Unsupervised Deduplication using Cross-Field Dependencies

Robert Hall
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
rhall@cs.umass.edu

Charles Sutton[*]
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
casutton@cs.umass.edu

Andrew McCallum
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
mccallum@cs.umass.edu

ABSTRACT

Recent work in deduplication has shown that collective deduplication of different attribute types can improve performance. But although these techniques *cluster* the attributes collectively, they do not *model* them collectively. For example, in citations in the research literature, canonical venue strings and title strings are dependent—because venues tend to focus on a few research areas—but this dependence is not modeled by current unsupervised techniques. We call this dependence between fields in a record a *cross-field dependence*. In this paper, we present an unsupervised generative model for the deduplication problem that explicitly models cross-field dependence. Our model uses a single set of latent variables to control two disparate clustering models: a Dirichlet-multinomial model over titles, and a non-exchangeable string-edit model over venues. We show that modeling cross-field dependence yields a substantial improvement in performance—a 58% reduction in error over a standard Dirichlet process mixture.

Categories and Subject Descriptors

H.2.8 [**Information Systems**]: Database Applications—*data mining*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*clustering*

1. INTRODUCTION

Deduplication is an important and difficult preprocessing step in knowledge discovery. For example, consider the venue portion of citations in the bibliographies of research papers. A single venue can be named by dissimilar strings—for example, *AAAI* and *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, and other variants caused by typographical errors. Alternatively, different entities can be denoted by identical strings—for example, *ISWC* is a commonly-used abbreviation both for the *International Semantic Web Conference*, and for the *International Symposium on Wearable Computers*. The deduplication problem is to group a set of these noisy strings, which are called *mentions*, by which underlying *entity* they refer to. In addition, entities may have attributes (also called fields), such as the title and venue of a research paper, that can be used to improve deduplication. Given clean venue data, one can imagine computing many interesting bibliographic measures, such as which venues are most concentrated around a small set of authors, which venues adopt new terminology most frequently, and so on. But such measures will always be suspect unless the deduplication problem is solved well.

Recent work has shown that collective inference—in which many deduplication decisions are made simultaneously and an optimal set of decisions chosen globally—can significantly improve performance over approaches in which individual decisions are made independently. This has been demonstrated in such areas as collective clustering over mentions [10, 15, 3], collective extraction and deduplication [18], and collective deduplication of different attribute types [13, 7, 14, 15]. But although some methods *compute* clusters collectively, there is an important sense in which they do not *model* clusters collectively. Current generative models focus on modeling the manner in which a canonical attribute, such as a paper's true title, is distorted in a noisy observation, such as a citation in a later paper. Crucially, however, current models do not incorporate the fact that different attribute types are dependent. For example, research venues tend to focus on specific research areas, and those areas are reflected in the titles of the papers that they publish. We call this a *cross-field dependence*, because the values of different fields are probabilistically dependent.

In this paper, we demonstrate the benefits of modeling cross field dependencies in the task of deduplicating research paper venues. We show that modeling the dependence between venue strings and paper titles yields a significant increase in deduplication performance from unlabeled data. In particular, we present a Dirichlet process mixture model that uses a single set of mixture components to combine two disparate clustering models: a Dirichlet-multinomial mixture for the titles, and a non-conjugate string-edit distortion model for the venues. In this way, each venue has a characteristic distribution not only of venue strings, but also of title strings. This encourages merging venue clusters with similar title distributions, even if their distributions over venue strings are somewhat different.

[*]Current Address: Computer Science Division, University of California, Berkeley, CA 94720

KDD'08, August 24–27, 2008, Las Vegas, Nevada, USA.

The two different distortion models for titles and venues reflect the fact that we expect different kinds of noise in both types of fields. For observed title strings, we expect that many citations will list the canonical title, while others have small, weakly correlated typographical errors. For observed venue strings, on the other hand, the edit distance between coreferent strings is much larger: several words may be added or deleted. Furthermore, while it is reasonable to model typos in title strings as independent, in venue strings often several variants appear equally commonly.

The performance of these models depends crucially on approximate inference. We describe inference based on Markov chain Monte Carlo (MCMC) methods, which are complicated by the nature of the string-edit distortion model that we use for venue strings. In addition, we compare an MCMC sampler based on Gibbs sampling to a recently-proposed split-merge sampler [8], and find that the split-merge sampler performs significantly better.

We evaluate our models on real-world citation data that is specifically designed to be hard for this task. A model that incorporates cross-field dependence performs substantially better than a standard DP mixture, yielding a 58% reduction in error over a standard DP mixture, and a 48% reduction in error over a reasonable heuristic baseline.

2. MODEL

In this section, we describe our model of venue and title mentions. Each mention m contains a paper's title $\mathbf{t}_m$ and venue $\mathbf{v}_m$, such as from the bibliography of a citing research paper. The task is to determine which venue strings refer to the same underlying venue. The data set as a whole is a set of mentions $\{(\mathbf{v}_m, \mathbf{t}_m)\}_{m=1}^{M}$. Each venue mention $\mathbf{v}_m$ is a sequence of words $(v_{m1}, v_{m2}, \dots v_{m,N(v_m)})$ and each title mention a sequence of words $(t_{m1}, t_{m2}, \dots t_{m,N(t_m)})$. This is an unsupervised problem, in the sense that we are not provided with training mentions which are known to be either duplicates or not.

We describe our model by incrementally augmenting a simple finite mixture model. All of our models are mixture models in which each mixture component is interpreted as an underlying venue. First, we describe a finite mixture model of the venue mentions only, using a string-edit model customized for this task (Section 2.1). Second, we modify this model to allow an infinite number of components by using a Dirichlet process mixture (Section 2.2). Then, we augment this model with title mentions that are drawn from a per-venue unigram model (Section 2.3), modeling a type of cross-field dependence. Finally, we describe a venue-title model in which the titles are drawn from a latent Dirichlet allocation (LDA) model [4] (Section 2.4).

2.1 Finite Mixture Model over Venues

First we describe a finite mixture model, where the number of venues C is chosen in advance. The main idea is that each true entity is modeled by a mixture component, where each component generates canonical strings and observed venue strings via a string-edit distortion model. More specifically, the mixture proportions β are sampled from a symmetric Dirichlet with concentration parameter α. Each cluster $c \in \{1 \dots C\}$ is associated with a canonical venue string $\mathbf{x}_c$, which is sampled from a unigram language model with uniform emission probabilities. For each mention, the model selects a venue assignment c_m (which is an index into the set of venues) according to the venue proportions β.

Finally, we generate the venue mentions $\mathbf{v}_m = v_{m,0} \cdots v_{m,a}$ for each mention of each cluster c. The venue mentions are generated by distorting the venue's canonical string $\mathbf{x}_c = x_{c,0} \cdots x_{c,b}$ by an HMM string-edit model denoted $p(\mathbf{v}_m|\mathbf{x}_c)$. Note that this model conditions on the canonical string of the cluster. The HMM string-edit model has three edit operations: *substitute* in which a token of the canonical string is replaced by a token of the observed string, *insert* which generates a token of $\mathbf{v}_m$, and *delete* which removes a token of $\mathbf{x}_c$. Each edit operation corresponds to a single state of the HMM. We choose transition probabilities $p(s_i = \text{insert}|s_{i-1}) = p(s_i = \text{delete}|s_{i-1}) = 0.3$ and $p(s_i = \text{substitute}|s_{i-1}) = 0.4$, so the model disfavors words that occur in only one of the two strings.

Now we describe the emission distributions for each state, that is, the distribution over the tokens that each state inserts into the observed string. The delete state deterministically emits the empty token. The insert state has uniform emission probability over the vocabulary of venue tokens. Finally, the substitute state has a custom emission distribution, to model the fact that acronyms are common in venue strings. If $v_{m,j}$ is the current venue token and $x_{c,i}$ the current canonical token, then the emission distribution is

$$p(v_{m,j}|s_{i,j} = \text{substitute}, x_{c,i}) = \begin{cases} a(x_{c,i})^{-1} & \text{if } x_{c,i} \text{ starts with } v_{m,j} \\ l(v_{m,j})^{-1} & \text{if } v_{m,j} \text{ starts with } x_{c,i}, \end{cases} \qquad (1)$$

where $a(w)$ is the number of words in the vocabulary that are prefixes of w, and $l(w)$ is the number of words for which w is a prefix. Calculating the probability of a canonical string generating a venue mention requires summing over all sequences of edit states, which can be done efficiently using the the forward algorithm for HMMs.

In summary, the finite-mixture model is

$$\begin{aligned} \beta &\sim \text{Dirichlet}(\alpha, \mathbf{u}) \\ \mathbf{x}_c &\sim \text{Unigram}(\mathbf{u}) \\ c_m \,|\, \beta &\sim \beta \\ \mathbf{v}_m \,|\, \mathbf{x}, c_m &\sim \text{StringEdit}(\mathbf{x}_{c_m}), \end{aligned} \qquad (2)$$

In this notation, $\mathbf{u}$ is a uniform vector over the vocabulary, that is, it is a vector of length V with each element $u_i = 1/V$. Here we write the Dirichlet distribution with two parameters, a base measure which we choose to be $\mathbf{u}$, and a scalar concentration parameter α, which we choose as described in Section 5.3. The notation Unigram denotes a unigram language model. Essentially, the prior on canonical strings $\mathbf{x}_c$ is a uniform distribution over strings. This an improper prior, but the posterior is still well-normalized.

The graphical model for the finite mixture is shown in Figure 1. This model requires choosing the number of venues in advance, which is undesirable. In the next section, we remove this requirement.

2.2 Dirichlet Process Mixture over Venues

The finite mixture model requires specifying a number of clusters a priori, which is unrealistic. For this reason, recent work in unsupervised coreference [5, 9, 2] has focused on nonparametric models, and in particular the infinite limit of (2), which is the Dirichlet process (DP) mixture. A DP mixture is attractive for two reasons: first, the number of

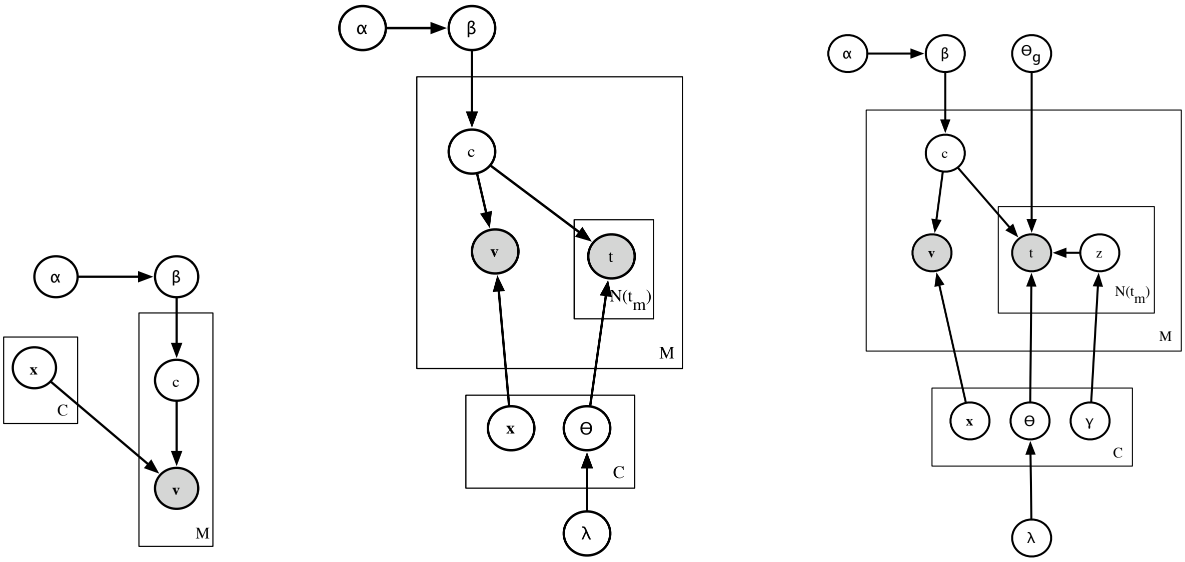

Figure 1: At left, finite mixture model over venues. The DP mixture is the infinite limit of this model. In middle, venue-title DP mixture (DPVT). At right, venue-title DP mixture with special words (DPVSW).

components of the mixture can be inferred from data; and second, samples from the induced cluster identities display a rich-get-richer property that is natural in many domains. For a review of modeling and inference using DPs and DP mixtures, see Teh et al. [17].

Probably the most intuitive way to understand the resulting distribution over venue strings is the Chinese restaurant process representation. This is a metaphor for describing how samples are drawn from a DP mixture. Here we imagine that each venue mention corresponds to a customer at a restaurant that contains an infinite number of tables. Each table c represents a cluster of mentions (in other words, a true venue), and associated with each table is a canonical string $\mathbf{x}_c$ (the "dish" served at that table).

Suppose we have a set of mentions that have already been generated at tables $1 \dots C$, where each venue c contains N_c mentions. To generate a new mention, first we generate which table the mention sits at, that is, which true venue is assigned to the mention. This table c_m is selected from the following distribution:

$$p(c_m = c | c_{1 \dots m-1}) \propto \begin{cases} N_c & \text{if } c \text{ is an existing table} \\ \alpha & \text{if } c = C+1, \text{ i.e., a new table} \end{cases} \tag{3}$$

The parameter $\alpha > 0$ is a parameter of the Dirichlet process, and affects how likely the model is to generate new tables. If the mention does sit at a new table, then we generate a canonical string for the new venue, from a uniform distribution over strings. Once that the mention has chosen a table c_m, it generates an observed venue string $\mathbf{v}_m$ from the canonical string $\mathbf{x}_{c_m}$ at that table. The observed string is generated from the string-edit model of the last section. This completes the description of the model.

It can be shown that this model is actually the infinite limit of the finite mixture model, as the number of mixture elements goes to infinity. This is the infinite limit of the graphical model shown in Figure 1 (left).

To describe this more formally, consider a random variable $\mathbf{c}$ that ranges over partitions of the integers $\{1 \dots M\}$. The Chinese restaurant process defines a distribution over $\mathbf{c}$—to see this, imagine labeling the customers $1 \dots M$. Denote this distribution as CRP(α). Using this representation we can describe the DP mixture model as

$$\begin{aligned} \mathbf{c} &\sim \text{CRP}(\alpha) \\ \mathbf{x}_c &\sim \text{Unigram}(\mathbf{1}) \qquad \text{for } c \text{ in } 1 \dots |\mathbf{c}| \\ \mathbf{v}_m \,|\, \mathbf{x}, c_m &\sim \text{StringEdit}(\mathbf{x}_{c_m}) \end{aligned} \tag{4}$$

In the above, c_m refers to the index of the set in the partition $\mathbf{c}$ that contains the integer m.

The advantage of the DP mixture is that it automatically determines the number of clusters from the data. This statement might seem disingenuous, because perhaps we have swapped the problem of selecting the number of clusters for the problem of selecting the parameter α. In practice, however, this is typically not an issue, because usually the number of clusters selected by the model is not sensitive to α, which indeed is the case in our setting (see Section 5.3).

2.3 DP Mixture over Venues and Titles

In the previous section, we presented an infinite mixture model over venue strings. But such a model can be improved dramatically by also considering information from paper titles, and demonstrating this is a key contribution of our work. In this section, we present a model that does this. We will call it the *DP venue-title model (DPVT)*.

The DPVT model jointly clusters venues and titles using a single set of latent variables that control both a string-edit model for the venues and a Dirichlet-multinomial distribution for the titles. Each venue c generates a distribution θ_c over title words, in other words, a probability vector with one element for each word in the vocabulary. Every mention m now generates all of its title word t_{mi} by a discrete distribution with parameters θ_{c_m}.

This model contains all of the factors in the venue-only model, and in addition:

$$\theta_c \sim \text{Dirichlet}(\lambda, \mathbf{u})$$
$$t_{mi} \,|\, c_m, \{\theta_c\} \sim \theta_{c_m}$$

As before, $\mathbf{u}$ is a uniform probability vector over the vocabulary, and λ is the concentration parameter of the Dirichlet.

To see how this model incorporates cross-field dependence between venues and titles, consider the graphical model in the middle of Figure 1. Note that although the canonical venues $\mathbf{x}_c$ and title distributions θ_c are independent in the prior, they are dependent in the posterior, because they are coupled by the c_m variables once the mentions are observed.

2.4 DP Mixture with Venues and Special-Word Title Model

In the first venue-title model, every venue has a multinomial distribution over title words. But we may hope to achieve better performance by using a more flexible model over title strings, for example, one that separates out common words from venue-specific words. Also, such a model allows reporting title words that are strongly associated with particular venues, which may be of interest in itself.

For this reason, in this section we describe an alternative title model in which the titles are generated by latent Dirichlet allocation [4]. One topic is dedicated solely to each venue, and a single "general English" topic is shared across all venues. This is a simple version of the special words model of Chemudugunta, Smyth, and Steyvers [6], so call this model the *DP mixture with venues and special-word title model (DPVSW)*.

This model includes all of the factors of the venue-only DP model, and in addition:

$$\begin{aligned}
\theta_g &\sim \text{Dirichlet}(\lambda_0, \mathbf{u}) \\
\theta_c &\sim \text{Dirichlet}(\lambda, \mathbf{u}) \qquad (5) \\
\gamma_c &\sim \text{Beta}(1,1) \\
z_{mi} \,|\, c, \gamma_c &\sim \text{Bernoulli}(\gamma_{c_m}) \\
t_{mi} \,|\, \theta_g, \theta_c, z_{mi} &\sim \begin{cases} \theta_g & \text{if } z_{mi} = 0 \\ \theta_c & \text{if } z_{mi} = 1 \end{cases}
\end{aligned}$$

Here θ_g is a single corpus-wide distribution over title words, while as before each θ_c is a venue-specific distribution over title words. For each word i of title mention m, this model includes an indicator variable z_{mi} which is 0 if word i was generated from the global distribution θ_g, and 1 if it was generated from the venue-specific distribution θ_c. Each γ_c controls how often venue c uses its venue-specific title distribution as opposed to the general distribution. The graphical representation of this model is shown in Figure 1.

3. INFERENCE

In this section we discuss two Markov chain Monte Carlo (MCMC) samplers for our models. Given a set of observed venue mentions $\mathbf{v} = \{\mathbf{v}_1 \ldots \mathbf{v}_M\}$, our concern will be to sample from the resulting posterior distribution $p(\{\mathbf{x}_c\}, \{c_m\}|\mathbf{v})$ over venue assignments c_m for each mention and canonical strings $\mathbf{x}_c$ for each venue. A common choice is to resample each c_m using a Gibbs sampler, but this is not straightforward in our models because the distribution $p(\mathbf{v}_m \,|\, \mathbf{x}_{c_m})$ is not conjugate to the prior over canonical strings $p(\mathbf{x}_{c_m})$. In many applications of the DP mixture model, the analogs of those two distributions are conjugate, and in those cases inference is simplified considerably.

3.1 DP Venue Model

First we describe the samplers for the basic DP venue model. The state of the sampler is the set of all cluster indices $\mathbf{c} = \{c_m\}$ for each mention m and of canonical strings $\mathbf{x} = \{\mathbf{x}_c\}$ for each cluster c from $1 \ldots C$. The main idea is to use a block Gibbs sampler, alternating between sampling the cluster identities and the canonical strings.

The first part of the outer block Gibbs sampler is to sample the cluster identities. For this we use a Metropolis-Hastings step. We consider two different proposals: one which is only a slight modification of the Gibbs sampler (so we call it *almost-Gibbs*) and another based on a split-merge proposal. The Gibbs proposal is a modification to Neal [12]. For every mention $m \in \{1 \ldots M\}$, we propose a new cluster c^*_m from the distribution:

$$p(c^*_m|c_{-m}, \mathbf{v}, \mathbf{x}) \propto \begin{cases} p(c^*_m \,|\, c_{-m}) p(\mathbf{v}_m|\mathbf{x}_{c^*_m}) \\ \qquad\qquad \text{if } c^*_m \text{ is an existing cluster} \\ p(c^*_m|c_{-m}) p(\mathbf{v}_m|\mathbf{x}_m = \mathbf{v}_m) \\ \qquad\qquad \text{if } c^*_m \notin \{1 \ldots C\} \end{cases} \quad (6)$$

That is, if the proposed cluster is one that already exists, the proposal is proportional to the prior $p(c^*_m \,|\, c_{-m})$ times the probability that the canonical string $x_{c^*_m}$ would be distorted into the observed string $\mathbf{v}_m$ of the current mention. This is exactly the Gibbs proposal. If the proposed cluster is new, then the proposal is proportional to the prior times the probability that the observed string would be distorted into itself. This is the part that is different from the Gibbs proposal. Ideally, we would sample $\mathbf{x}_{c^*_m}$ in this case; Neal [12] suggests using the prior $p(\mathbf{x})$, but this would lead to a string that would hardly ever be close to $\mathbf{v}_m$. Here we are mainly interested in finding a high-probability configuration, which makes our choice seem reasonable.

The second proposal distribution that we use is the split-merge proposal of Dahl [8]. Here the basic idea is to pick two mentions m and m' at random. If the mentions are in different clusters, the proposal merges them. If the mentions are in the same cluster, the proposal splits that cluster into two, one containing m and one containing m', using a procedure similar to sequential importance sampling.

The second part of the outer block Gibbs sampler is to sample the canonical strings $\mathbf{x}_c$. Here we use a Gibbs step, but with the restriction that $\mathbf{x}_c$ must be identical to one of the observed venue strings in the cluster. This restriction is a slight abuse, but seems to work well in practice. More specifically, let k in $1 \ldots N_c$ index the mentions assigned to cluster c. Then the new canonical string $\mathbf{x}^*_c$ is sampled from the distribution

$$p(\mathbf{x}^*_c \,|\, \mathbf{c}, \mathbf{v}) \propto \prod_{k=1}^{N_c} p(\mathbf{v}_k|\mathbf{x}^*_c) 1\{\mathbf{x}^*_c = \mathbf{v}_k \text{ for some } k \text{ in } 1 \ldots N_c\}, \quad (7)$$

where $1\{\cdots\}$ is an indicator function that enforces the restriction that canonical strings be somewhere observed.

All the MCMC methods described here require a choice of starting configuration, which can greatly impact their effectiveness. We initialize the samplers by placing each venue mention in its own cluster.

3.2 Venue-title model

For the venue-title model, we use the same samplers as above, except that when we propose a new cluster assignment c_m we must take into account the distribution over title words, integrating out the mean vector θ_c. Denote by $p(\mathbf{t}_m|c^*_m, \mathbf{t}_{-m})$ the probability of the title mention $\mathbf{t}_m$ being generated by the proposed cluster, conditioned on the titles of the other mentions in that cluster, and integrating out the distribution θ_c over title words. This probability can be computed using a Polya urn scheme:

$$p(\mathbf{t}_m|c^*_m, \mathbf{t}_{-m}) = \prod_{i=1}^{N(t_m)} p(t_{mi}|t_{m1}, \ldots, t_{m,i-1}) \tag{8}$$

$$= \prod_{i=1}^{N(t_m)} \frac{N_{\{t_{mi}=t_{mj};j<i\}} + N_{\{t_{-m,c}=t_{mi}\}} + \lambda}{(i-1) + \sum_{m' \in c_m \setminus m} |\mathbf{t}_{m'}| + V\lambda}, \tag{9}$$

where $N_{\{t_{mi}=t_{mj};j<i\}}$ is the number of words in $\mathbf{t}_m$ that precede word i and are identical to it; and $N_{\{t_{-m,c}=t_{mi}\}}$ is the number of occurrences of the token t_{mi} in the other titles $\mathbf{t}_{-m,c}$ in cluster c; and V the vocabulary size. Finally, recall that $N(t_m)$ is the number of words in the title mention t_m.

Now in the almost-Gibbs proposal, we sample a new cluster assignment c^*_m from the distribution

$$p(c^*_m|c_{-m}, \mathbf{v}) \propto \begin{cases} p(c^*_m \mid c_{-m})p(\mathbf{v}_m|c^*_m)p(\mathbf{t}_m|c^*_m, \mathbf{t}_{-m}) \\ \qquad\qquad \text{if } c^*_m = c_j \text{ for some } j \neq m \\ p(c^*_m|c_{-m})p(\mathbf{v}_m|c^*_m, \mathbf{x}_m = \mathbf{v}_m)p(\mathbf{t}_m|c^*_m) \\ \qquad\qquad \text{if } c^*_m \notin \{1 \ldots C\}, \end{cases} \tag{10}$$

The difference from the venue-only version is the inclusion of the term $p(\mathbf{t}_m|c^*_m)$.

3.3 Venue-Special words title model

Finally, we describe the modifications to the sampler for the venue-special words title model. For this model, we add to the state of the sampler the indicator variables $\mathbf{z}_m = \{z_{mi}\}$, which for each title word i in mention m, indicate whether the word is to be sampled from the venue multinomial with mean θ_c or from the general multinomial with mean θ_g.

The venue-special words model requires two changes to the sampler for the venue-title model. First, we add a step to the outer block Gibbs sampler that resamples all of the $\mathbf{z}$ variables given the cluster assignments and canonical strings. As before, let t_{mi} be the title word of mention m in position i; $\theta_c(t_{mi})$ be the element of the title mean vector θ_c for the word t_{mi}; and $\theta_g(t_{mi})$ be the analogous quantity in the general English multinomial vector. Then, during the additional Gibbs step, each z_{mi} is sampled from the distribution

$$p(z_{mi} = 1|\mathbf{z}_{m,-i}, \mathbf{z}_{-m}, \mathbf{c}, \mathbf{x}) \propto \frac{\theta_{c_m}(t_{mi}) + 1}{\theta_{c_m}(t_{mi}) + \theta_g(t_{mi}) + 2}, \tag{11}$$

where $z_{mi} = 1$ indicates that t_{mi} is sampled from the venue-specific distribution.

The second change is that the proposal distribution in the cluster assignment step changes slightly, because now the distribution over titles depends on $\mathbf{z}$. The new proposal distribution is

$$p(c^*_m|c_{-m}, \mathbf{v}, \mathbf{z}) \propto \begin{cases} p(c^*_m \mid c_{-m})p(\mathbf{v}_m|c^*_m)p(\mathbf{t}_m|c^*_m, \mathbf{t}_{-m}, \mathbf{z}_m) \\ \qquad\qquad \text{if } c^*_m = c_j \text{ for existing cluster } j \\ p(c^*_m|c_{-m})p(\mathbf{v}_m|c^*_m, \mathbf{x}_m = \mathbf{v}_m)p(\mathbf{t}_m|c^*_m, \mathbf{z}_m) \\ \qquad\qquad \text{if } c^*_m \notin \{1 \ldots C\}. \end{cases} \tag{12}$$

Observe that this Gibbs sampling scheme depends crucially on the fact that it is reasonable to change the cluster identity c_m but without changing z_m. Such a move is in fact reasonable because the semantics of the $\mathbf{z}_m$ variables do not depend on the venue identity.

4. RELATED WORK

The literature on deduplication is extensive. (Ironically, deduplication is also known as coreference, record linkage, and identity uncertainty.) A growing body of work has shown that incorporating global information can improve coreference. For example, McCallum and Wellner [10] show that incorporating transitive closure improves performance over making pairwise coreference decisions independently. Culotta and McCallum [7] have applied similar models to venue coreference, finding that jointly modeling coreference of records and fields improves performance. Additionally, Singla and Domingos [14, 15] perform simultaneous coreference of authors, papers, and venues using conditional undirected models, and find a similar improvement. These models are all supervised, so our approaches have the advantage of not requiring labeled training data, although they can readily exploit labeled coreference data if it is available.

Several authors have used DP mixture models for deduplication. The general framework of using a per-cluster mixture model for coreference of research papers was introduced by Pasula et al. [13]. A more detailed description of a similar model is given by Milch [11]. These models generate the number of venues from a log normal distribution. A variant of this model which models the venue assignments with a hierarchical DP was reported by Carbonetto et al. [5], although they do not report a comparison with the log normal model. Similarly, Bhattacharya and Getoor [2] use a DP mixture model to perform deduplication of authors in research papers.

None of the models above, however, incorporate cross-field dependencies. For example, in the model of Carbonetto et al., every canonical paper has a true title and distribution over observed title strings, and every canonical author has a distribution over observed author strings. But the model does not represent that canonical authors tend to favor certain words in their canonical titles. Similarly, the Singla and Domingos models [15] do incorporate the constraint that if two paper mentions are identical, then so are the corresponding venue mentions. But they do not have weights that say if one title appears in a venue, then distinct titles with similar words are likely to also appear in that venue. The key contribution of our work is to explicitly model this cross-field dependence, showing that this leads to dramatically better performance on venue coreference.

Another related model is by Haghighi and Klein [9], which applies DP mixtures to noun-phrase corference, which is the problem of determining which noun phrases in a document refer to the same entity, such as *George W. Bush* and *he*. This work is in similar spirit to ours, in that it augments the basic DP mixture with additional variables tailored to a specific coreference task. However, the specifics of their model are very different, because they need to model notions such as that pronouns can only refer to entities of a particular gender, and that more salient entities in the discourse are referred to using different language than less salient ones.

5. EXPERIMENTS

In this section, we compare the performance of the various DP mixture models on citation data. We first obtain a list of automatically extracted citations from the Rexa database (`http://rexa.info`). The citations are first segmented automatically using a conditional random field documents into plaintext. This process is imperfect, so the fields contain extraction errors as well as the expected typographical errors. The data consists of the resulting were mapped onto venue-title pairs, and duplicate citations (those that were string identical in both fields) were collapsed.

We choose a dozen test venues, and assemble a corpus of about 180 citations per venue, for a total of 2190 citations[1]. Reflecting the coverage of a large-scale digital library, the venues cover a range of topics including: artificial intelligence, machine learning, computational physics, biology, the semantic web, and wearable computing. The venues are also specifically chosen to be hard for the coreference task; for example, several venue pairs are included that have string-identical abbreviations. After removal of stopwords and punctuation symbols, there are 262 unique venue strings. Also, in any mention that consists entirely of a string of capital letters, we treat each capital letter as a separate word. This allows the distortion model to more easily align acronyms with their full names.

We compare four models: (i) the DPV model, which is the Dirichlet process mixture only (Section 2.2), (ii) the DPVT model (Section 2.3), (iii) the DPVSW model (Section 2.4), and (iv) a baseline heuristic STR. In this heuristic, first remove stopwords such as “an”, “of”, and “proceedings”, then merge string-identical venues, and finally, merge all venue clusters that contain string-identical titles. For each of the generative models we performed 1000 iterations of block Gibbs sampling. To select the hyperparameters α and λ, we perform a parameter sweep on a separate small validation set.

We measure performance using the B^3 metric of Bagga and Baldwin [1]. For each mention m_i, let c_i be the set of predicted coreferent mentions, and t_i be the set of truly coreferent mentions. The precision for m_i is the number of correct mentions of entity i—that is, those that appear in both c_i and t_i—divided by the number of mentions in c_i. The recall is the number of correct mentions of entity i divided by the size of t_i. These are averaged over all mentions in the corpus to obtain a single pair of precision and recall numbers. F1 is the harmonic mean of precision and recall.

5.1 Comparison of Models

Coreference performance for each of the four systems is shown in Table 1. The baseline STR heuristic demonstrates the difficulty of performing coreference on this dataset: string identical mentions are not necessarily coreferent, and different strings often refer to the same venue. The best performance overall is obtained by the DPVT system, where we set the concentration over title unigram distributions to be $\lambda = 0.6$. This setting has the effect of favoring more peaked unigram distributions over title words, where the peaks correspond to words particular to that cluster. These results demonstrate a marked improvement in coreference performance by modeling the titles of the papers. F1 is increased from 63.8% to 84.9% by adding title modeling to the DP mixture, a 58% error reduction. The error reduction over the STR baseline is 48%.

Model	Precision	Recall	F1
DPVT	86.4±0.84	83.4±1.32	**84.9**±1.06
DPVSW	88.5±1.06	72.9±1.70	80.0±1.39
DPV	84.1±0.37	51.4±0.31	63.8±0.20
STR	88.9	56.5	69.9

Table 1: Percent B^3 venue coreference performance for the four systems. The smaller numbers indicate the standard deviation across five independent realizations of the Markov chain.

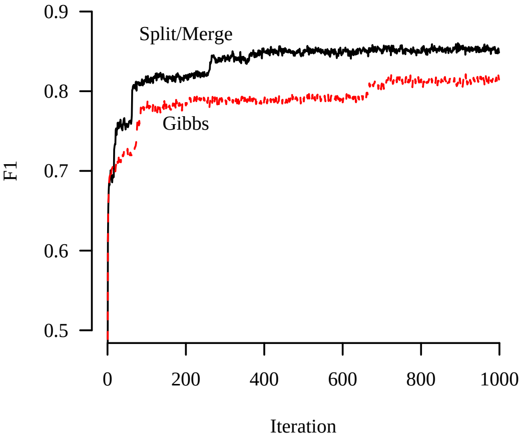

Figure 2: Comparison of Gibbs and split/merge samplers on the DPVT model.

The DPVSW model has slightly higher precision than the DVPT model, but at a high cost to recall. Also, the data set contains two venues which share the name ”ISWC” (for *International Semantic Web Conference* and *International Symposium on Wearable Computing*), which the DPVSW model is able to disambiguate more accurately, as shown in Table 3. The standard Dirchlet process mixture, on the other hand, will almost always merge identical venue strings. Shown in Table 2 are some example per-venue distributions over words that were generated by the DPVSW model. While common words such as “a” and “for” are highly weighted in these clusters, so are less frequent venue-specific words.

We test statistical significance using a stratified bootstrap sampler. Namely, we bootstrap a confidence interval for the $B^3 F1$ of all the methods by, for each true venue V with n_v mentions, we sample n_v new mentions from V uniformly with replacement. The performance difference between DPVT and STR is highly significant ($p < 0.01$).

5.2 Comparison of Sampling Algorithms

We also compare the split-merge and Gibbs proposal distributions, described in Section 3.1. Although in the infinite

[1] The dataset is available on the web at `http://www.cs.umass.edu/~rhall/rexa/ven_coref.zip`

ICDAR	J. Comput. Phys.	Intl. Symp. Wearable Comp.
a	equations	realtime
document	for	positioning
recognition	numerical	a
handwritten	and	system
on	in	wearable
and	method	novel
using	with	sensing
of	of	for
for	the	computers
line	a	personal

Table 2: Most probable title words for three example clusters. ICDAR is the *International Conference on Document Analysis and Recognition.*

Intl. Symp. on Wearable Comp.	**Intl. Semantic Web Conf.**
Realtime Personal Positioning System for Wearable Computers.	*Benchmarking DAML+OIL Repositories*
Acceleration Sensing Glove (ASG).	*TRIPLE - A Query, Inference, and Transformation Language for the Semantic Web.*

Table 3: Examples of ambiguous acronyms that are correctly disambiguated by the DPVSW model. Shown in bold are the canonical strings for the clusters. All of these mentions have the venue string "ISWC."

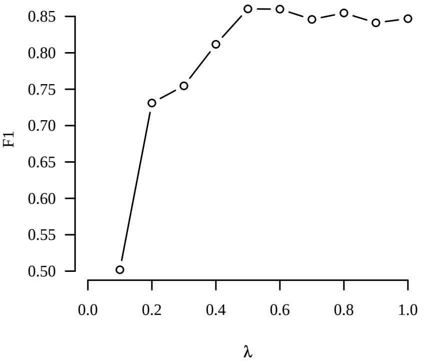

Figure 3: Sensitivity of DPVT model to title-Dirichlet parameter λ.

Model	Split-Merge	Gibbs
DPVT	84.9±1.06	80.5±1.20
DPVSW	80.0±1.39	78.1±0.41
DPV	63.8±0.20	63.4±0.61

Table 4: Comparison of B^3 F1 performance between the two proposals used in Metropolis Hastings sampling. Means and standard deviations are computed from 5 independent trials.

limit both techniques sample from the same distribution, for finite sample sizes one sampler might converge significantly faster to the posterior. The split-merge sampler that we use [8] has been relatively recently proposed, and has not to our knowledge been used for DP models of coreference, so it is interesting to see if it performs better than more typical inference algorithms. We measure this by the B^3 performance after each iteration of both samplers. This is shown in Figure 2. Clearly, the samples from split-merge perform much better than those of Gibbs sampling. The split-merge sampler at 300 iterations finds venue assignments that are comparable to those of the Gibbs sampler at 1000 iterations.

Interestingly, however, the split-merge sampler shows the greatest benefit for the DPVT model (see Table 4). For the other models, the improvement due to the split-merge sampler is modest.

5.3 Sensitivity to Hyperparameters

There are two hyperparameters that must be tuned in the DP venue-topic models: the strength parameter α of the DP prior, and the concentration parameter λ that controls how similar the per-venue title distributions are across venues. We choose these parameters by parameter sweep on a small development set of 200 mentions and 4 venues. (This is about 10% of the size of the test set.) The results of the parameter sweep are shown in Figure 3. The model is somewhat sensitive to the choice of the concentration λ of the title Dirichlet prior. It is possible to sample λ based on the training data [16], so that no labeled validation set is required, but we leave that to future work.

The model is, however, not sensitive to choice of the DP parameter α (not shown in the figure). Over 2 orders of magnitude, taking $\alpha \in \{0.01, 0.1, 1.0\}$ yields comparable performance (85.9, 84.9, and 85.6 F1 respectively).

6. CONCLUSIONS AND FUTURE WORK

We present an unsupervised nonparametric Bayesian model for coreference of research venues. Although related models have been applied to coreference of paper titles and authors, research venues have several unique characteristics that warrant special modeling. By exploiting the fact that research venues have a characteristic distribution over titles, we obtain a dramatic increase in performance on venue coreference. In particular, the model is even able to accurately split up venues that have string-identical abbreviations.

Several directions are available for future work. First, if labeled training data is available, then this model readily lends itself to semi-supervised prediction. This could be necessary to match the performance of discriminative coreference systems. It would also be interesting to extend this model to deduplicate papers and authors.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval, in part by Lockheed Martin through prime contract #FA8650-06-C-7605 from the Air Force Office of Scientific Research, in part by The Central Intelligence Agency, the National Security Agency and National Science Foundation under NSF grant #IIS-0326249, and in part by the Defense Advanced Research Projects Agency (DARPA), through the Department of the Interior, NBC, Acquisition Services Division, under contract number NBCHD030010, and AFRL #FA8750-07-D-0185. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

7. REFERENCES

[1] A. Bagga and B. Baldwin. Algorithms for scoring coreference chains. In *Proceedings of MUC7*, 1998.

[2] I. Bhattacharya and L. Getoor. A latent dirichlet model for unsupervised entity resolution. In *SIAM Conference on Data Mining (SDM)*, 2006.

[3] I. Bhattacharya and L. Getoor. Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data*, 1(1):1–36, March 2007.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993, 2003.

[5] P. Carbonetto, J. Kisynski, N. de Freitas, and D. Poole. Nonparametric Bayesian logic. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2005.

[6] C. Chemudugunta, P. Smyth, and M. Steyvers. Modeling general and specific aspects of documents with a probabilistic topic model. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 241–248. MIT Press, Cambridge, MA, 2007.

[7] A. Culotta and A. McCallum. Joint deduplication of multiple record types in relational data. In *CIKM*, pages 257–258, 2005.

[8] D. B. Dahl. Sequentially-allocated merge-split sampler for conjugate and nonconjugate dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 2005.

[9] A. Haghighi and D. Klein. Unsupervised coreference resolution in a nonparametric Bayesian model. In *ACL*, 2007.

[10] A. McCallum and B. Wellner. Conditional models of identity uncertainty with application to noun coreference. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 905–912. MIT Press, Cambridge, MA, 2005.

[11] B. Milch. *Probabilistic Models with Unknown Objects*. PhD thesis, University of California, Berkeley, 2006.

[12] R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249–265, 2000.

[13] H. M. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. Identity uncertainty and citation matching. In *NIPS*, 2003.

[14] P. Singla and P. Domingos. Multi-relational record linkage. In *KDD-2004 Workshop on Multi-Relational Data Mining*, pages 31–48, 2004.

[15] P. Singla and P. Domingos. Entity resolution with markov logic. In *Sixth International Conference on Data Mining*, pages 572–582, 2006.

[16] Y. W. Teh. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 985–992, 2006.

[17] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[18] B. Wellner, A. McCallum, F. Peng, and M. Hay. An integrated, conditional model of information extraction and coreference with application to citation graph construction. In *20th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2004.