# Unsupervised Coreference Resolution in a Nonparametric Bayesian Model

**Aria Haghighi and Dan Klein**
Computer Science Division
UC Berkeley
{aria42, klein}@cs.berkeley.edu

## Abstract

We present an unsupervised, nonparametric Bayesian approach to coreference resolution which models both global entity identity across a corpus as well as the sequential anaphoric structure within each document. While most existing coreference work is driven by pairwise decisions, our model is fully generative, producing each mention from a combination of global entity properties and local attentional state. Despite being unsupervised, our system achieves a 70.3 MUC $F_1$ measure on the MUC-6 test set, broadly in the range of some recent supervised results.

## 1 Introduction

Referring to an entity in natural language can broadly be decomposed into two processes. First, speakers directly introduce new entities into discourse, entities which may be shared across discourses. This initial reference is typically accomplished with proper or nominal expressions. Second, speakers refer back to entities already introduced. This anaphoric reference is canonically, though of course not always, accomplished with pronouns, and is governed by linguistic and cognitive constraints. In this paper, we present a nonparametric generative model of a document corpus which naturally connects these two processes.

Most recent coreference resolution work has focused on the task of deciding which *mentions* (noun phrases) in a document are coreferent. The dominant approach is to decompose the task into a collection of pairwise coreference decisions. One then applies discriminative learning methods to pairs of mentions, using features which encode properties such as distance, syntactic environment, and so on (Soon et al., 2001; Ng and Cardie, 2002). Although such approaches have been successful, they have several liabilities. First, rich features require plentiful labeled data, which we do not have for coreference tasks in most domains and languages. Second, coreference is inherently a clustering or partitioning task. Naive pairwise methods can and do fail to produce coherent partitions. One classic solution is to make greedy left-to-right linkage decisions. Recent work has addressed this issue in more global ways. McCallum and Wellner (2004) use graph partioning in order to reconcile pairwise scores into a final coherent clustering. Nonetheless, all these systems crucially rely on pairwise models because cluster-level models are much harder to work with, combinatorially, in discriminative approaches.

Another thread of coreference work has focused on the problem of identifying matches between documents (Milch et al., 2005; Bhattacharya and Getoor, 2006; Daume and Marcu, 2005). These methods ignore the sequential anaphoric structure inside documents, but construct models of how and when entities are shared between them.[1] These models, as ours, are generative ones, since the focus is on cluster discovery and the data is generally unlabeled.

In this paper, we present a novel, fully generative, nonparametric Bayesian model of mentions in a document corpus. Our model captures both within- and cross-document coreference. At the top, a hierarchical Dirichlet process (Teh et al., 2006) cap-

[1] Milch et al. (2005) works with citations rather than discourses and does model the linear structure of the citations.



*Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 848–855,
Prague, Czech Republic, June 2007. 

tures cross-document entity (and parameter) sharing, while, at the bottom, a sequential model of salience captures within-document sequential structure. As a joint model of several kinds of discourse variables, it can be used to make predictions about either kind of coreference, though we focus experimentally on within-document measures. To the best of our ability to compare, our model achieves the best unsupervised coreference performance.

## 2 Experimental Setup

We adopt the terminology of the Automatic Context Extraction (ACE) task (NIST, 2004). For this paper, we assume that each document in a corpus consists of a set of *mentions*, typically noun phrases. Each mention is a *reference* to some entity in the domain of discourse. The coreference resolution task is to partition the mentions according to referent. Mentions can be divided into three categories, *proper* mentions (names), *nominal* mentions (descriptions), and *pronominal* mentions (pronouns).

In section 3, we present a sequence of increasingly enriched models, motivating each from shortcomings of the previous. As we go, we will indicate the performance of each model on data from ACE 2004 (NIST, 2004). In particular, we used as our development corpus the English translations of the Arabic and Chinese treebanks, comprising 95 documents and about 3,905 mentions. This data was used heavily for model design and hyperparameter selection. In section 5, we present final results for new test data from MUC-6 on which no tuning or development was performed. This test data will form our basis for comparison to previous work.

In all experiments, as is common, we will assume that we have been given as part of our input the true mention boundaries, the head word of each mention and the mention type (proper, nominal, or pronominal). For the ACE data sets, the head and mention type are given as part of the mention annotation. For the MUC data, the head was crudely chosen to be the rightmost mention token, and the mention type was automatically detected. We will not assume any other information to be present in the data beyond the text itself. In particular, unlike much related work, we do not assume gold named entity recognition (NER) labels; indeed we do not assume observed NER labels or POS tags at all. Our primary performance metric will be the MUC $F_1$ measure (Vilain et al., 1995), commonly used to evaluate coreference systems on a within-document basis. Since our system relies on sampling, all results are averaged over five random runs.

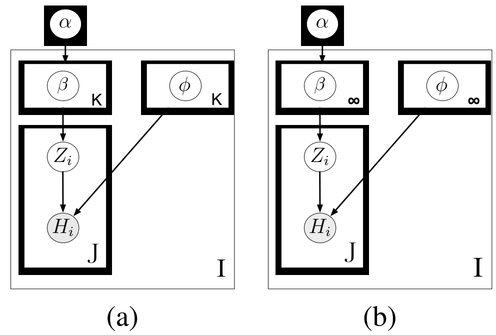



Figure 1: Graphical model depiction of document level entity models described in sections 3.1 and 3.2 respectively. The shaded nodes indicate observed variables.

## 3 Coreference Resolution Models

In this section, we present a sequence of generative coreference resolution models for document corpora. All are essentially mixture models, where the mixture components correspond to entities. As far as notation, we assume a collection of $I$ documents, each with $J_i$ mentions. We use random variables $Z$ to refer to (indices of) entities. We will use $\phi_z$ to denote the parameters for an entity $z$, and $\boldsymbol{\phi}$ to refer to the concatenation of all such $\phi_z$. $X$ will refer somewhat loosely to the collection of variables associated with a mention in our model (such as the head or gender). We will be explicit about $X$ and $\phi_z$ shortly.

Our goal will be to find the setting of the entity indices which maximize the posterior probability:

$$\begin{aligned}\mathbf{Z}^* &= \arg\max_{\mathbf{Z}} P(\mathbf{Z}|\mathbf{X}) = \arg\max_{\mathbf{Z}} P(\mathbf{Z}, \mathbf{X}) \\ &= \arg\max_{\mathbf{Z}} \int P(\mathbf{Z}, \mathbf{X}, \boldsymbol{\phi})\, dP(\boldsymbol{\phi})\end{aligned}$$

where $\mathbf{Z}, \mathbf{X}$, and $\boldsymbol{\phi}$ denote all the entity indices, observed values, and parameters of the model. Note that we take a Bayesian approach in which all parameters are integrated out (or sampled). The inference task is thus primarily a search problem over the index labels $\mathbf{Z}$.

**(a)** 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.

**(b)** 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.

**(c)** 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$_5$ will be situated in Rudong$_6$, Jiangsu$_7$, has an annual generation capacity of 2.4 million kilowatts.

Figure 2: Example output from various models. The output from (a) is from the infinite mixture model of section 3.2. It incorrectly labels both boxed cases of anaphora. The output from (b) uses the pronoun head model of section 3.3. It correctly labels the first case of anaphora but incorrectly labels the second pronominal as being coreferent with the dominant document entity *The Weir Group*. This error is fixed by adding the salience feature component from section 3.4 as can be seen in (c).

### 3.1 A Finite Mixture Model

Our first, overly simplistic, corpus model is the standard finite mixture of multinomials shown in figure 1(a). In this model, each document is independent save for some global hyperparameters. Inside each document, there is a finite mixture model with a fixed number $K$ of components. The distribution $\beta$ over components (entities) is a draw from a symmetric Dirichlet distribution with concentration $\alpha$. For each mention in the document, we choose a component (an entity index) $z$ from $\beta$. Entity $z$ is then associated with a multinomial emission distribution over head words with parameters $\phi^h_Z$, which are drawn from a symmetric Dirichlet over possible mention heads with concentration $\lambda_H$.[2] Note that here the $X$ for a mention consists only of the mention head $H$.

As we enrich our models, we simultaneously develop an accompanying Gibbs sampling procedure to obtain samples from $P(\mathbf{Z}|\mathbf{X})$.[3] For now, all heads $H$ are observed and all parameters ($\beta$ and $\boldsymbol{\phi}$) can be integrated out analytically: for details see Teh et al. (2006). The only sampling is for the values of $Z_{i,j}$, the entity index of mention $j$ in document $i$. The relevant conditional distribution is:[4]

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

where $H_{i,j}$ is the head of mention $j$ in document $i$. Expanding each term, we have the contribution of the prior:

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

where $n_z$ is the number of elements of $\mathbf{Z}^{-i,j}$ with entity index $z$. Similarly we have for the contribution of the emissions:

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

where $n_{h,z}$ is the number of times we have seen head $h$ associated with entity index $z$ in $(\mathbf{Z}, \mathbf{H}^{-i,j})$.

### 3.2 An Infinite Mixture Model

A clear drawback of the finite mixture model is the requirement that we specify a priori a number of entities $K$ for a document. We would like our model to select $K$ in an effective, principled way. A mechanism for doing so is to replace the finite Dirichlet prior on $\beta$ with the non-parametric Dirichlet process (DP) prior (Ferguson, 1973).[5] Doing so gives the model in figure 1(b). Note that we now list an infinite number of mixture components in this model since there can be an unbounded number of entities. Rather than a finite $\beta$ with a symmetric Dirichlet distribution, in which draws tend to have balanced clusters, we now have an infinite $\beta$. However, most draws will have weights which decay exponentially quickly in the prior (though not necessarily in the posterior). Therefore, there is a natural penalty for each cluster which is actually used.

With $\mathbf{Z}$ observed during sampling, we can integrate out $\beta$ and calculate $P(Z_{i,j}|\mathbf{Z}^{-i,j})$ analytically, using the Chinese restaurant process representation:

$$P(Z_{i,j} = z|\mathbf{Z}^{-i,j}) \propto \begin{cases} \alpha, & \text{if } z = z_{new} \\ n_z, & \text{otherwise} \end{cases} \quad (1)$$

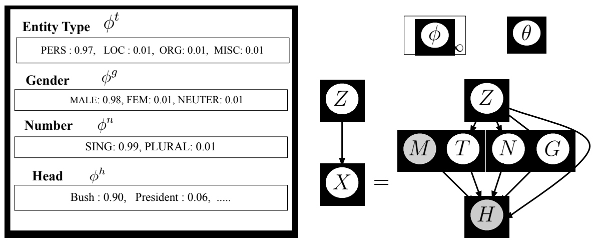
where $z_{new}$ is a new entity index not used in $\mathbf{Z}^{-i,j}$ and $n_z$ is the number of mentions that have entity index $z$. Aside from this change, sampling is identical

[2]In general, we will use a subscripted $\lambda$ to indicate concentration for finite Dirichlet distributions. Unless otherwise specified, $\lambda$ concentration parameters will be set to $e^{-4}$ and omitted from diagrams.

[3]One could use the EM algorithm with this model, but EM will not extend effectively to the subsequent models.

[4]Here, $\mathbf{Z}^{-i,j}$ denotes $\mathbf{Z} - \{Z_{i,j}\}$

[5]We do not give a detailed presentation of the Dirichlet process here, but see Teh et al. (2006) for a presentation.




(a) (b)

Figure 3: (a) An entity and its parameters. (b)The head model described in section 3.3. The shaded nodes indicate observed variables. The mention type determines which set of parents are used. The dependence of mention variable on entity parameters $\boldsymbol{\phi}$ and pronoun head model $\boldsymbol{\theta}$ is omitted.

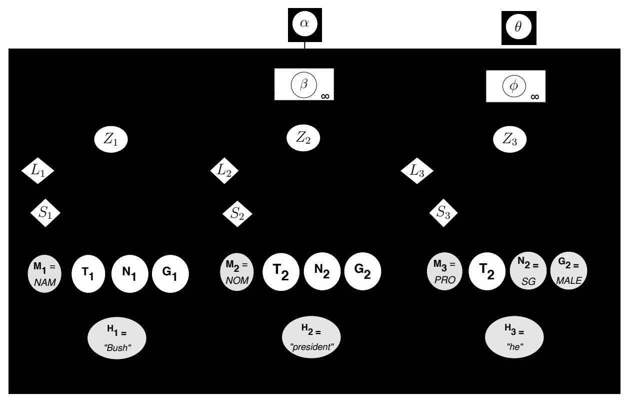



Figure 4: Coreference model at the document level with entity properties as well salience lists used for mention type distributions. The diamond nodes indicate deterministic functions. Shaded nodes indicate observed variables. Although it appears that each mention head node has many parents, for a given mention type, the mention head depends on only a small subset. Dependencies involving parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ are omitted.

to the finite mixture case, though with the number of clusters actually occupied in each sample drifting upwards or downwards.

This model yielded a 54.5 $F_1$ on our development data.[6] This model is, however, hopelessly crude, capturing nothing of the structure of coreference. Its largest empirical problem is that, unsurprisingly, pronoun mentions such as *he* are given their own clusters, not labeled as coreferent with any non-pronominal mention (see figure 2(a)).

### 3.3 Pronoun Head Model

While an entity-specific multinomial distribution over heads makes sense for proper, and some nominal, mention heads, it does not make sense to generate pronominal mentions this same way. I.e., all entities can be referred to by generic pronouns, the choice of which depends on entity properties such as gender, not the specific entity.

We therefore enrich an entity's parameters $\phi$ to contain not only a distribution over lexical heads $\phi^h$, but also distributions $(\phi^t, \phi^g, \phi^n)$ over properties, where $\phi^t$ parametrizes a distribution over entity types (PER, LOC, ORG, MISC), and $\phi^g$ for gender (MALE, FEMALE, NEUTER), and $\phi^n$ for number (SG, PL).[7] We assume each of these property distributions is drawn from a symmetric Dirichlet distribution with small concentration parameter in order to encourage a peaked posterior distribution.

[6] See section 4 for inference details.

[7] It might seem that entities should simply have, for example, a gender $g$ rather than a distribution over genders $\phi^g$. There are two reasons to adopt the softer approach. First, one can rationalize it in principle, for entities like cars or ships whose grammatical gender is not deterministic. However, the real reason is that inference is simplified. In any event, we found these property distributions to be highly determinized in the posterior.

Previously, when an entity $z$ generated a mention, it drew a head word from $\phi^h_z$. It now undergoes a more complex and structured process. It first draws an entity type $T$, a gender $G$, a number $N$ from the distributions $\phi^t$, $\phi^g$, and $\phi^n$, respectively. Once the properties are fetched, a mention type $M$ is chosen (*proper*, *nominal*, *pronoun*), according to a global multinomial (again with a symmetric Dirichlet prior and parameter $\lambda_M$). This corresponds to the (temporary) assumption that the speaker makes a random i.i.d. choice for the type of each mention.

Our head model will then generate a head, conditioning on the entity, its properties, and the mention type, as shown in figure 3(b). If $M$ is not a pronoun, the head is drawn directly from the entity head multinomial with parameters $\phi^h_z$. Otherwise, it is drawn based on a global pronoun head distribution, conditioning on the entity properties and parametrized by $\boldsymbol{\theta}$. Formally, it is given by:

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

Although we can observe the number and gender draws for some mentions, like personal pronouns, there are some for which properties aren't observed (e.g., *it*). Because the entity property draws are not (all) observed, we must now sample the unobserved ones as well as the entity indices $Z$. For instance, we could sample

| 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 |

Table 1: Posterior distribution of mention type given salience by bucketing entity activation rank. Pronouns are preferred for entities which have high salience and non-pronominal mentions are preferred for inactive entities.

$T_{i,j}$, the entity type of pronominal mention $j$ in document $i$, using, $P(T_{i,j}|\mathbf{Z}, \mathbf{N}, \mathbf{G}, \mathbf{H}, \mathbf{T}^{-i,j}) \propto P(T_{i,j}|\mathbf{Z})P(H_{i,j}|\mathbf{T}, \mathbf{N}, \mathbf{G}, \mathbf{H})$, where the posterior distributions on the right hand side are straightforward because the parameter priors are all finite Dirichlet. Sampling $G$ and $N$ are identical.

Of course we have prior knowledge about the relationship between entity type and pronoun head choice. For example, we expect that *he* is used for mentions with $T = \text{PERSON}$. In general, we assume that for each pronominal head we have a list of compatible entity types, which we encode via the prior on $\boldsymbol{\theta}$. We assume $\boldsymbol{\theta}$ is drawn from a Dirichlet distribution where each pronoun head is given a synthetic count of $(1 + \lambda_P)$ for each $(t, g, n)$ where $t$ is compatible with the pronoun and given $\lambda_P$ otherwise. So, while it will be possible in the posterior to use *he* to refer to a non-person, it will be biased towards being used with persons.

This model gives substantially improved predictions: 64.1 $F_1$ on our development data. As can be seen in figure 2(b), this model does correct the systematic problem of pronouns being considered their own entities. However, it still does not have a preference for associating pronominal references to entities which are in any way local.

### 3.4 Adding Salience

We would like our model to capture how mention types are generated for a given entity in a robust and somewhat language independent way. The choice of entities may reasonably be considered to be independent given the mixing weights $\beta$, but how we *realize* an entity is strongly dependent on context (Ge et al., 1998).

In order to capture this in our model, we enrich it as shown in figure 4. As we proceed through a document, generating entities and their mentions, we maintain a list of the active entities and their *saliences*, or activity scores. Every time an entity is mentioned, we increment its activity score by 1, and every time we move to generate the next mention, all activity scores decay by a constant factor of $0.5$. This gives rise to an ordered list of entity activations, $L$, where the rank of an entity decays exponentially as new mentions are generated. We call this list a *salience list*. Given a salience list, $L$, each possible entity $z$ has some rank on this list. We discretize these ranks into five buckets $S$: TOP (1), HIGH (2-3), MID (4-6), LOW (7+), and NONE. Given the entity choices $\mathbf{Z}$, both the list $L$ and buckets $S$ are deterministic (see figure 4). We assume that the mention type $M$ is conditioned on $S$ as shown in figure 4.

We note that correctly sampling an entity now requires that we incorporate terms for how a change will affect all future salience values. This changes our sampling equation for existing entities:

$$P(Z_{i,j} = z|\mathbf{Z}^{-i,j}) \propto n_z \prod_{j' \geq j} P(M_{i,j'}|S_{i,j'}, \mathbf{Z}) \quad (2)$$

where the product ranges over future mentions in the document and $S_{i,j'}$ is the value of future salience feature given the setting of all entities, including setting the current entity $Z_{i,j}$ to $z$. A similar equation holds for sampling a new entity. Note that, as discussed below, this full product can be truncated as an approximation.

This model gives a 71.5 $F_1$ on our development data. Table 1 shows the posterior distribution of the mention type given the salience feature. This model fixes many anaphora errors and in particular fixes the second anaphora error in figure 2(c).

### 3.5 Cross Document Coreference

One advantage of a fully generative approach is that we can allow entities to be shared between documents in a principled way, giving us the capacity to do cross-document coreference. Moreover, sharing across documents pools information about the properties of an entity across documents.

We can easily link entities across a corpus by assuming that the pool of entities is global, with global mixing weights $\beta_0$ drawn from a DP prior with concentration parameter $\gamma$. Each document uses

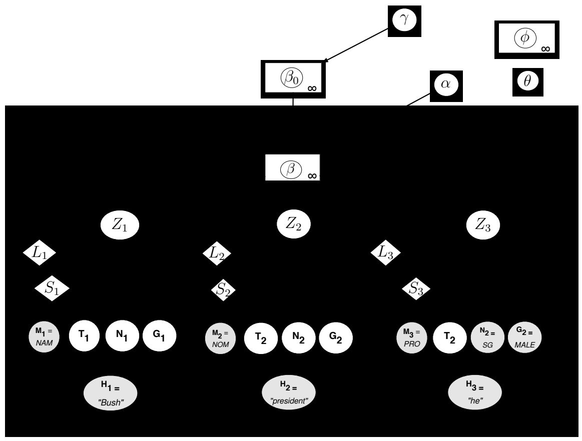



Figure 5: Graphical depiction of the HDP coreference model described in section 3.5. The dependencies between the global entity parameters $\phi$ and pronoun head parameters $\theta$ on the mention observations are not depicted.

the same global entities, but each has a document-specific distribution $\beta_i$ drawn from a DP centered on $\beta_0$ with concentration parameter $\alpha$. Up to the point where entities are chosen, this formulation follows the basic hierarchical Dirichlet process prior of Teh et al. (2006). Once the entities are chosen, our model for the realization of the mentions is as before. This model is depicted graphically in figure 5.

Although it is possible to integrate out $\beta_0$ as we did the individual $\beta_i$, we instead choose for efficiency and simplicity to sample the global mixture distribution $\beta_0$ from the posterior distribution $P(\beta_0|\mathbf{Z})$.[8] The mention generation terms in the model and sampler are unchanged.

In the full hierarchical model, our equation (1) for sampling entities, ignoring the salience component of section 3.4, becomes:

$$P(Z_{i,j} = z|\mathbf{Z}^{-i,j}, \beta_0) \propto \begin{cases} \alpha\beta_0^u, & \text{if } z = z_{new} \\ n_z + \alpha\beta_0^z, & \text{otherwise} \end{cases}$$

where $\beta_0^z$ is the probability of the entity $z$ under the sampled global entity distribution and $\beta_0^u$ is the unknown component mass of this distribution.

The HDP layer of sharing improves the model's predictions to 72.5 $F_1$ on our development data. We should emphasize that our evaluation is of course per-document and does not reflect cross-document coreference decisions, only the gains through cross-document sharing (see section 6.2).

## 4 Inference Details

Up until now, we've discussed Gibbs sampling, but we are not interested in sampling from the posterior $P(\mathbf{Z}|\mathbf{X})$, but in finding its mode. Instead of sampling directly from the posterior distribution, we instead sample entities proportionally to exponentiated entity posteriors. The exponent is given by $\exp \frac{ci}{k-1}$, where $i$ is the current round number (starting at $i = 0$), $c = 1.5$ and $k = 20$ is the total number of sampling epochs. This slowly raises the posterior exponent from $1.0$ to $e^c$. In our experiments, we found this procedure to outperform simulated annealing. We also found sampling the $T$, $G$, and $N$ variables to be particularly inefficient, so instead we maintain soft counts over each of these variables and use these in place of a hard sampling scheme. We also found that correctly accounting for the future impact of salience changes to be particularly inefficient. However, ignoring those terms entirely made negligible difference in final accuracy.[9]

## 5 Final Experiments

We present our final experiments using the full model developed in section 3. As in section 3, we use true mention boundaries and evaluate using the MUC $F_1$ measure (Vilain et al., 1995). All hyperparameters were tuned on the development set only. The document concentration parameter $\alpha$ was set by taking a constant proportion of the average number of *mentions* in a document across the corpus. This number was chosen to minimize the squared error between the number of proposed entities and true entities in a document. It was not tuned to maximize the $F_1$ measure. A coefficient of $0.4$ was chosen. The global concentration coefficient $\gamma$ was chosen to be a constant proportion of $\alpha M$, where $M$ is the number of *documents* in the corpus. We found $0.15$ to be a good value using the same least-square procedure. The values for these coefficients were not changed for the experiments on the test sets.

### 5.1 MUC-6

Our main evaluation is on the standard MUC-6 formal test set.[10] The standard experimental setup for

[8] We do not give the details here; see Teh et al. (2006) for details on how to implement this component of the sampler (called "direct assignment" in that reference).

[9] This corresponds to truncating equation (2) at $j' = j$.

[10] Since the MUC data is not annotated with mention types, we automatically detect this information in the same way as Luo

| 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 |

(a)

| 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 |

(b)

Table 2: Formal Results: Our system evaluated using the MUC model theoretic measure Vilain et al. (1995). The table in (a) is our performance on the thirty document MUC-6 formal test set with increasing amounts of training data. In all cases for the table, we are evaluating on the same thirty document test set which is included in our training set, since our system in unsupervised. The table in (b) is our performance on the ACE 2004 training sets.

this data is a 30/30 document train/test split. Training our system on all 60 documents of the training and test set (as this is in an unsupervised system, the unlabeled test documents are present at training time), but evaluating only on the test documents, gave 63.9 $F_1$ and is labeled MUC-6 in table 2(a).

One advantage of an unsupervised approach is that we can easily utilize more data when learning a model. We demonstrate the effectiveness of this fact by evaluating on the MUC-6 test documents with increasing amounts of unannotated training data. We first added the 191 documents from the MUC-6 dryrun training set (which were not part of the training data for official MUC-6 evaluation). This model gave 68.0 $F_1$ and is labeled +DRYRUN-TRAIN in table 2(a). We then added the ACE ENGLISH-NWIRE training data, which is from a different corpora than the MUC-6 test set and from a different time period. This model gave 70.3 $F_1$ and is labeled +ENGLISH-NWIRE in table 2(a).

Our results on this test set are surprisingly comparable to, though slightly lower than, some recent supervised systems. McCallum and Wellner (2004) report 73.4 $F_1$ on the formal MUC-6 test set, which is reasonably close to our best MUC-6 number of 70.3 $F_1$. McCallum and Wellner (2004) also report a much lower 91.6 $F_1$ on only proper nouns mentions. Our system achieves a 89.8 $F_1$ when evaluation is restricted to only proper mentions.[11] The closest comparable unsupervised system is Cardie and Wagstaff (1999) who use pairwise NP distances to cluster document mentions. They report a 53.6 $F_1$ on MUC6 when tuning distance metric weights to maximize $F_1$ on the development set.

et al. (2004). A mention is proper if it is annotated with NER information. It is a pronoun if the head is on the list of English pronouns. Otherwise, it is a nominal mention. Note we do not use the NER information for any purpose but determining whether the mention is proper.

[11]The best results we know on the MUC-6 test set using the standard setting are due to Luo et al. (2004) who report a 81.3 $F_1$ (much higher than others). However, it is not clear this is a comparable number, due to the apparent use of gold NER features, which provide a strong clue to coreference. Regardless, it is unsurprising that their system, which has many rich features, would outperform ours.

| HEAD | ENT TYPE | GENDER | NUMBER |
|---|---|---|---|
| *Bush: 1.0* | PERS | MALE | SG |
| *AP: 1.0* | ORG | NEUTER | PL |
| *viacom*: 0.64, *company*: 0.36 | ORG | NEUTER | SG |
| *teamsters*: 0.22, *union*: 0.78, | MISC | NEUTER | PL |

Table 3: Frequent entities occurring across documents along with head distribution and mode of property distributions.

### 5.2 ACE 2004

We also performed experiments on ACE 2004 data. Due to licensing restrictions, we did not have access to the ACE 2004 formal development and test sets, and so the results presented are on the training sets.

We report results on the newswire section (NWIRE in table 2b) and the broadcast news section (BNEWS in table 2b). These datasets include the *prenominal* mention type, which is not present in the MUC-6 data. We treated prenominals analogously to the treatment of proper and nominal mentions.

We also tested our system on the Chinese newswire and broadcast news sections of the ACE 2004 training sets. Our relatively higher performance on Chinese compared to English is perhaps due to the lack of prenominal mentions in the Chinese data, as well as the presence of fewer pronouns compared to English.

Our ACE results are difficult to compare exactly to previous work because we did not have access to the restricted formal test set. However, we can perform a rough comparison between our results on the training data (without coreference annotation) to supervised work which has used the same training data (with coreference annotation) and evaluated on the formal test set. Denis and Baldridge (2007) report 67.1 $F_1$ and 69.2 $F_1$ on the English NWIRE and BNEWS respectively using true mention boundaries. While our system underperforms the supervised systems, its accuracy is nonetheless promising.

# 6 Discussion

## 6.1 Error Analysis

The largest source of error in our system is between coreferent proper and nominal mentions. The most common examples of this kind of error are appositive usages e.g. *George W. Bush, president of the US, visited Idaho.* Another error of this sort can be seen in figure 2, where the *corporation* mention is not labeled coreferent with the *The Weir Group* mention. Examples such as these illustrate the regular (at least in newswire) phenomenon that nominal mentions are used with informative intent, even when the entity is salient and a pronoun could have been used unambiguously. This aspect of nominal mentions is entirely unmodeled in our system.

## 6.2 Global Coreference

Since we do not have labeled cross-document coreference data, we cannot evaluate our system's cross-document performance quantitatively. However, in addition to observing the within-document gains from sharing shown in section 3, we can manually inspect the most frequently occurring entities in our corpora. Table 3 shows some of the most frequently occurring entities across the English ACE NWIRE corpus. Note that *Bush* is the most frequent entity, though his (and others') nominal cluster *president* is mistakenly its own entity. Merging of proper and nominal clusters does occur as can be seen in table 3.

## 6.3 Unsupervised NER

We can use our model to for unsupervised NER tagging: for each proper mention, assign the mode of the generating entity's distribution over entity types. Note that in our model the only way an entity becomes associated with an entity type is by the pronouns used to refer to it.[12] If we evaluate our system as an unsupervised NER tagger for the proper mentions in the MUC-6 test set, it yields a per-label accuracy of 61.2% (on MUC labels). Although nowhere near the performance of state-of-the-art systems, this result beats a simple baseline of always guessing PERSON (the most common entity type), which yields 46.4%. This result is interesting given that the model was not developed for the purpose of inferring entity types whatsoever.

[12] Ge et al. (1998) exploit a similar idea to assign gender to proper mentions.

# 7 Conclusion

We have presented a novel, unsupervised approach to coreference resolution: global entities are shared across documents, the number of entities is determined by the model, and mentions are generated by a sequential salience model and a model of pronoun-entity association. Although our system does not perform quite as well as state-of-the-art supervised systems, its performance is in the same general range, despite the system being unsupervised.

# References


I. Bhattacharya and L. Getoor. 2006. A latent dirichlet model for unsupervised entity resolution. *SIAM conference on data mining*.

Claire Cardie and Kiri Wagstaff. 1999. Noun phrase coreference as clustering. *EMNLP*.

Hal Daume and Daniel Marcu. 2005. A Bayesian model for supervised clustering with the Dirichlet process prior. *JMLR*.

Pascal Denis and Jason Baldridge. 2007. Global, joint determination of anaphoricity and coreference resolution using integer programming. *HLT-NAACL*.

Thomas Ferguson. 1973. A bayesian analysis of some non-parametric problems. *Annals of Statistics*.

Niyu Ge, John Hale, and Eugene Charniak. 1998. A statistical approach to anaphora resolution. *Sixth Workshop on Very Large Corpora*.

Xiaoqiang Luo, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla, and Salim Roukos. 2004. A mention-synchronous coreference resolution algorithm based on the bell tree. *ACL*.

Andrew McCallum and Ben Wellner. 2004. Conditional models of identity uncertainty with application to noun coreference. *NIPS*.

Brian Milch, Bhaskara Marthi, Stuart Russell, David Sontag, Daniel L. Ong, and Andrey Kolobov. 2005. Blog: Probabilistic models with unknown objects. *IJCAI*.

Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. *ACL*.

NIST. 2004. The ACE evaluation plan.

W. Soon, H. Ng, and D. Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*.

Yee Whye Teh, Michael Jordan, Matthew Beal, and David Blei. 2006. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. *MUC-6*.