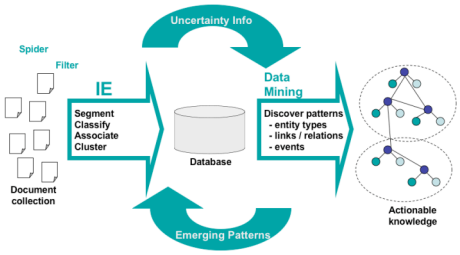
Probabilistic First Order Models for Coreference

Aron Culotta
Information Extraction & Synthesis Lab
University of Massachusetts

joint work with advisor *Andrew McCallum*

Motivation

- Beyond local representation of language
  - Information Extraction
    - Reason about extracted records, not just fields
  - Identity Uncertainty (Coreference resolution)
    - Reason about entities, not just mentions
  - Parsing
    - Global semantic/discourse constraints
  - Joint Extraction and Data Mining



Toward High-Order Representations

Identity Uncertainty

..Howard Dean..

..H Dean..

..Dean Martin..

..Howard Martin..

..Dino..

..Howard..

Toward High-Order Representations

Identity Uncertainty

..Howard Dean..

..H Dean..

..Dean Martin..

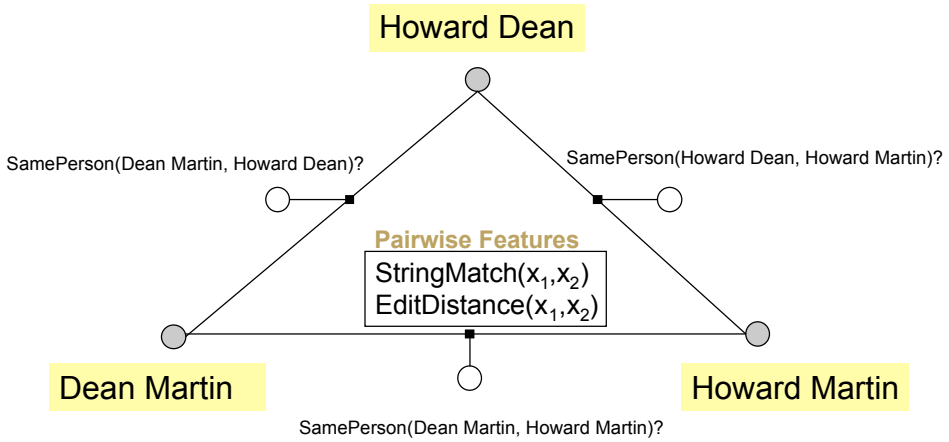
..Dino..

..Howard Martin..

..Howard..

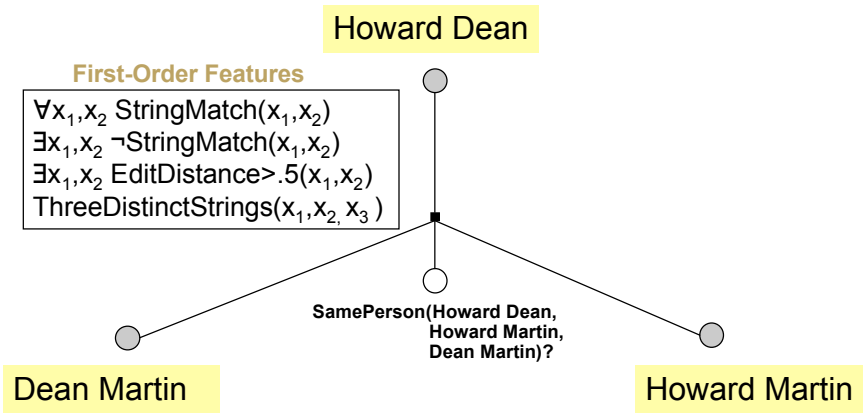
Toward High-Order Representations

Identity Uncertainty



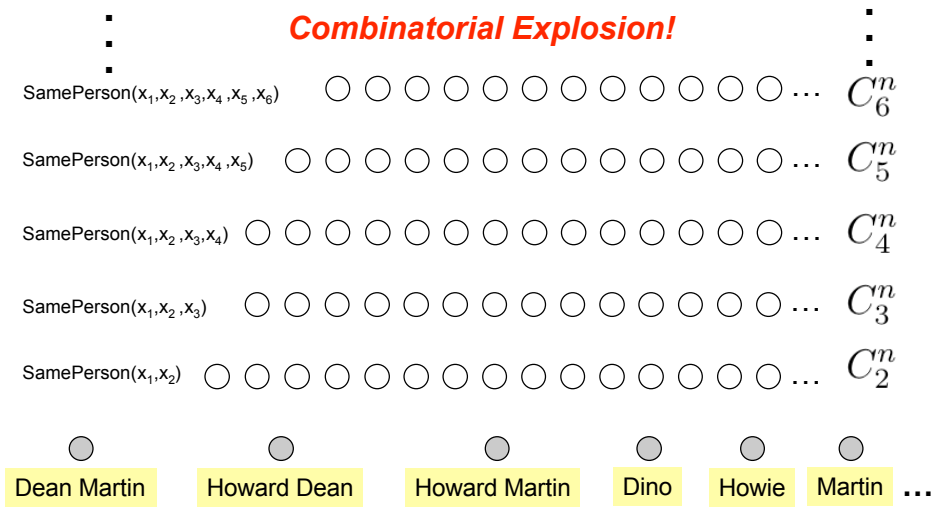
Toward High-Order Representations

Identity Uncertainty



Toward High-Order Representations

Identity Uncertainty



This space complexity is common in first-order probabilistic models

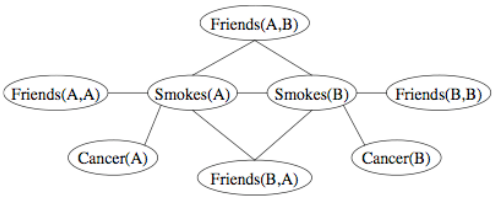
Markov Logic as a Template to Construct a Markov Network using First-Order Logic

[Richardson & Domingos 2005]

English	First-Order Logic	Clausal Form	Weight
Friends of friends are friends.	$\forall x \forall y \forall z\ \mathtt{Fr}(x,y) \wedge \mathtt{Fr}(y,z) \Rightarrow \mathtt{Fr}(x,z)$	$\neg\mathtt{Fr}(x,y) \vee \neg\mathtt{Fr}(y,z) \vee \mathtt{Fr}(x,z)$	0.7
Friendless people smoke.	$\forall x\ (\neg(\exists y\ \mathtt{Fr}(x,y)) \Rightarrow \mathtt{Sm}(x))$	$\mathtt{Fr}(x,\mathtt{g}(x)) \vee \mathtt{Sm}(x)$	2.3
Smoking causes cancer.	$\forall x\ \mathtt{Sm}(x) \Rightarrow \mathtt{Ca}(x)$	$\neg\mathtt{Sm}(x) \vee \mathtt{Ca}(x)$	1.5
If two people are friends, either both smoke or neither does.	$\forall x \forall y\ \mathtt{Fr}(x,y) \Rightarrow (\mathtt{Sm}(x) \Leftrightarrow \mathtt{Sm}(y))$	$\neg\mathtt{Fr}(x,y) \vee \mathtt{Sm}(x) \vee \neg\mathtt{Sm}(y)$, $\neg\mathtt{Fr}(x,y) \vee \neg\mathtt{Sm}(x) \vee \mathtt{Sm}(y)$	1.1 1.1

***ground* Markov network**

$$P(X = x) = \frac{1}{Z} \exp\left(\sum w_i n_i(x)\right)$$



grounding Markov network requires space O(n^r)

***n* = number constants**

***r* = highest clause arity**

How can we perform inference and learning in models that cannot be grounded?

Inference in First-Order Models

SAT Solvers

- Weighted SAT solvers [Kautz et al 1997]
  - Requires complete grounding of network
- LazySAT [Singla & Domingos 2006]
  - Saves memory by only storing clauses that may become unsatisfied

Inference in First-Order Models

MCMC

- Gibbs Sampling
  - Difficult to move between high probability configurations by changing single variables
    - Although, consider MC-SAT [Poon & Domingos '06]
- An alternative: Metropolis-Hastings sampling
  - Can be extended to partial configurations
    - **Only instantiate relevant variables**
  - Successfully used in BLOG models [Milch et al 2005]

Learning in First-Order Models

- Sampling
- Pseudo-likelihood
- Voted Perceptron

- We propose:
  - Conditional model to rank configurations
  - Intuitive objective function for Metropolis-Hastings

Contributions

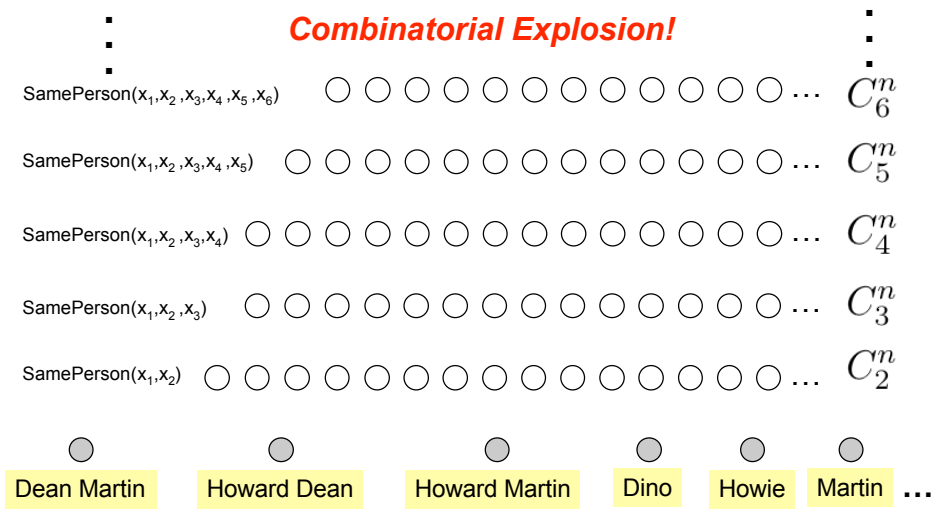
- Metropolis-Hastings sampling in an undirected model with first-order features

- Discriminative training for Metropolis-Hastings

An Undirected Model of Identity Uncertainty

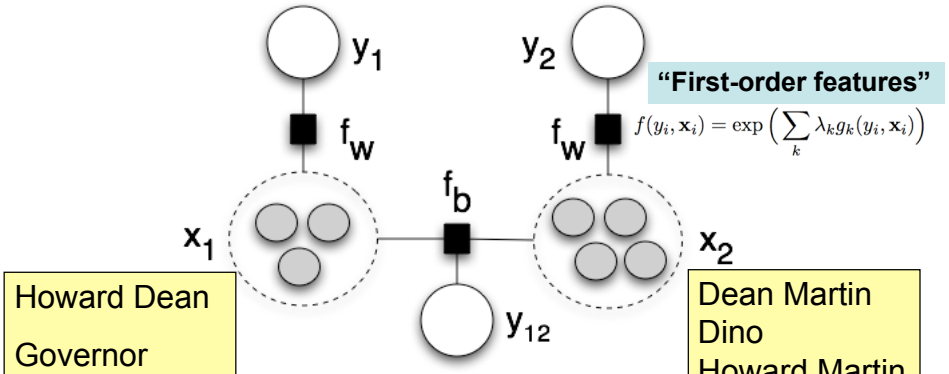
Toward High-Order Representations

Identity Uncertainty



Model

$$P(Y|X) = \frac{1}{Z_X} \prod_{y_i \in Y} f_w(y_i, \mathbf{x}_i) \prod_{y_i, y_j \in Y} f_b(y_{ij}, \mathbf{x}_{ij})$$



"First-order features"

$$f(y_i, \mathbf{x}_i) = \exp\left(\sum_k \lambda_k g_k(y_i, \mathbf{x}_i)\right)$$

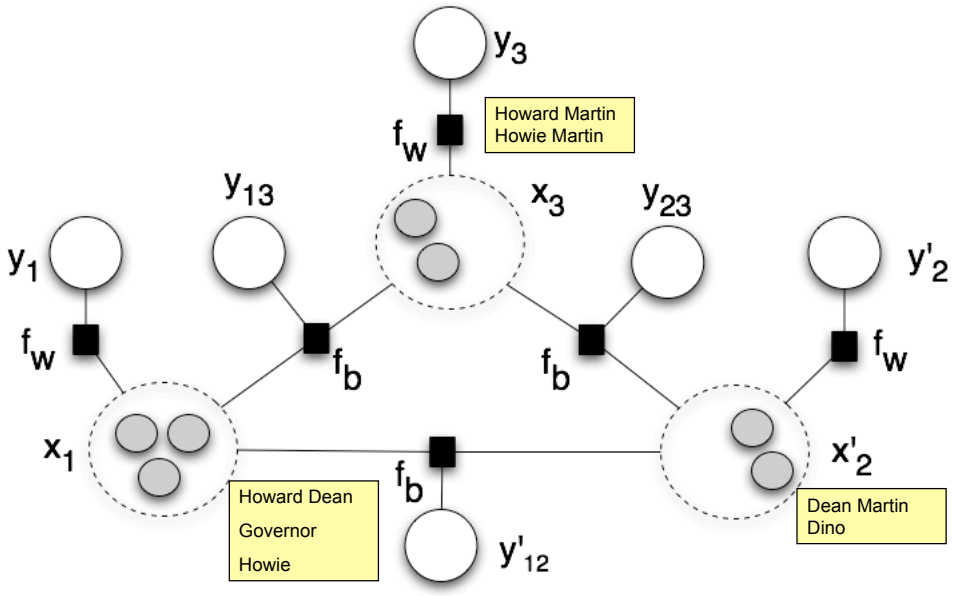
Howard Dean
Governor
Howie

Dean Martin
Dino
Howard Martin
Howie Martin

$\mathbf{f_w}$: SamePerson(**x**)

$\mathbf{f_b}$: DifferentPerson(**x, x'**)

Model



Model

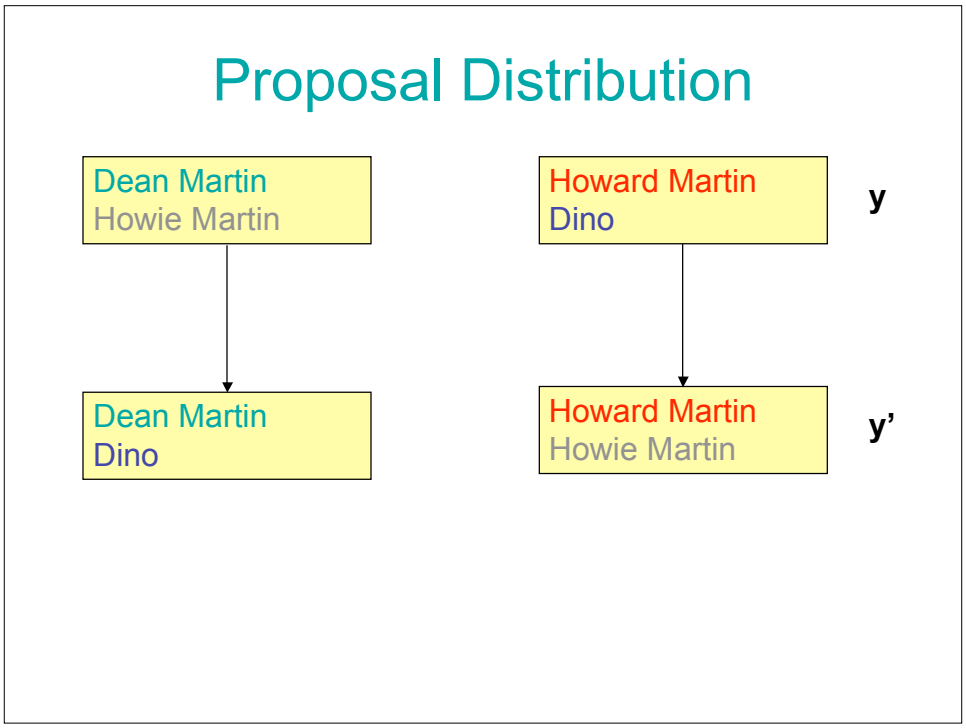
$$P(Y|X) = \frac{1}{Z_X} \prod_{y_i \in Y} f_w(y_i, \mathbf{x}_i) \prod_{y_i, y_j \in Y} f_b(y_{ij}, \mathbf{x}_{ij})$$

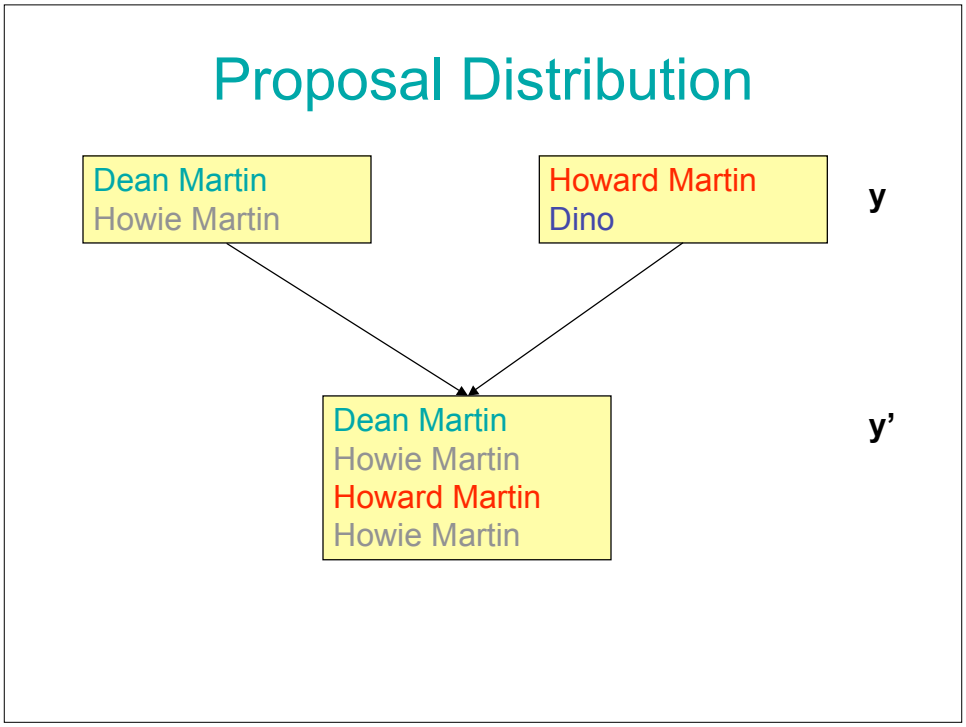
Z_X: Sum over all possible configurations!

Inference with Metropolis-Hastings

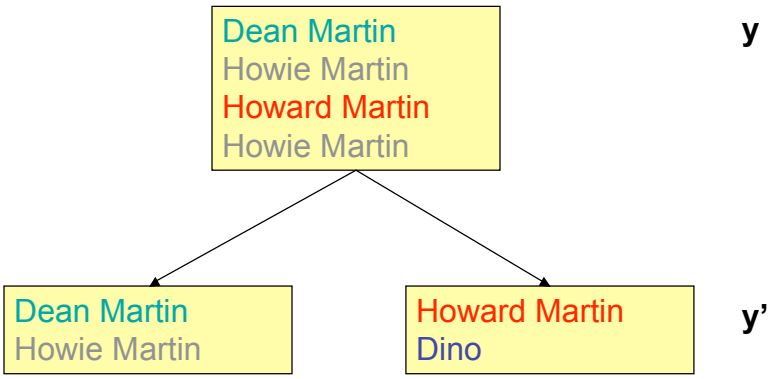
$$\alpha(y'|y) = \frac{p(y')}{p(y)} \cdot \frac{q(y'|y)}{q(y|y')}$$

- y : configuration
- p(y')/p(y) : **likelihood ratio**
  - Ratio of P(Y|X)
  - Z_X cancels
- q(y'|y) : **proposal distribution**
  - probability of proposing move y ⇒y'


Proposal Distribution
Dean Martin
Howie Martin
Howard Martin
Dino
y
Dean Martin
Dino
Howard Martin
Howie Martin
y'


Proposal Distribution
Dean Martin
Howie Martin
Howard Martin
Dino
y
Dean Martin
Howie Martin
Howard Martin
Howie Martin
y'

Proposal Distribution



Learning the Likelihood Ratio

$$\alpha(y'|y) = \frac{p(y')}{p(y)} \cdot \frac{q(y'|y)}{q(y|y')}$$

Given a pair of configurations, learn to rank the "better" configuration higher.

Learning the Likelihood Ratio

$$P(Y|X) \propto S(Y) = \prod_{y_i \in Y} f_w(y_i, \mathbf{x}_i) \prod_{y_i, y_j \in Y} f_b(y_{ij}, \mathbf{x}_{ij})$$

S*(Y) = true evaluation of configuration (e.g. F1)

$$Z_{p>q} = 1 \Leftrightarrow S^*(Y_p) > S^*(Y_q)$$

$$P(Z_{p>q} = 1) = \frac{S(Y_p)}{S(Y_p) + S(Y_q)}$$

Sampling Training Examples

- Run sampler on training data
- Generate training example for each proposed move
- Iteratively retrain during sampling

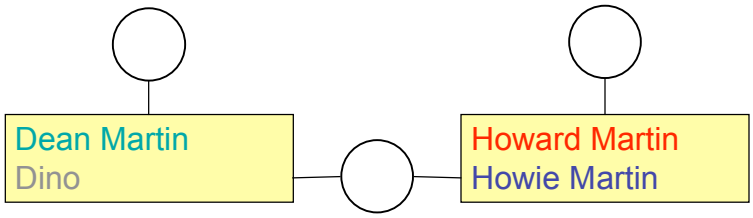
Tying Parameters with Proposal Distribution

- Proposal distribution q(y'|y) "cheap" approximation to p(y)
- Reuse subset of parameters in p(y)
- E.g. in identity uncertainty model
  - Sample two clusters
  - Stochastic agglomerative clustering to propose new configuration

Experiments

Simplified Model

- Use only **within-cluster** factors.
- Inference with agglomerative clustering



Experiments

- Paper citation coreference
- Author coreference
- First-order features
  - All Titles Match, Exists Year MisMatch, Average String Edit Distance > X, …
  - Number of mentions

Results on Citation Data

	First-Order	Pairwise
constraint	**82.3**	76.7
reinforce	**93.4**	78.7
face	**88.9**	83.2
reason	81.0	**84.9**

Citeseer paper coreference results (pair F1)

	First-Order	Pairwise
miller_d	41.9	**61.7**
li_w	**43.2**	36.2
smith_b	**65.4**	25.4

Author coreference results (pair F1)

Conclusions

- Enable tractable training of first-order features in relational models
- Higher-order representations can help identity uncertainty

Related Work

- MLNs [Richardson et al 2006]
- BLOG [Milch et al 2005]
- Lifted Inference [Poole '03] [Braz et al '05]
  - Inference over populations to avoid grounding network
  - Difficult to answer queries about one specific input
- SEARN [Daume et al 2005]:
  - Learns distribution over possible moves in search-based inference
  - Assumes can enumerate all local moves
- Reinforcement learning for combinatorial search
  - [Zhang and Dietterich '95] [Boyan '98]