Unsupervised Coreference Resolution in a Nonparametric Bayesian Model

Aria Haghighi and Dan Klein

Presented by Brandon Norick

Overview

- Introduction
- Preliminaries
- Coreference Resolution Models
- Experiments
- Conclusion

Introduction

- When speaking or writing natural language there are two processes which govern references to entities
  - New entities are introduced, generally with proper or nominal expressions
  - References are made back to entities which have already been introduced, generally with pronouns
- Problem: how can a computer determine which entity references actually refer to the same entity (i.e., are coreferent)?

Introduction

An example

The Weir Group, whose headquarters is in the US, is a large, specialized corporation investing in the area of electricity generation. This power plant, which will be situated in Rudong, Jiangsu, has an annual generation capacity of 2.4 million kilowatts.

Introduction

An example

The Weir Group, whose headquarters is in the US, is a large, specialized corporation investing in the area of electricity generation. This power plant, which will be situated in Rudong, Jiangsu, has an annual generation capacity of 2.4 million kilowatts.

Introduction

An example

$\boxed{\text{The Weir Group}}$, $\boxed{\text{whose}}$ $\boxed{\text{headquarters}}$ … $\boxed{\text{the US}}$

… $\boxed{\text{corporation}}$ …

… $\boxed{\text{power plant}}$,

$\boxed{\text{which}}$ … $\boxed{\text{Rudong}}$, $\boxed{\text{Jiangsu}}$ …

…

For the problem of coreference resolution, we are only interested in entity references and the rest of the text is ignored.

Background

Related work

- Primary approach is to treat the problem as a set of pairwise coreference decisions
  – Use discriminative learning with features encoding properties such as distance and environment
- However, there are several problems with this approach
  – In order to have rich features, a large amount of data is required, which is typically unavailable
  – In order to partition, a greedy approach is generally taken which relies solely on the pairwise model

Preliminaries

- **Each document consists of a set of *mentions* (usually noun phrases)**
- **A *mention* is a *reference* to some entity**
- **There are three types of mentions:**
  - ***proper* (names)**
  - ***nominal* (descriptions)**
  - ***pronominal* (pronouns)**
- **Therefore, the coreference resolution problem is to partition the mentions according to their referents**

Preliminaries

- During the design process for the final model, the authors used data from the Automatic Context Extraction (ACE) 2004 task
  - This data was used to test performance, as well as for hyperparameter selection
  - Used English translations of the Arabic and Chinese treebanks
  - 95 documents, 3905 mentions

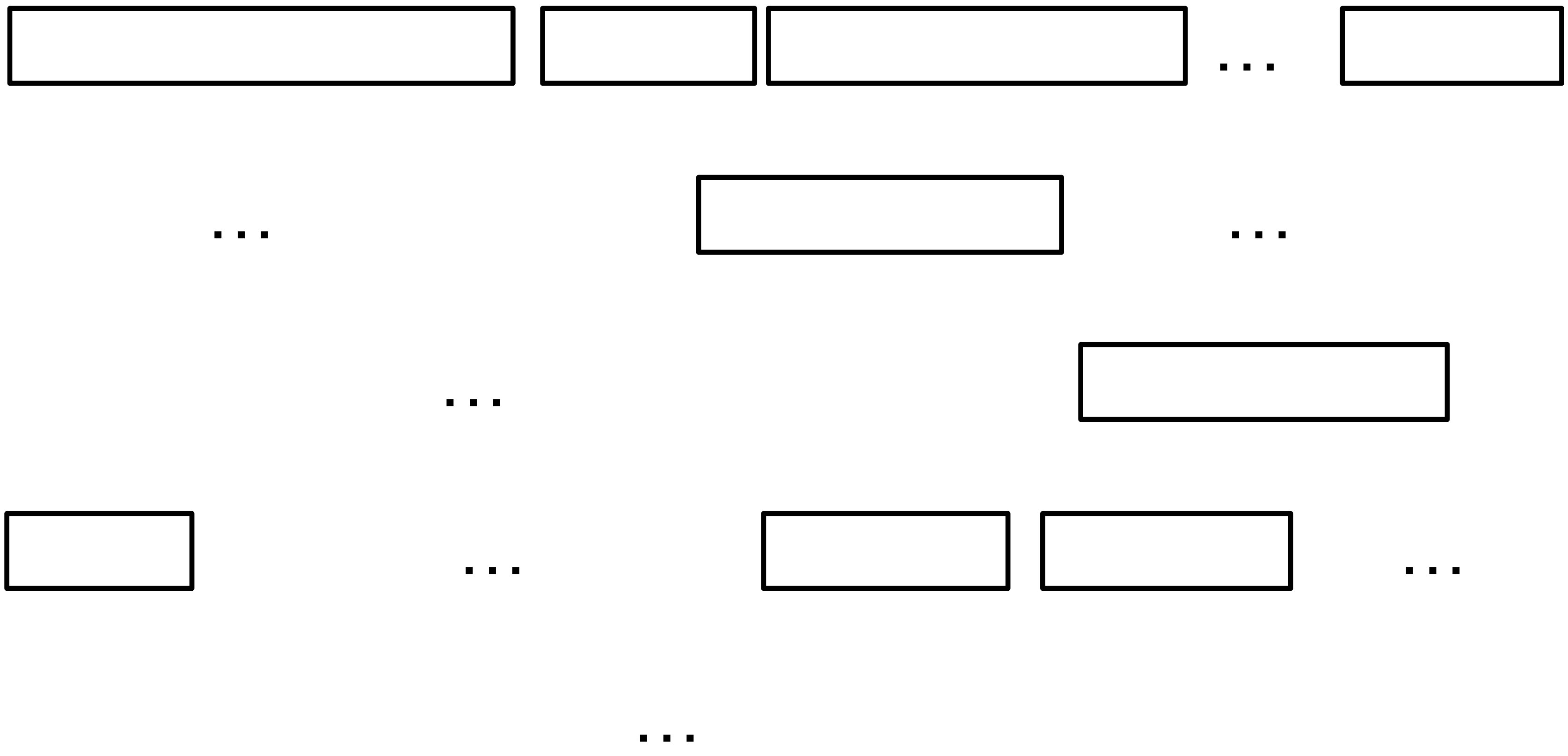
Preliminaries

Some assumptions

- The system assumes that the following data is provided as input:
  - The true mention boundaries
  - The head words for mentions (i.e., the "main" word of a mention, such as "a big sheep dog)
  - The mention types
- Unlike related work, named entity recognition labels and part of speech tags are <u>not</u> required

Coreference Resolution Models

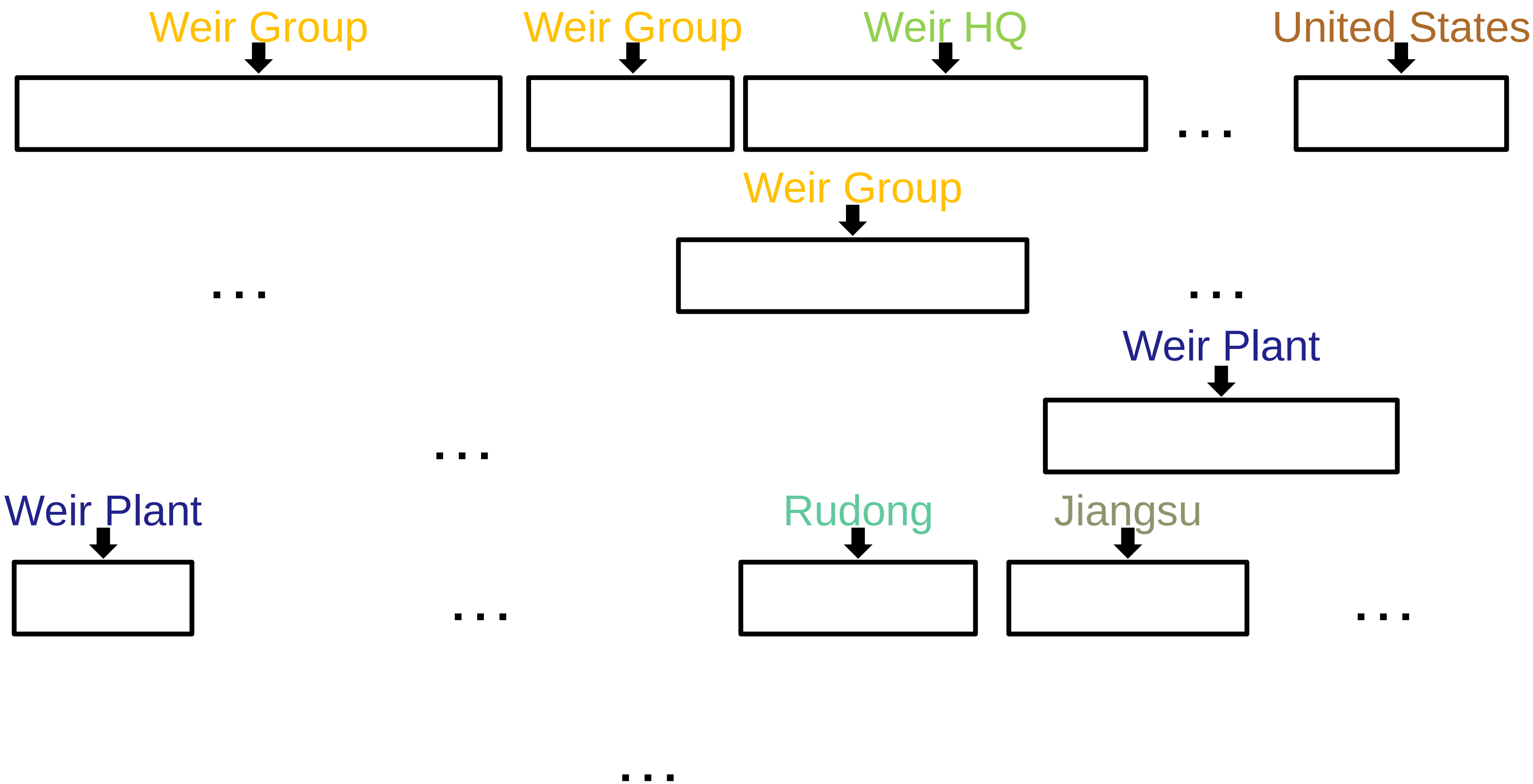
Generative story



Coreference Resolution Models

Generative story

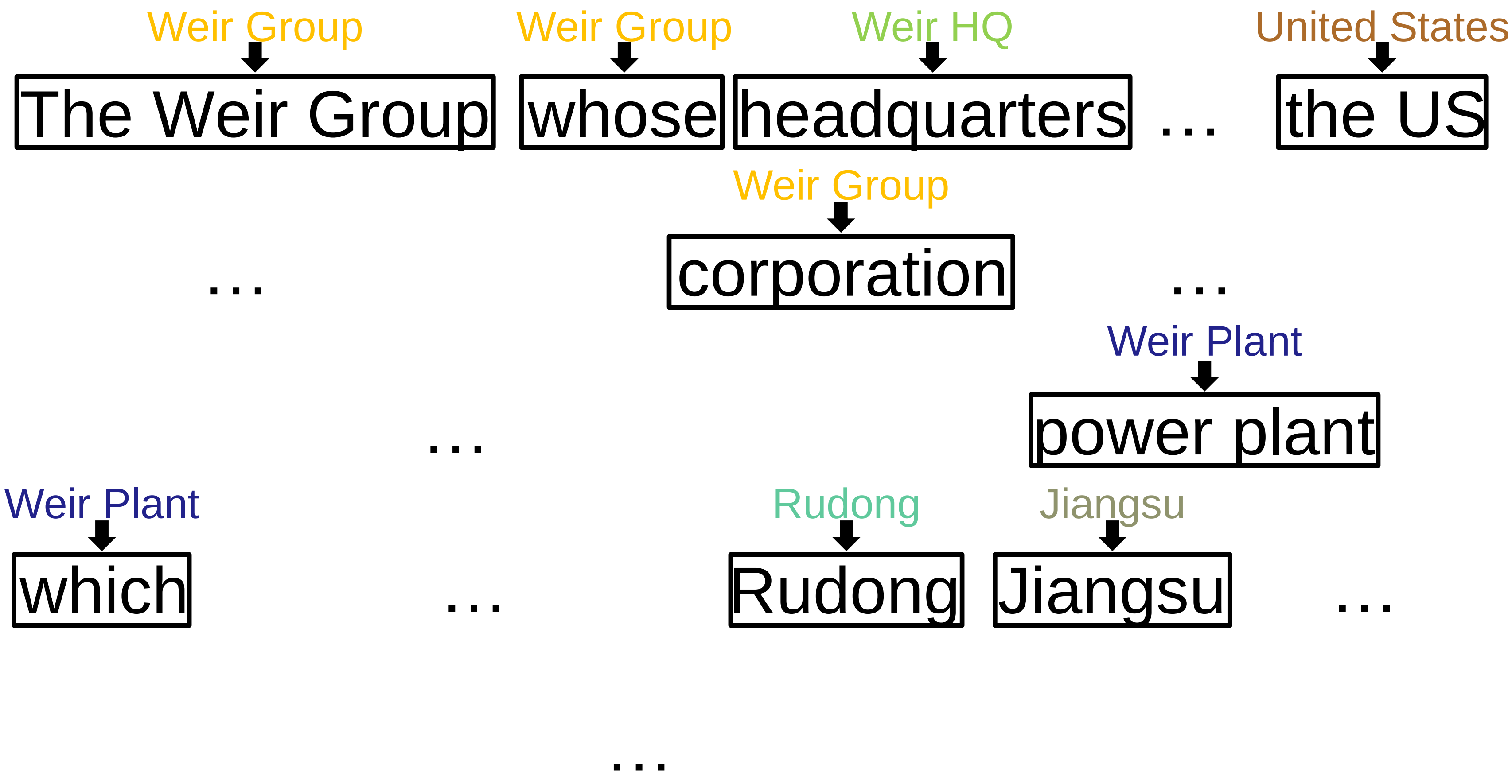
First, generate entities



Coreference Resolution Models

Generative story

Then, generate mentions according to these entities



Coreference Resolution Models

Finite mixture model

- **Documents are independent, with the exception of some global hyperparameters**
- **Each document is a mixture of a fixed number of components, *K***
- **The distribution over entities is drawn from a symmetric Dirichlet distribution**

$$\beta \sim Dir(\alpha)$$

- **The entity for each mention is drawn from beta**

$$z \sim \beta$$

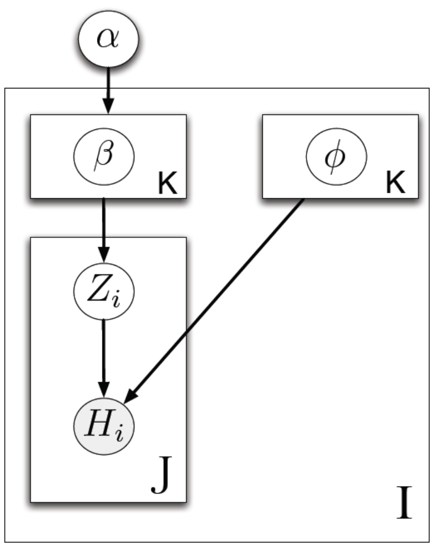
Coreference Resolution Models

Finite mixture model

- **Each entity is associated with a multinomial distribution over head words, these are also drawn from a symmetric Dirichlet distribution**

$$\phi_Z^h \sim Dir(\lambda_H)$$

- **The head word for each mention is drawn from the associated multinomial**
- **The graphical model for this approach, where shaded nodes represent observed variables**



Coreference Resolution Models

Finite mixture model

- **Gibbs sampling to obtain samples from *P(*Z|X*)* where X represents the variables associated with mentions, in this case only the head words**

$$P(Z_{i,j}|\mathbf{Z}^{-i,j},\mathbf{H}) \propto P(Z_{i,j}|\mathbf{Z}^{-i,j})P(H_{i,j}|\mathbf{Z},\mathbf{H}^{-i,j})$$

$$P(Z_{i,j}=z|\mathbf{Z}^{-i,j}) \propto n_z + \alpha$$

$$P(H_{i,j}=h|\mathbf{Z},\mathbf{H}^{-i,j}) \propto n_{h,z} + \lambda_H$$

Coreference Resolution Models

Finite mixture model

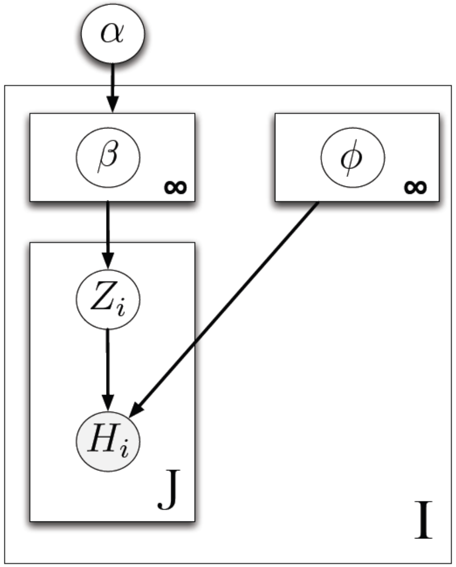
- **A big problem with this model is that the number of entities, *K*, must be fixed a priori**
- **What we want is for the model to be able to select *K* itself, in a manner which fits the data**
- **In order to accomplish this in a principled manner, the authors suggest the use of a Dirichlet process (DP), which allows for a countably infinite number of entities**

Coreference Resolution Models

Infinite mixture model

- **The new graphical model, where the Dirichlet priors have been replaced**
- **Now:**

$$P(Z_{i,j} = z|\mathbf{Z}^{-i,j}) \propto \begin{cases} \alpha, & \text{if } z = z_{new} \\ n_z, & \text{otherwise} \end{cases}$$



Coreference Resolution Models

Infinite mixture model

- **This approach is still rather crude, and has trouble with pronominal mentions**

The Weir Group$_1$, whose$_2$ headquarters$_3$ is in the US$_4$, is a large, specialized corporation$_5$ investing in the area of electricity generation. This power plant$_6$, which$_7$ will be situated in Rudong$_8$, Jiangsu$_9$, has an annual generation capacity of 2.4 million kilowatts.

- **The entity specific multinomials in this approach are effective for proper and some nominal mentions, but do not make sense for pronominal mentions**
  - **All entities can be referred to with pronouns, and the choice depends on entity properties rather than the specific entity**

Coreference Resolution Models

Pronoun head model

- Now, when generating a head word for a mention we consider more than the entity specific multinomial distribution over head words
- Also consider entity specific distributions over the properties
  - Entity type (Person, Location, Organization, Misc.)
  - Gener (Male, Female, Neuter)
  - Number (Single, Plural)

Coreference Resolution Models

Pronoun head model

- **Each of these property distributions is assumed to be a draw from symmetric Dirichlet distributions with small concentration parameters, encouraging peakedness**

Coreference Resolution Models

Pronoun head model

- The generative story for mentions is now slightly different
  - Draw an entity type *T*, a gender *G*, and a number *N* from the appropriate distributions
  - Draw a mention type *M* from a global multinomial (sym. Dir. with λ_M)
  - A head word is then generated conditioned on these properties and the mention type
    - If *M* is not pronoun, the head word is drawn directly from the entity head word multinomial as before
    - Otherwise, the head word is drawn based on the global pronoun head distribution, conditioning on the properties

Coreference Resolution Models

Pronoun head model

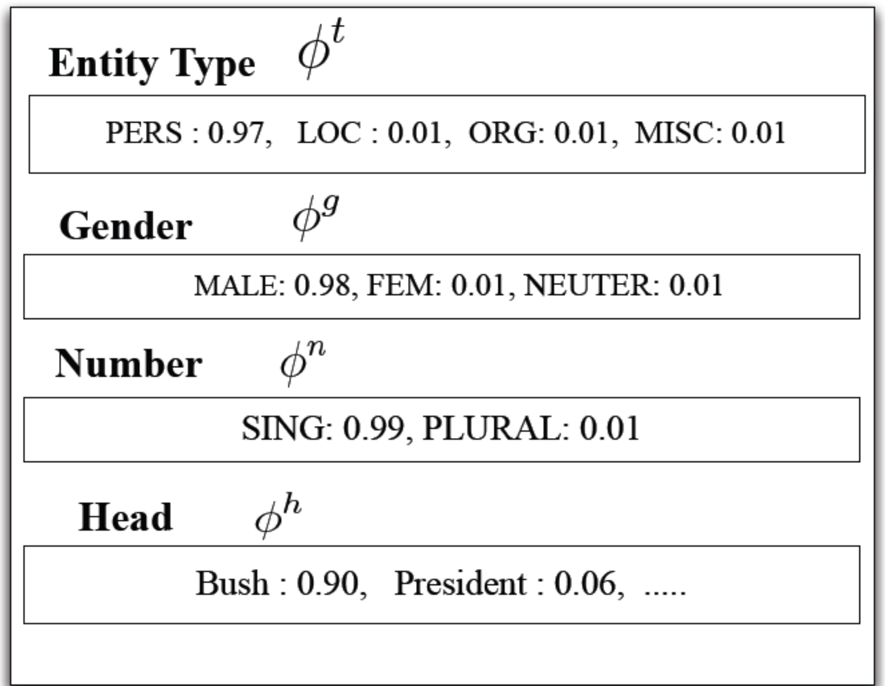
- **More specifically,**

$$P(H|Z,T,G,N,M,\phi,\boldsymbol{\theta}) = \begin{cases} P(H|T,G,N,\boldsymbol{\theta}), & \text{if } M = \text{PRO} \\ P(H|\phi_Z^h), & \text{otherwise} \end{cases}$$

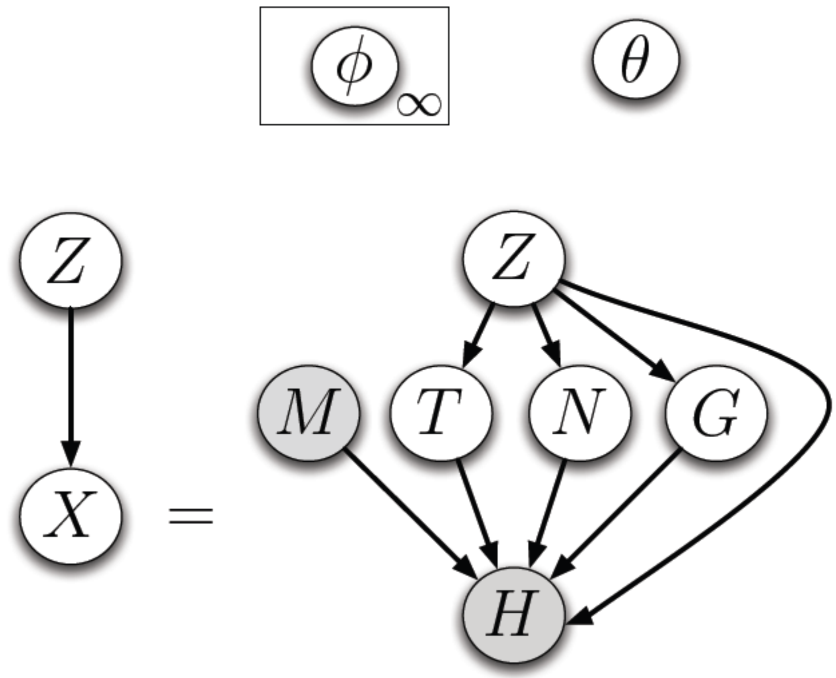
- **Use the prior on theta, the parameters for the global pronoun head distribution, to encode compatible entity types for a pronoun (e.g., "he" with "Person")**

Coreference Resolution Models

Pronoun head model



An example of the parameters associated with an entity



The graphical model for this approach

Coreference Resolution Models

Pronoun head model

- Substantial improvement, achieving a MUC F_1 of 64.1

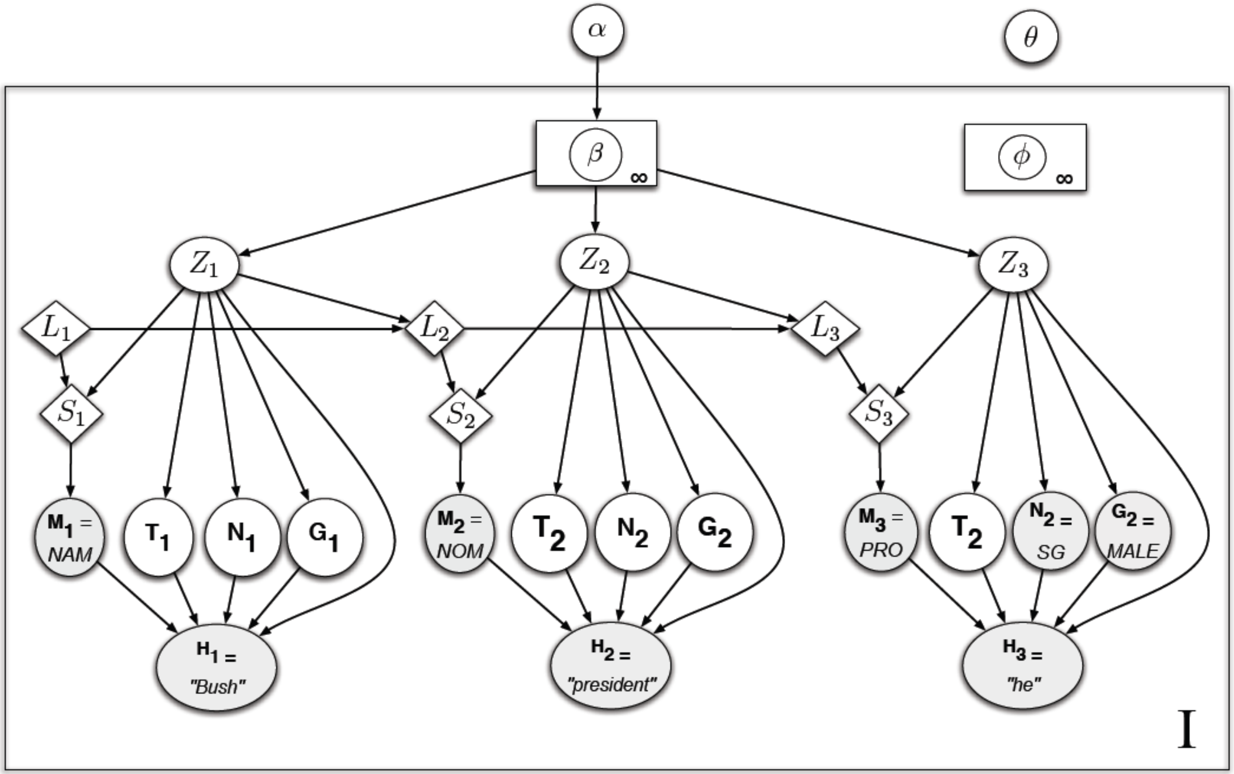
The Weir Group$_1$, whose$_1$ headquarters$_2$ is in the US$_3$, is a large, specialized corporation$_4$ investing in the area of electricity generation. This power plant$_5$, which$_1$ will be situated in Rudong$_6$, Jiangsu$_7$, has an annual generation capacity of 2.4 million kilowatts.

- However, there is no local preference for pronominal mentions exists in this model
- Introduce salience to address this issue

Coreference Resolution Models

Adding salience

- **The new graphical model is as follows:**



Coreference Resolution Models

Adding salience

- As the mentions in a document are generated, a list of active entities and their salience scores is maintained
  - When an entity is mentioned, its score is incremented by 1
  - When moving to generate the next mention, all scores decay by a factor of 0.5
- Based on the list of scores, *L*, each entity *z* has a rank on this list which can be in one of five buckets: Top (1), High (2-3), Mid (4-6), Low (7+), or None

Coreference Resolution Models

Adding salience

- **This changes the sampling equation, which now has to account for how future salience values change when sampling an entity**

$$P(Z_{i,j} = z|\mathbf{Z}^{-i,j}) \propto n_z \prod_{j' \geq j} P(M_{i,j'}|S_{i,j'}, \mathbf{Z})$$

- **This approach fixes the final error exhibited by the previous models, and gives an F_1 of 71.5**

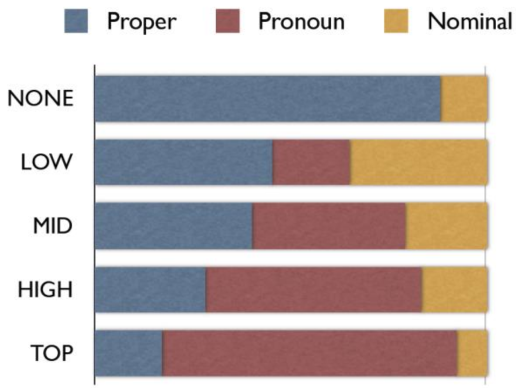
$\boxed{\text{The Weir Group}_1\text{, whose}_1}$ headquarters$_2$ is in the US$_3$, is a large, specialized corporation$_4$ investing in the area of electricity generation. This $\boxed{\text{power plant}_5\text{, which}_5}$ will be situated in Rudong$_6$, Jiangsu$_7$, has an annual generation capacity of 2.4 million kilowatts.

Coreference Resolution Models

Adding salience

- **The posterior distribution of mention type *M* given salience *S* is described in the following table**

Salience Feature	Pronoun	Proper	Nominal
TOP	0.75	0.17	0.08
HIGH	0.55	0.28	0.17
MID	0.39	0.40	0.21
LOW	0.20	0.45	0.35
NONE	0.00	0.88	0.12



- **Pronoun type is preferred for the entities with Top or High salience, whereas proper and nominal types are preferred otherwise**

Figure from slides by Aria Haghighi

Coreference Resolution Models

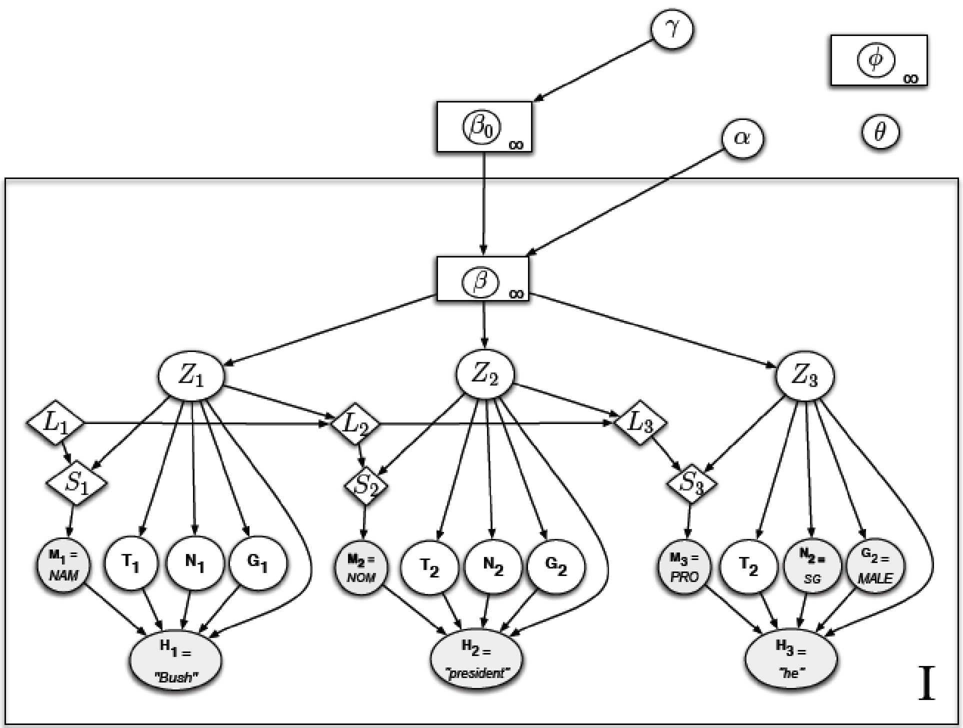
Cross document coreference

- Sharing data across documents is desirable, allowing for information about the properties of entities to be pooled across all documents
- This can easily be accomplished with a hierarchical Dirichlet process for entity selection
  – Assume the pool of entities is global, with global mixing weights β_0 drawn from a DP prior with parameter
  – Each document draws its own distribution β_i from a DP centered on β_0

Coreference Resolution Models

Cross document coreference

- **The graphical model for this approach:**



- **Results improved to an F1 score of 72.5**

Experiments

MUC-6

Dataset	Num Docs.	Prec.	Recall	F_1
MUC-6	60	80.8	52.8	63.9
+DRYRUN-TRAIN	251	79.1	59.7	68.0
+ENGLISH-NWIRE	381	80.4	62.4	70.3

- **As this is an unsupervised method, it is able to make use of unannotated data (with respect to coreferences)**
  - **The result labeled +DRYRUN-TRAIN displays this by including 191 unannotated documents from the MUC-6 dryrun training set**

Experiments
MUC-6

Dataset	Num Docs.	Prec.	Recall	F_1
MUC-6	60	80.8	52.8	63.9
+DRYRUN-TRAIN	251	79.1	59.7	68.0
+ENGLISH-NWIRE	381	80.4	62.4	70.3

- **Including data from a different corpora can even improve results**
  - **The result labeled +ENGLISH-NWIRE includes data from the ACE dataset, a different corpora from a different time period, and results still improve**

Experiments

MUC-6

Dataset	Num Docs.	Prec.	Recall	F_1
MUC-6	60	80.8	52.8	63.9
+DRYRUN-TRAIN	251	79.1	59.7	68.0
+ENGLISH-NWIRE	381	80.4	62.4	70.3

- **Recent supervised results gave an F_1 score of 73.4 on the MUC-6 test**
  - **Relatively close the best unsupervised result of 70.3**

Experiments

ACE 2004

Dataset	Prec.	Recall	F_1
ENGLISH-NWIRE	66.7	62.3	64.2
ENGLISH-BNEWS	63.2	61.3	62.3
CHINESE-NWIRE	71.6	63.3	67.2
CHINESE-BNEWS	71.2	61.8	66.2

- **Recent supervised results are 67.1 F_1 and 69.2 F_1 for the English NWIRE and BNEWS respectively**

Discussion

Errors

- The largest source of error is from coreferent proper and nominal mentions
  - George W. Bush, president of the US, visited Idaho
- This is unmodeled in the proposed system

Conclusion

- **A nonparametric Bayesian approach is proposed for entity coreference**
- **The proposed model accounts for the tendency to favor pronominal head words for coreferences in close proximity**
- **A hierarchical Dirichlet process is used to share data across documents**
- **Results comparable to supervised methods are achieved**