselines**

1. **Unique Strings**
   - Mention with identical mention strings are considered coreferent
   - Often used as approximate cross-document coreference
2. **Distributed Clustering**
   - Related work performs clustering on the mentions
   - Distributed clustering with same *distance* as ours
   - **Subsquare** is a graph-based approach [Bshouty & Long, ICML 2010]

| Method | Pairwise | | $B^3$ Score | |
|---|---|---|---|---|
| | P/ R | F1 | P/ R | F1 |
| Unique Strings | 30.0 / 66.7 | 41.5 | 82.7 / 43.8 | 57.3 |

# Large-Scale Experiments

**Baselines**

1. **Unique Strings**
   - Mention with identical mention strings are considered coreferent
   - Often used as approximate cross-document coreference
2. **Distributed Clustering**
   - Related work performs clustering on the mentions
   - Distributed clustering with same *distance* as ours
   - **Subsquare** is a graph-based approach [Bshouty & Long, ICML 2010]

| **Method** | **Pairwise** | | **$B^3$ Score** | |
|---|---|---|---|---|
| | P/ R | F1 | P/ R | F1 |
| Unique Strings | 30.0 / 66.7 | 41.5 | 82.7 / 43.8 | 57.3 |
| Subsquare | 38.2 / 49.1 | 43.0 | 87.6 / 51.4 | 64.8 |

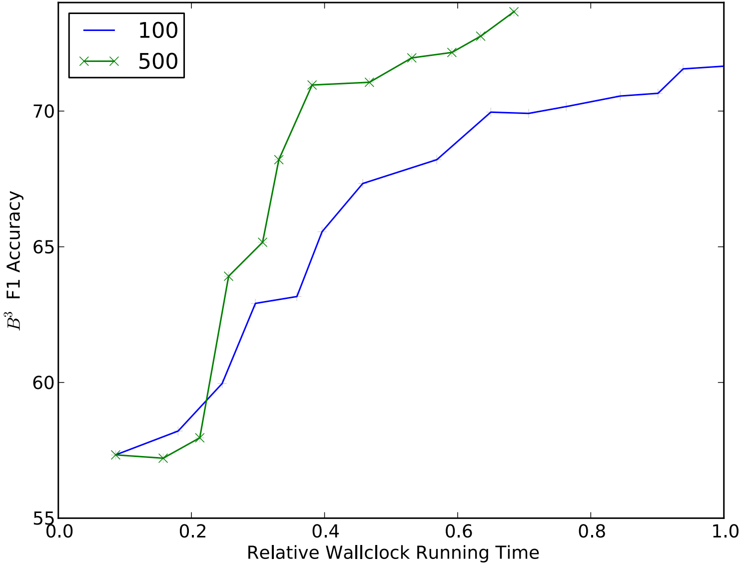
# Large-Scale Experiments

**Baselines**

1. **Unique Strings**
   - Mention with identical mention strings are considered coreferent
   - Often used as approximate cross-document coreference
2. **Distributed Clustering**
   - Related work performs clustering on the mentions
   - Distributed clustering with same *distance* as ours
   - **Subsquare** is a graph-based approach [Bshouty & Long, ICML 2010]

| Method | Pairwise | | $B^3$ Score | |
|---|---|---|---|---|
| | P/ R | F1 | P/ R | F1 |
| Unique Strings | 30.0 / 66.7 | 41.5 | 82.7 / 43.8 | 57.3 |
| Subsquare | 38.2 / 49.1 | 43.0 | 87.6 / 51.4 | 64.8 |
| **Our Model** | 44.2 / 61.4 | **51.4** | 89.4 / 62.5 | **73.7** |

# Large-Scale Experiments



# Conclusions

1. represent cross-doc coreference as a graphical model
2. propose a distributed inference algorithm
3. improve inference with latent hierarchical variables
4. demonstrate utility on large datasets

# Conclusions

1. represent cross-doc coreference as a graphical model
2. propose a distributed inference algorithm
3. improve inference with latent hierarchical variables
4. demonstrate utility on large datasets

**Future Work:**

- more scalability experiments
- study mixing and convergence properties
- add more expressive factors
- supervision: labeled data, noisy evidence

# Thanks!

**Sameer Singh**
sameer@cs.umass.edu

Amarnag Subramanya
asubram@google.com

Fernando Pereira
pereira@google.com

Andrew McCallum
mccallum@cs.umass.edu