Distributed MAP Inference for Undirected Graphical 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

Workshop on Learning on Cores, Clusters and Clouds (LCCC)
Neural Information Processing Systems (NIPS) 2010

Motivation

- Graphical models are used in a number of information extraction tasks
- Recently, models are getting larger and denser
  - Coreference Resolution [Culotta et al. NAACL 2007]
  - Relation Extraction [Riedel et al. EMNLP 2010, Poon & Domingos EMNLP 2009]
  - Joint Inference [Finkel & Manning. NAACL 2009, Singh et al. ECML 2009]
- Inference is difficult, and approximations have been proposed
  - LP-Relaxations [Martins et al. EMNLP 2010]
  - Dual Decomposition [Rush et al. EMNLP 2010]
  - MCMC-Based [McCallum et al. NIPS 2009, Poon et al. AAAI 2008]

Without parallelization, these approaches have restricted scalability

Motivation

Contributions:

1. Distribute MAP Inference for a large, dense factor graph
   - 1 million variables, 250 machines
2. Incorporate sharding as variables in the model

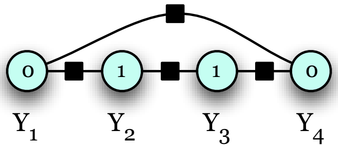
Outline

Factor Graphs

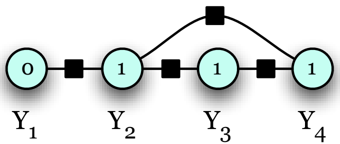
Represent distribution over variables Y using factors ψ.

$$p(Y = y) \propto \exp \sum_{y_c \subseteq y} \psi_c(y_c)$$

Note: Set of factors is different of every assignment $Y = y$ $(\{\psi\}_y)$



$\{\psi\}_{0110} = \{\psi_{12}^{01}, \psi_{23}^{11}, \psi_{34}^{10}, \psi_{14}^{00}\}$



$\{\psi\}_{0111} = \{\psi_{12}^{01}, \psi_{23}^{11}, \psi_{34}^{11}, \psi_{24}^{11}\}$

MAP[1] Inference

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

$$\begin{aligned}\hat{y} &= \arg\max_y p(Y = y) \\ &= \arg\max_y \exp \sum_{y_c \subseteq y} \psi_c(y_c)\end{aligned}$$

Computational bottlenecks:

1. Space of Y is usually enormous (exponential)
2. Even evaluating $\sum_{y_c \subseteq y} \psi_c(y_c)$ for each y may be polynomial

[1] MAP = maximum a posteriori

MCMC for MAP Inference

```
Initial Configuration y = y_0
 for (num_samples):
  1 Propose a change to y to get configuration y'
    (Usually a small change)
  2 Acceptance probability:  α(y, y') = min(1, (p(y')/p(y))^{1/t})
    (Only involve computations local to the change)
  3 if Toss(α):  Accept the change, y = y'
return y
```

Initial Configuration $y = y_0$

for (num_samples):

1. Propose a change to y to get configuration y' (Usually a *small* change)
2. Acceptance probability: $\alpha(y, y') = \min\left(1, \left(\frac{p(y')}{p(y)}\right)^{1/t}\right)$ (Only involve computations local to the change)
3. if Toss(α): Accept the change, $y = y'$

return y

$$\frac{p(y')}{p(y)} = \exp\left\{\sum_{y'_c \subseteq y'} \psi_c(y'_c) - \sum_{y_c \subseteq y} \psi_c(y_c)\right\}$$

Mutually Exclusive Proposals

Let $\{\psi\}_y^{y'}$ be the set of factors used to evaluate a proposal $y \rightarrow y'$

$$\text{i.e. } \{\psi\}_y^{y'} = \left(\{\psi\}_y \cup \{\psi\}_{y'}\right) - \left(\{\psi\}_y \cap \{\psi\}_{y'}\right)$$

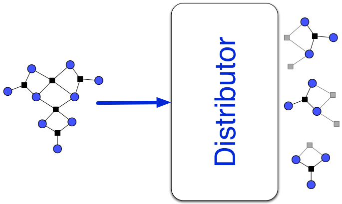
Consider two proposals $y \rightarrow y_a$ and $y \rightarrow y_b$ such that,

$$\{\psi\}_y^{y_a} \cap \{\psi\}_y^{y_b} = \{\}$$

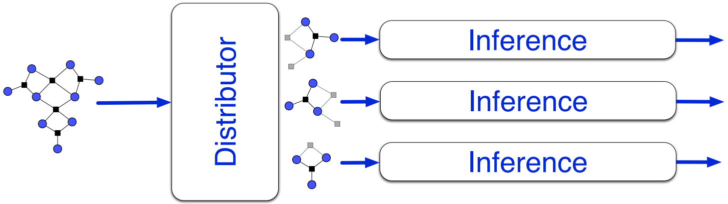
Completely different set of factors are required to evaluate these proposals.

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

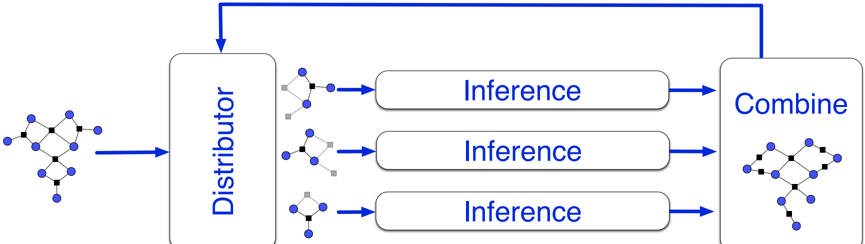
Distributed Inference



Distributed Inference

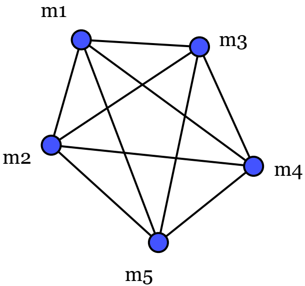


Distributed Inference



Outline

Input Features



Define similarity between mentions, $\phi : \mathcal{M}^2 \rightarrow \mathcal{R}$

- $\phi(m_i, m_j) > 0$: m_i, m_j are similar
- $\phi(m_i, m_j) < 0$: m_i, m_j are dissimilar

We use cosine similarity of the context bag of words:

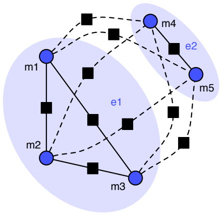
$$\phi(m_i, m_j) = cosSim(\{c\}_i, \{c\}_j) - b$$

Graphical Model

The random variables in our model are entities (E) and mentions (M)
For any assignment to these entities ($E = e$), we define the model score:

$$p(E=e) \propto \exp\left\{ \sum_{m_i \sim m_j} \psi_a(m_i, m_j) + \sum_{m_i \nsim m_j} \psi_r(m_i, m_j) \right\}$$

where $\psi_a(m_i, m_j) = w_a \phi(m_i, m_j)$, and
$\psi_r(m_i, m_j) = -w_r \phi(m_i, m_j)$

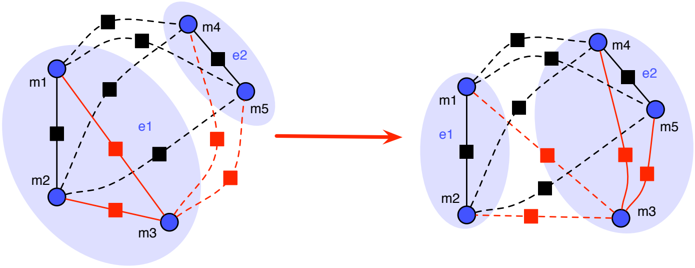


For the following configuration,

$$p(e_1, e_2) \propto \exp \left\{ w_a(\phi_{12} + \phi_{13} + \phi_{23} + \phi_{45}) - w_r(\phi_{15} + \phi_{25} + \phi_{35} + \phi_{14} + \phi_{24} + \phi_{34}) \right\}$$

1. Space of E is Bell Number(n) in number of mentions
2. Evaluating model score for each $E = e$ is $O(n^2)$

MCMC for MAP Inference

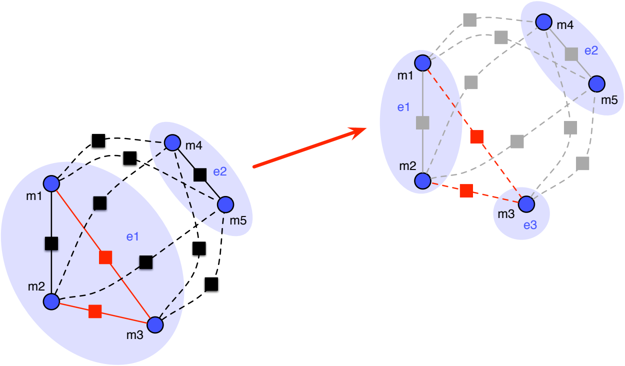


$$p(e) \propto \exp\{w_a\,(\phi_{12} + \phi_{13} + \phi_{23} + \phi_{45}) - w_r(\phi_{15} + \phi_{25} + \phi_{35} + \phi_{14} + \phi_{24} + \phi_{34})\}$$

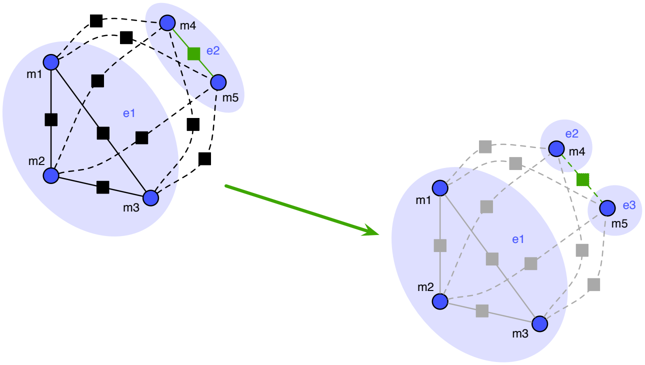
$$p(\acute{e}) \propto \exp\{w_a\,(\phi_{12} + \phi_{34} + \phi_{35} + \phi_{45}) - w_r(\phi_{15} + \phi_{25} + \phi_{13} + \phi_{14} + \phi_{24} + \phi_{23})$$

$$\log \frac{p(\acute{e})}{p(e)} = w_a\,(\phi_{34} + \phi_{35} - \phi_{13} - \phi_{23}) - w_r(\phi_{13} + \phi_{23} - \phi_{34} - \phi_{35})$$

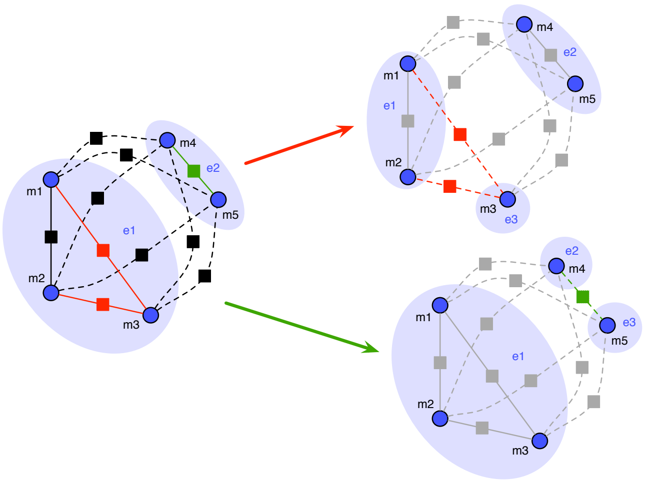
Mutually Exclusive Proposals



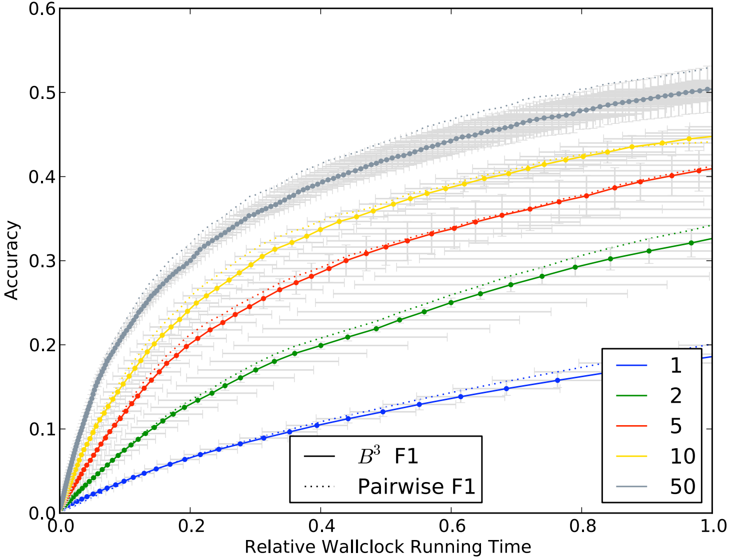
Mutually Exclusive Proposals



Mutually Exclusive Proposals

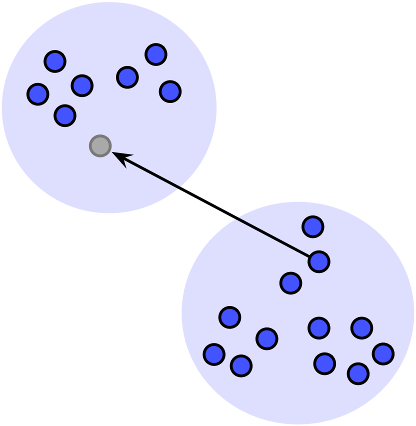


Results


0.6
0.5
0.4
0.3
0.2
0.1
0.0
Accuracy
0.0
0.2
0.4
0.6
0.8
1.0
Relative Wallclock Running Time
B^3 F1
Pairwise F1
1
2
5
10
50

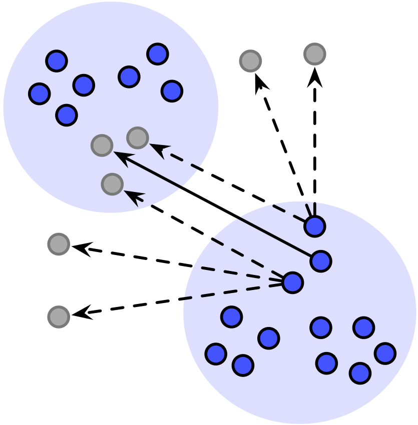
Outline

Sub-Entities



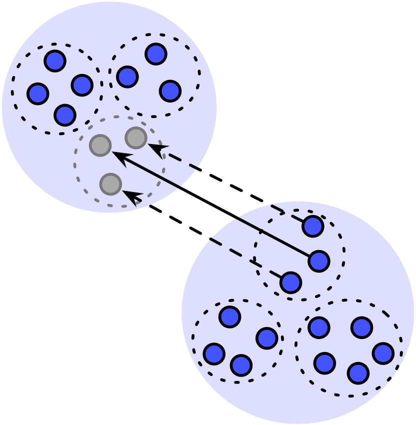
- Consider an accepted move for a mention

Sub-Entities



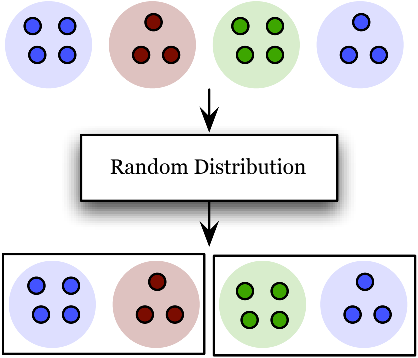
- Ideally, *similar* mentions should also move to the same entity
- Default proposal function does not utilize this
- *Good* proposals become more rare with larger datasets

Sub-Entities



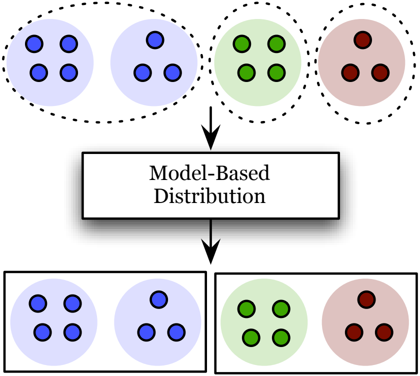
- Include Sub-Entity variables
- Model score is used to sample sub-entity variables
- Propose moves of mentions in a sub-entity simultaneously

Super-Entities



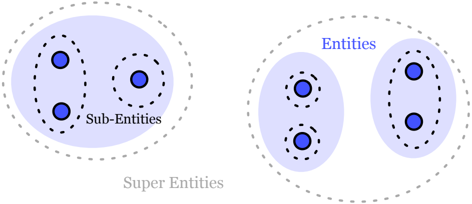
- Random distribution may not assign *similar* entities to the same machine
- Probability that similar entities will be assigned to the same machine is small

Super-Entities



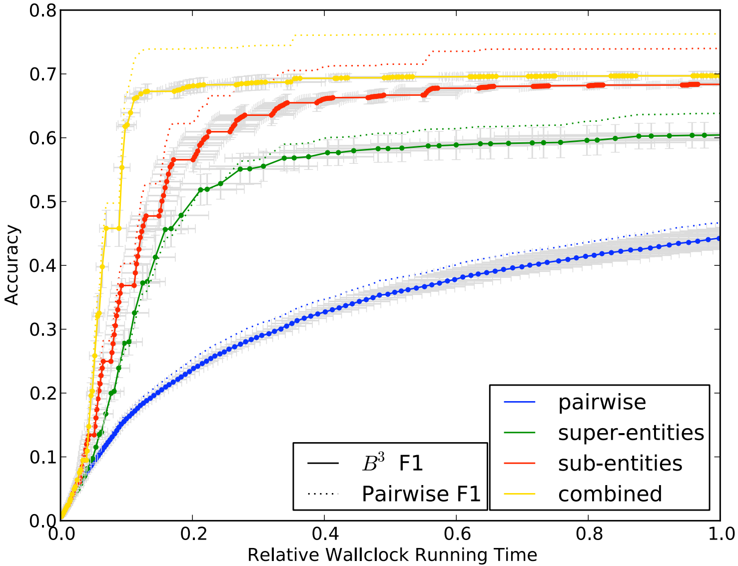
- Augment model with Super-Entities variables
- Entities in the same super-entity are assigned the same machine
- Model score is used to sample super-entity variables

Hierarchical Representation



- **Factors**
  - Affinity factors between mentions / sub-entities / entities in the same sub-entities / entities / super-entities
  - Repulsion factors are similarly symmetric across levels
- **Sampling:** Fix variables of two levels, sample the remaining level

Evaluation


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Accuracy
0.0
0.2
0.4
0.6
0.8
1.0
Relative Wallclock Running Time
B^3 F1
Pairwise F1
pairwise
super-entities
sub-entities
combined

Outline

Preliminary Large-Scale Experiments

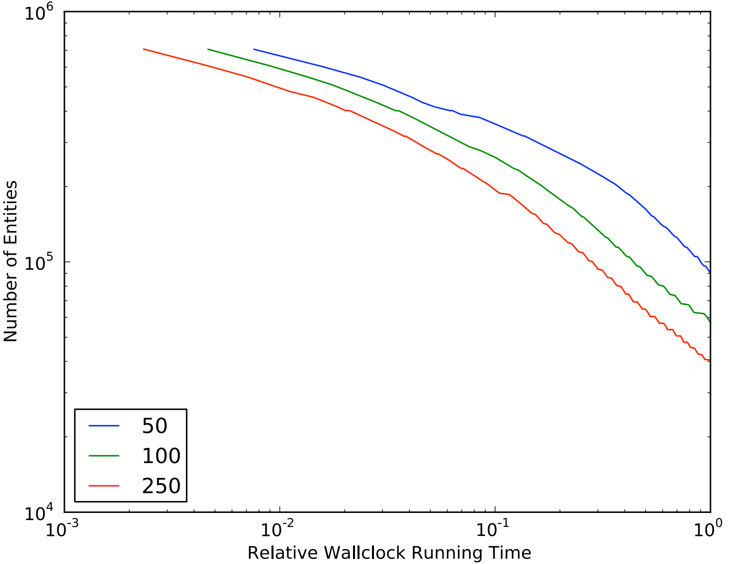
Data

- *New York Times Annotated Corpus* [Sandhous LDC 2008]
  20 years of articles (1987-2007)
- prune rare names ($<$1000): $\sim$1 million person name mentions

Evaluation

- Automated labels are too noisy for evaluation
- Instead, we estimate the speed of inference
  - trust the model to accept good proposals
  - observe the number of predicted entities

Speed of Inference



Related Work

- GraphLab [Low et al. UAI 2010]
  - how do we represent dynamic graphs
  - how do we represent hierarchical models
- Graph Splashing [Gonzalez et al. UAI 2009]
  - graph structure changes with every configuration
  - BP messages are enormous for exponential-domain variables
- Topic Models [Smola & Narayanmurthy. VLDB 2010, Asuncion et al. NIPS 2009]
  - restrictions since they are calculating probabilities
  - we allow non-random distribution and customized proposals

Conclusions

1. propose distributed inference for graphical models
2. enable distributed cross-document coreference
3. improve sharding 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 evidences

Thanks!

Sameer Singh
sameer@cs.umass.edu

Amarnag Subramanya
asubram@google.com

Fernando Pereira
pereira@google.com

Andrew McCallum
mccallum@cs.umass.edu