A Conditional Model of Deduplication for Multi-Type Relational Data

Aron Culotta, Andrew McCallum
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
{culotta, mccallum}@cs.umass.edu

Abstract

Record deduplication is the task of merging database records that refer to the same underlying entity. In relational databases, accurate deduplication for records of one type is often dependent on the merge decisions made for records of other types. Whereas nearly all previous approaches have merged records of different types independently, this work models these inter-dependencies explicitly to collectively deduplicate records of multiple types. We construct a conditional random field model of deduplication that captures these relational dependencies, and then employ a novel relational partitioning algorithm to jointly deduplicate records.

We evaluate the system on two citation matching datasets, for which we deduplicate both papers and venues. We show that by collectively deduplicating paper and venue records, we obtain up to a 30% error reduction in venue deduplication, and up to a 20% error reduction in paper deduplication over competing methods.

1 Introduction

A common prerequisite for knowledge discovery is accurately combining data from multiple, heterogeneous sources into a unified, mineable database. An important step in creating such a database is *record deduplication*: consolidating multiple records that refer to the same abstract entity. The difficulty in this task arises both from data errors (e.g. misspellings and missing fields) and from variants in field values (e.g. abbreviations).

Most historical approaches have framed the deduplication problem as a set of independent decisions. For each pair of records, a similarity score is calculated, and the records are merged if the similarity is above some threshold [8]. The decisions are combined by taking the transitive closure of the resulting adjacency matrix.

More recently, McCallum and Wellner [13] and Parag and Domingos [20] have demonstrated that making multiple deduplication decisions collectively can provide better results than historical approaches. These models are types of conditional random fields (CRFs) [9], where the observed nodes are mentions, and the predicted nodes are the deduplication decisions for each pair of nodes. By framing inference as an instance of graph partitioning, the models are "collective" in the sense that mentions are clustered based not only on their distance to each other, but also on their distance from all other partitions. By treating

deduplication decisions in *dependent relation* to each other, inconsistencies and noise in the similarity metric may be overcome.

This paper presents a model for collective deduplication, extended to the important and ubiquitous case of relational databases, where records have *types*, and where there exist *relations* between records of different types. These relations provide useful evidence for deduplication decisions because the identity of a record often depends on the identities of related records.

For example, consider a database of research papers, where records can be of type *paper*, *venue*, or *author*. If two *paper* records are labeled as duplicates, then it follows that the *venue* records corresponding to those papers should also be labeled as duplicates. The reverse is more subtly true: if two *venues* are duplicates, then this may slightly increase the probability that their corresponding *papers* are duplicates.

We propose a model that leverages these subtle interdependencies to make deduplication decisions collectively across multiple record types.

In particular, we present a CRF for the citation domain that provides a conditional probabilistic model of deduplication decisions over records of multiple types given observed record mentions and the relations among them. We propose a novel, relational graph partitioning algorithm for inference that not only ensures that deduplication decisions made for different record types are consistent, but also allows the decisions from one record type to inform the decisions for another record type.

Parameter estimation consists of maximizing the product of local marginals for pairs of records of different types. That is, we parameterize the CRF to learn weights over 4-tuples consisting of a record pair *and* a related record pair of a different type. In the citation domain, these 4-tuples consist of a pair of paper records, and a pair of related venue records. In this way, the model learns parameters to trade-off paper and venue deduplication decisions.

We provide results on a database of research papers, where we show that modeling deduplication of paper and venue records collectively improves deduplication performance for each type, providing up to a 30% error reduction in venue deduplication, and up to a 20% error reduction in paper deduplication over a previously proposed collective model [13] that does not model the dependencies between record types.

2 Related Work

To the best of our knowledge, this is the first paper to present a discriminative, collective model of deduplication for multiple, related record types and demonstrate empirically the performance gains attainable over independent models. We briefly review classical work in deduplication, then discuss recent efforts in collective deduplication.

Record deduplication, known variously as record linkage, coreference resolution, deduplication, and identity uncertainty, is prevalent in many fields, including computer vision, databases, and natural language processing.

Originally introduced in the database community as "record linkage" [17], record deduplication was later formalized by Fellegi and Sunter [8] as the computation over features between pairs of records, and further extended by Winkler [25, 26]. This previous work calculates a similarity score for record pairs, collapses those above a similarity threshold, then performs transitive closure. It is not relational in the sense that one deduplication decision does not directly affect another.

More recent record linkage work has considered the deduplication of categorical data, allowing attributes to be deduplicated along with records [1]; however, that work does not utilize machine learning and requires thresholds to be set manually.

Methods of learning a better similarity score have been investigated recently in the database community [5, 7]. Similar trends exist in natural language processing for the task of coreference resolution, where research has focused on learning more useful similarity metrics and applying them to thresholding techniques analogous to those found in the database community [18, 16].

Only recently have collective deduplication models been investigated. Milch et. al have introduced generative models to reason in worlds with an unknown number of objects, enabling probability distributions to be defined over relational data with many object types [11, 15]. While these models have appealing formal semantics, their generative nature forces the model to make conditional independence assumptions among features.

Our work can be viewed as extensions to recent models applying conditional random fields to the deduplication task [13, 24, 20]. McCallum and Wellner [13] have presented CRFs which perform collective coreference by equating inference in the CRF as a graph partitioning problem, resulting in collective coreference of records of one type. This previous work demonstrated the advantages collective coreference has over classical approaches; however, it does not model multiple types of coreferent objects.

Parag and Domingos [20] present a CRF model similar to that in McCallum and Wellner [13] in that it collectively deduplicates records of the same type using graph partitioning. Additionally, this method allows information to propagate between records by way of their shared attributes. However, the Parag and Domingos model does not treat attributes as first-class objects. In particular, their model collapses string identical attribute nodes and creates "information nodes" to model whether or not attributes match. The model does not explicitly optimize deduplication decisions for attributes; rather, the "information nodes" can be viewed as an input variables to record deduplication. An important distinction with our work is that in the Parag and Domingos model, *joint deduplication only occurs among records sharing an identical attribute*. This is often not the case in real data.

In a sense, the Parag and Domingos model can be viewed as a discriminative version of a recently proposed hierarchical model for deduplication by Ravikumar and Cohen [22], which introduces latent match nodes for attributes. Here again, determining whether attribute values are coreferent is viewed as a local decision used as input to the record deduplication decision. Our model instead

PID	Author	Title	Venue	VID
0	X. Li	*Predicting the stock market*	CIKM	10
1	X. Li	*Predicting the stock market*	Conf on Information Management	20
2	J. Smith	*Semi-Definite Programming*	CIKM	30
3	Smith, J.	*Semi-Definate Programing*	Conference on Info Management	40

Table 1: An example of four papers with the same venues in a publications database. **PID** is the paper id, and **VID** is the venue id.

treats attributes as records themselves, performing full deduplication on them as well as their related records.

Other recent work in knowledge discovery has leveraged relational information to perform coreference [3, 4]. These models define a similarity metric between records that considers the identity of related records. This is similar in spirit to our model, since it uses deduplication decisions of related records to calculate the similarity between records. However, their model is mainly concerned with deduplicating authors alone, and does not explicitly model deduplication of multiple record types. Also, the training methods described in those models do not capture the rich set of features available to the model presented in this paper.

Our model can also be described as a type of relational Markov network (RMN) [23], which have been employed successfully in relational domains, although not multi-type deduplication. Another important distinction is that our training and inference methods differ substantially from the loopy belief propagation algorithm commonly used in RMNs.

3 Motivating Example

We first provide an example to motivate the potential benefits of collective deduplication for databases with multiple record types.

Consider again a database of research papers, with author, paper, and venue records. Our task is to deduplicate the various mentions of these records into unique entities. Table 1 shows a database of four papers and four venues, each with unique ids. Papers 0 and 1 and papers 2 and 3 should be merged; all the venues should be merged.

Imagine an agglomerative deduplication system which begins by assuming each record is unique. Suppose the system first considers merging papers 0 and 1. Although the venues do not match, all the other fields are exact matches, so it is feasible that the system may overcome this discrepancy. After merging papers 0 and 1, the system also merges the corresponding venues 10 and 20 into the same cluster, since the venues of duplicate papers must themselves be duplicates.

Imagine the system next merges venues 10 and 30 because they are string identical. The system must now decide if papers 2 and 3 are duplicates. Treated in isolation, a system may have a hard time correctly detecting that 2 and 3 are duplicates: the authors are highly similar, but the title contains two

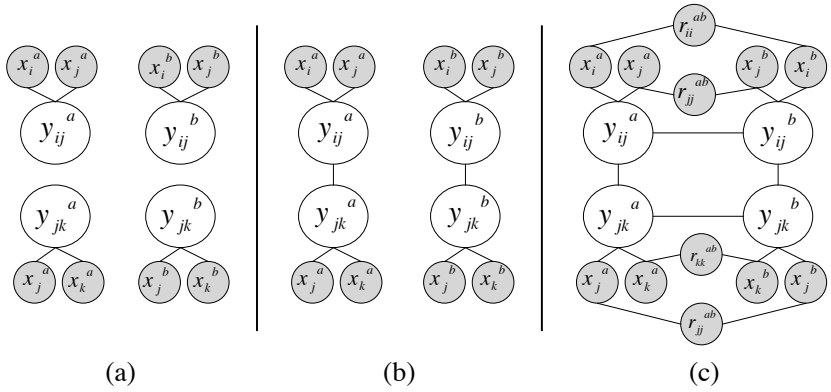

Figure 1: Three increasingly complex models of paper and venue deduplication. $\mathbf{X}^a$ and $\mathbf{X}^b$ are paper and venue records, respectively. $\mathbf{Y}$ is a binary random variable indicating duplicate records, and $\mathbf{R}^{ab}_{ij}$ denotes whether paper X^a_i was published in venue X^b_j.

misspellings, and the venues are extremely dissimilar.

However, the system we have described so far is fortunate to have more information at its disposal. It has already merged venues 10 and 20, which are highly similar to venues 30 and 40. By consulting its database of deduplicated venues, it could determine that 30 and 40 are in fact the same venue. With this information in hand, it may be more forgiving of the spelling mistakes in the title, finally merging papers 2 and 3 correctly.

This example illustrates the notion that the identity of an object is dependent on the identity of related objects. Notice that by using relational information, the system not only merged papers it might not have otherwise (2 and 3), but also merged venues it might not have (10, 20 and 30, 40). Indeed, the chain of deduplication decisions which led to optimal performance interleaved paper and venue decisions: (0,1), (10,20), (30,40), (2,3).

The work presented here describes a system that models the deduplication decisions of related records collectively, enabling the sort of probabilistic trade-offs instrumental to the success of the system in this example.

4 Model

The model is an instance of a conditional random field that jointly models the conditional probability of multiple deduplication decisions given an observed relational database.

We begin with a brief review of conditional random fields, followed by a formal description of the model. We then describe the approximations used to make inference and parameter estimation tractable for this model.

4.1 Conditional Random Fields

Conditional random fields (CRFs) [9] are undirected graphical models encoding the conditional probability of a set of output variables $\mathbf{Y}$ given a set of evidence variables $\mathbf{X}$. The set of distributions expressible by a CRF is specified by an undirected graph $\mathcal{G}$, where each vertex corresponds to a random variable. If $C = \{\{\mathbf{y}_c, \mathbf{x}_c\}\}$ is the set of cliques in $\mathcal{G}$, then the conditional probability of $\mathbf{y}$ given $\mathbf{x}$ is

$$p_\Lambda(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_\mathbf{x}} \prod_{c \in C} \phi_c(\mathbf{y}_c, \mathbf{x}_c; \Lambda)$$

where ϕ is a potential function parameterized by Λ and $Z_x = \sum_\mathbf{y} \prod_{c \in C} \phi(\mathbf{y}_c, \mathbf{x}_c)$ is a normalization factor. We assume ϕ_c factorizes as a log-linear combination of arbitrary features computed over clique c, therefore

$$\phi_c(\mathbf{y}_c, \mathbf{x}_c; \Lambda) = \exp\left(\sum_k \lambda_k f_k(\mathbf{y}_c, \mathbf{x}_c)\right)$$

The model parameters $\Lambda = \{\lambda_k\}$ are a set of real-valued weights typically learned from labeled training data by maximum likelihood estimation.

4.2 CRFs for multi-type deduplication

Let $\mathbf{X}$ be a collection of random variables representing observed record mentions in a database that requires deduplication. For clarity, assume there are only two types of records, $\mathbf{X} = (\mathbf{X}^a, \mathbf{X}^b)$, where $\mathbf{X}^a = (X_1^a, \ldots, X_n^a)$, $\mathbf{X}^b = (X_1^b, \ldots, X_m^b)$. The goal of deduplication is to partition $\mathbf{X}$ into clusters of records that all refer to the same abstract entity.

To this end, we define a collection of binary random variables $\mathbf{Y} = (\mathbf{Y}^a, \mathbf{Y}^b)$ that indicate whether or not two records are duplicates. For example, Y_{ij}^a indicates whether or not records X_i^a and X_j^a are coreferent. We also define the binary random variables $\mathbf{R}$, where R_{ij}^{ab} indicates whether some arbitrary relation R holds between record mentions X_i^a and X_j^b.

For example, in a research paper database, $\mathbf{X}^a$ represents the set of paper records, $\mathbf{X}^b$ represents the set venue records, Y_{ij}^a indicates whether X_i^a and X_j^a are duplicates, and R_{ij}^{ab} indicates whether paper X_i^a was published at venue X_j^b.

In the general case where $\mathbf{R}$ is unobserved, one could construct the conditional distribution $P(\mathbf{Y}^a, \mathbf{Y}^b, \mathbf{R}|\mathbf{X})$. With this model, one can infer from an observed set of records the most probable set of duplicate records *and* the most probable set of relations between records. For example, in the publications domain, one may want to model the *advisor_of* relation between authors, while also modeling author deduplication.

We postpone this investigation for future work, and instead focus on the case where $\mathbf{R}$ is observed. For instance, in citation data, we know which venues records are related to which paper records. Thus, we desire to model the conditional distribution $P(\mathbf{Y}^a, \mathbf{Y}^b|\mathbf{X}, \mathbf{R})$.

Figure 1 displays three increasingly complex graphical models of this conditional distribution. Vertices are random variables, edges indicate a possible probabilistic dependence between variables, and shaded vertices indicate observed variables. Model (a) corresponds to the classical approach, which treats each duplication decision independently. Model (b) is the approach evaluated in McCallum and Wellner [13], extended here to the case of multiple record types. Note that in this model deduplication decisions for records of the same type are made collectively.

Model (c) is the one this paper advocates. Not only are deduplication decisions for records of the *same* type made collectively, but also the decisions for one type of record are dependent on decisions made for related records. For ease of presentation, we have only included the observed relation variables $\mathbf{R}$ which are true.

We now provide a more precise description of model (c).

Let $\mathbf{x}_{ij}^{ab} = \langle x_i^a, x_j^a, x_i^b, x_j^b \rangle$ be a pair of observed paper record mentions and their corresponding venue records. To capture the dependence between y_{ij}^a and y_{ij}^b, we factorize the potential functions to consider them jointly, resulting in the model:

$$p(\mathbf{y}^a, \mathbf{y}^b | \mathbf{x}, \mathbf{r}) = \frac{1}{Z_{\mathbf{x}}} \exp\Big(\sum_{i,j,l} \lambda_l f_l(\mathbf{x}_{ij}^{ab}, y_{ij}^a, y_{ij}^b, r_{ij}^{ab}) + \sum_{i,j,k} \lambda_* f_*(y_{ij}^a, y_{jk}^a, y_{ik}^a, y_{ij}^b, y_{jk}^b, y_{ik}^b) \Big)$$

where the features f_* are consistency checking functions used to enforce transitivity among deduplication decisions. For example, if papers x_i^a and x_j^a are coreferent, and x_j^a and x_k^a are coreferent, then not only must papers x_i^a and x_k^a be coreferent, but venues x_i^b, x_j^b, x_k^b must also be coreferent. (Note f_* is of notational use only — in practice, the inference algorithm simply avoids these impossible configurations.)

Because both y_{ij}^a and y_{ij}^b are arguments to the feature functions f_l, these potentials capture the cross-product of paper and venue deduplication decisions. This allows the learned weights to encourage merging paper records which have equivalent venue records, and to discourage merging papers with different venues.

The cost of modeling these interdependencies is a highly connected graphical model, which necessitates approximations in both inference and parameter estimation. We describe these approximations below.

4.3 Inference

Inference in this model corresponds to finding the solution to

$$\mathbf{y}^* = (\mathbf{y}^{a*}, \mathbf{y}^{b*}) = \operatorname*{argmax}_{\mathbf{y}} p_\Lambda(\mathbf{y}^a, \mathbf{y}^b | \mathbf{x}^a, \mathbf{x}^b, \mathbf{r})$$

that is, finding the most probable deduplication decisions $\mathbf{y}^*$ given $\mathbf{x}^a, \mathbf{x}^b, \mathbf{r}$ and the learned parameters Λ.

Exact inference in this model is intractable because the space of possible $\mathbf{y}$ is exponential in the number of records $\mathbf{x}$, and the the high connectivity of the graph precludes a feasible dynamic program to make this search tractable.

One common approximate inference technique for such a predicament is to perform *loopy belief propagation*; that is, perform standard belief propagation [21], ignoring the "message double-counting" caused by the cycles in the graph. However, the severe cyclicity of this model may require a prohibitive amount of time for belief propagation to converge, if it converges at all.

Instead, we follow recent work which finds an equivalence between graph partitioning algorithms and inference in certain undirected graphical models [6]. We first transform our graph to a weighted, undirected graph that only contains vertices for variables $\mathbf{x}$ and has edges weighted by the (log) clique potential for each pair of vertices. The value on these edges depends on which type of records they join.

For paper edges, we define the weight

$$w_{ij}^a = \sum_{y_{ij}^b \in \{0,1\}} \Big(\sum_l \lambda_l f_l(\mathbf{x}_{ij}^{ab}, y_{ij}^a = 1, y_{ij}^b, r_{ij}^{ab}) - \sum_l \lambda_l f_l(\mathbf{x}_{ij}^{ab}, y_{ij}^a = 0, y_{ij}^b, r_{ij}^{ab}) \Big)$$

and similarly for venue edges:

$$w_{ij}^b = \sum_{y_{ij}^a \in \{0,1\}} \Big(\sum_l \lambda_l f_l(\mathbf{x}_{ij}^{ab}, y_{ij}^a, y_{ij}^b = 1, r_{ij}^{ab}) - \sum_l \lambda_l f_l(\mathbf{x}_{ij}^{ab}, y_{ij}^a, y_{ij}^b = 0, r_{ij}^{ab}) \Big)$$

Intuitively, the paper weights w_{ij}^a can be thought of as the compatibility of papers x_i^a, x_j^a, summed over possible deduplication decisions for venues x_i^b, x_j^b. Similarly, the venue weights w_{ij}^b can be thought of as the compatibility of venues x_i^b, x_j^b, summed over the possible deduplication decisions for papers x_i^a, x_j^a. Interpreting the weights as the *similarity* between two records, we can see that the similarity of paper records considers the similarity of their venue records, and vice versa.

This results in a weighted, undirected graph with edge weights ranging from $-\infty$ to $+\infty$. It can be shown that finding an optimal partitioning of this graph corresponds to finding the optimal configuration $\mathbf{y}^*$ in the original undirected graphical model. Here, the number of partitions is unknown, as it corresponds to the number of unique records.

Although graph partitioning with positive and negative edge weights is NP-hard, there exist several good approximations, including recent work in correlation clustering [2]. Additionally, McCallum and Wellner [13] have found that greedy agglomerative clustering with an average link criterion works well in practice.

However, traditional partitioning algorithms would not account for the known dependencies between clusters that exist in our data. Therefore, we develop a novel, *relational agglomerative* clustering algorithm that exploits these dependencies.

Traditional greedy agglomerative clustering first initializes each vertex to its own cluster, then iteratively merges the clusters that are "closest," where the distance between clusters is often defined as the average of the edge weights connecting the two clusters. We augment this algorithm with two enhancements.

First, we must enforce the constraint that duplicate papers have duplicate venues. This is straight-forwardly enforced by the following rule: Whenever a pair of paper clusters are merged, their corresponding venue clusters must also be merged.

The second enhancement redefines the distance between clusters to more accurately reflect the impact of the first enhancement. Let C_i^a, C_j^a be two paper clusters that are candidates to be merged, and let C_i^b, C_j^b be the venue clusters corresponding to these papers. The first enhancement requires that if we merge C_i^a, C_j^a, we must also merge C_i^b, C_j^b. However, the current distance metric between C_i^a, C_j^a does not reflect this fact.

To remedy this, we redefine the distance between two paper clusters (C_i^a, C_j^b) to be the average of (1) the traditional distance between the paper clusters (C_i^a, C_j^b) and (2) the traditional distance between their corresponding venue clusters (C_i^b, C_j^b). (Note that we choose the average rather than the sum to deal with papers that have no venue information.) This metric is likely to better approximate the effect merging C_i^a, C_j^a will have on the objective function $p_\Lambda(\mathbf{y}|\mathbf{x}, \mathbf{r})$, since it accounts for the merger of the corresponding venue clusters.

This new clustering algorithm provides benefits to both paper and venue deduplication that would be unavailable in an independent clustering algorithm. As illustrated in our motivating example in Section 3, it is often the case that paper duplicates are not detected because they have venues with decidedly different surface forms (e.g. "CIKM" and "Conference on Knowledge and Information Management"). The second enhancement addresses this problem by using the evidence from previous venue clusterings to inform paper deduplication. Specifically, if "CIKM" has already been resolved with " Conference on Knowledge and Information Management," then merging papers with venues "CIKM" and "Conference on Knowledge and Information Management" will be encouraged, since there will be a high similarity between their associated venue clusters.

Conversely, by the hard constraint introduced in the first enhancement, difficult venue deduplication decisions are informed by confident paper deduplication decisions, as was also illustrated in Section 3. In this way, deduplication decisions for both record types simultaneously grow more accurate.

4.4 Parameter Estimation

Given a labeled corpus of fully clustered data, maximum likelihood parameter estimation corresponds to finding the parameters Λ which maximize the log-likelihood of the labeled training data. Exact estimation is intractable here because it requires calculating the normalization term $Z_{\mathbf{x}}$, a sum over all possible values of $\mathbf{y}$, which is a sum over all possible partitionings of the data. Due to the nature of the data and the high connectivity of the graph, this cannot be efficiently computed with a dynamic program.

One could perform stochastic gradient ascent on an approximation of the likelihood. However, it has been noted that maximizing a product of local marginals performs at least as well as this approximation on a similar coreference task, if not better [24, 12]. Whereas in [24] the local marginals are over single coreference decisions, here we maximize a product of joint conditional probabilities for decisions y^a_{ij} and y^b_{ij}, which we define as

$$P_\Lambda(y^a_{ij}, y^b_{ij} | \mathbf{x}^a, \mathbf{x}^b, \mathbf{r}) = \frac{1}{Z'_{\mathbf{x}}} \exp\left(\sum_{i,j,l} \lambda_l f_l(\mathbf{x}^{ab}_{ij}, y^a_{ij}, y^b_{ij}, r_{ij})\right)$$

where $Z'_{\mathbf{x}}$ is a normalization constant summing over possible values for the pair $\langle y^a_{ij}, y^b_{ij}\rangle$.

For a labeled dataset $\mathcal{D}$, the log-likelihood is defined as

$$\mathcal{L}_\Lambda(\mathcal{D}) = \log\left(\prod_{\langle y^a_{ij}, y^b_{ij}\rangle \in \mathcal{D}} P_\Lambda(y^a_{ij}, y^b_{ij} | \mathbf{x}^a, \mathbf{x}^b, \mathbf{r})\right)$$

We perform gradient ascent on $\mathcal{L}$ by maximizing its derivative:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \lambda_l} = \sum_{\langle \mathbf{x}, \mathbf{y}, \mathbf{r}\rangle \in \mathcal{D}} \Big(& \sum_{i,j,l} \lambda_l f_l(\mathbf{x}^{ab}_{ij}, y^a_{ij}, y^b_{ij}, r^{ab}_{ij}) \\ & - \sum_{\langle y'^a_{ij}, y'^b_{ij}\rangle} P_\Lambda(y'^a_{ij}, y'^b_{ij} | \mathbf{x}^a, \mathbf{x}^b, \mathbf{r}) \\ & \sum_{i,j,l} \lambda_l f_l(\mathbf{x}^{ab}_{ij}, y'^a_{ij}, y'^b_{ij}, r^{ab}_{ij})\Big)\end{aligned}$$

Because the defined likelihood is a convex function, we can perform gradient ascent using any suitable optimization algorithm. In particular, we use limited-memory BFGS, which iteratively approximates second-order curvature information to speed up convergence [19].

The estimation method can also be viewed as learning a distance metric between paper-venue pairs. As explained in Section 4.3, this metric is used to weight the edges in the deduplication graph, which is then partitioned at inference time.

5 Experiments

We evaluate our model on two datasets of research paper citations. The first is from Citeseer [10], containing approximately 1500 citations, with 900 unique papers and 350 unique venues. The second is from the Cora Computer Science Research Paper Engine[1], containing about 1800 citations, with 600 unique papers and 200 unique venues.

Both datasets are manually labeled for both paper coreference and venue coreference, as well as manually segmented into fields, such as author, title, etc. The data were collected by searching for certain authors and topic, and they are split into subsets with non-overlapping papers for the sake of cross-validation experiments.

We used a number of feature functions, including exact and approximate string match[2] on normalized and unnormalized values for the following citation fields: title, booktitle, journal, authors, venue, date, editors, institution, and the entire unsegmented citation string. We also calculated an unweighted cosine similarity between tokens in the title and author fields. Additional features include whether or not the papers have the same publication type (e.g. journal or conference), as well as the numerical distance between fields such as year and volume. All real values were binned and converted into binary-valued features.

To evaluate performance, we compare the clusters output by our system with the true clustering using pairwise metrics. Pairwise precision is the fraction of pairs in the same cluster that are coreferent; pairwise recall is the fraction of coreferent papers that were placed in the same cluster. Pairwise F1 is the harmonic mean of pairwise precision and pairwise recall.

Tables 2 and 3 show the F1 performance of two systems: **joint** is the system we have advocated in this paper, and **indep** is the system which deduplicates records of different types *independently*. Note that this system corresponds to model (b) in Figure 1, so deduplication decisions *are* made collectively for records of the same type, as in McCallum and Wellner [13]. Since the McCallum and Wellner model has been shown to consistently outperform the classical transitive closure model, we do not compare with the classical model here.

Results are listed by the name of each test set; the remaining sections are used for training.

Venue performance improves considerably in the joint model, which is plausible considering the strong influence paper deduplication has on venue deduplication. Because paper deduplication often has more evidence at its disposal than does venue deduplication, the joint model dramatically enhances venue recall, obtaining a 5% absolute recall boost in Citeseer, and a 9% boost in Cora data. This is especially noticeable when paper deduplication performance is high: The hard constraint requiring the venues of duplicate papers to be merged often merges venues that otherwise would have seemed too dissimilar to merge on their own. Indeed, error analysis confirms that many of the venue

[1] http://www.cs.umass.edu/~mccallum/data/cora-refs.tar.gz

[2] We used the Secondstring package, found at http://secondstring.sourceforge.net

	Paper		Venue	
	indep	**joint**	**indep**	**joint**
constraint	88.9	**91.0**	79.4	**94.1**
reinforce	92.2	92.2	56.5	**60.1**
face	88.2	**93.7**	80.9	**82.8**
reason	**97.4**	97.0	75.6	**79.5**
Micro Avg.	91.7	**93.4**	73.1	**79.1**

Table 2: Pairwise F1 deduplication performance on Citeseer data.

	Paper		Venue	
	indep	**joint**	**indep**	**joint**
kibl	92.9	**93.3**	93.6	**99.3**
fahl	**95.5**	95.0	87.3	**99.7**
utgo	79.9	**84.0**	51.7	**60.4**
Micro Avg.	89.4	**90.8**	77.5	**84.5**

Table 3: Pairwise F1 deduplication performance on Cora data.

deduplication errors our model avoids are those where venues are dissimilar in form, but are related to papers that are similar in form.

More interestingly, a noticeable improvement in paper deduplication is attained by the collective model. Part of this is due to the precision enhancement provided by the constrained clustering algorithm. Workshop and technical report versions of journal or conference papers with the same title are correctly not merged when the venues are accurately identified. Also, error analysis suggests that papers that would not have been otherwise merged were merged because their venues were determined to be coreferent.

It is worth noting that many of the errors made by the **joint** model have causes similar to those that on average have improved performance. For example, if paper deduplication accuracy is poor, the relational clustering algorithm can result in many venues being merged that would not have been otherwise. Future work should investigate how to detect poor paper accuracy and adjust accordingly.

5.1 Scalability

While the datasets used in our experiments are of reasonable size, we would ultimately like to apply this model to large databases. Here we briefly discuss performance issues and describe methods to scale our model to real-world data.

Because parameter estimation maximizes the product of local node potentials, it is likely to be much faster than a global approximate training method such as loopy belief propagation. For the data used in these experiments, training time averaged about 15 minutes on dual-processor, 3.06 GHz Xeon machines

with 4 GB of RAM. Inference time ranged from 20 minutes to about an hour, depending on the size of the testing data.

To make inference scalable, a practical implementation would make use of "canopies" [14]. This technique reduces the connectivity of the graph of records by defining a cheap similarity metric between records (often using an inverted index or tf-idf). When constructing the record graph, edges are only added between those records that have similarity above some threshold, where similarity is the output of the cheap metric. In this way, records that are very unlikely to be duplicates are not considered by the model. This provides an efficient, accurate way of pruning the search space.

Besides performance issues, there is reason to believe that the advantages of the model presented in this paper will be even more noticeable in larger data sets, where there is more heterogeneity in field values, more interesting relational patterns, and larger record clusters. In fact, the data used in experiments presented here contain many singleton clusters, which is not truly reflective of the large clusters of records found in real-world data.

6 Conclusions

We have introduced a collective model for deduplication of related records of multiple types and demonstrated empirically the advantages it has over methods that do not address the interdependencies inherent in relational data.

Based on these results, two promising areas of future research are (1) extending the model to databases with more than two types of records, and (2) modeling the relation variables $\mathbf{R}$. In addition to paper and venue deduplication, author deduplication is also a difficult problem that would likely benefit from this approach, and we are in the process of harvesting data to allow us to model author, venue, and paper deduplication jointly.

In the publications domain, the connections between author, venue, and paper deduplication become more interesting with larger databases, where communities and relations become more visible. In particular, exciting challenges include building a model to predict *advisor_of* relations between authors, suggest possible venues for a paper, identify fruitful author collaborations, match recent graduates with potential research labs, and discover the dynamics of research communities.

The challenge as usual will lie in developing a model that is complex enough to model these long-distance relations, but is still tractable enough to perform on real data. We feel that the model proposed here is a productive step in that long-term direction.

7 Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval, in part by U.S. Government contract #NBCH040171 through a subcon-

tract with BBNT Solutions LLC, 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. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s)' and do not necessarily reflect those of the sponsor.

References

[1] R. Ananthakrishna, S. Chaudhuri, and V. Ganti. Eliminating fuzzy duplicates in data warehouses. In *In Proceedings of the 28th International Conference on Very Large Databases (VLDB 2002)*, 2002.

[2] Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation clustering. *Machine Learining*, 56:89–113, 2004.

[3] Indrajit Bhattacharya and Lise Getoor. Deduplication and groupd detection using links. In *10th ACM SIGKDD Workshop on Link Analysis and Group Detection (LinkKDD-04)*, 2004.

[4] Indrajit Bhattacharya and Lise Getoor. Iterative record linkage for cleaning and integration. In *Proceedings of the SIGMOD 2004 Workshop on Research Issues on Data Mining and Knowledge Discovery*, June 2004.

[5] Mikhail Bilenko and Raymond J. Mooney. Adaptive duplicate detection using learnable string similarity measures. In *9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003)*, pages 39–48, 2003.

[6] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *In IEEE transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 23(11):1222–1239, 2001.

[7] W. Cohen and J. Richman. Learning to match and cluster entity names. In *ACM SIGIR'01 workshop on Mathematical / Formal Methods in IR*, 2001.

[8] I. P. Fellegi and A. B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64:1183–1210, 1969.

[9] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA, 2001.

[10] S. Lawrence, C. L. Giles, and K. Bollaker. Digital libraries and autonomous citation indexing. *IEEE Computer*, 32:67–71, 1999.

[11] Bhaskara Marthi, Brian Milch, and Stuart Russell. First-order probabilistic models for information extraction. In *Proceedings of the IJCAI-2003 Workshop on Learning Statistical Models from Relational Data*, pages 71–78, Acapulco, Mexico, August 2003.

[12] Andrew McCallum and Charles Sutton. Piecewise training with parameter independence diagrams: Comparing globally- and locally-trained linear-chain crfs. In *NIPS 2004 Workshop on Learning with Structured Outputs*, 2004.

[13] Andrew McCallum and Ben Wellner. Conditional models of identity uncertainty with application to noun coreference. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*. MIT Press, Cambridge, MA, 2005.

[14] Andrew K. McCallum, Kamal Nigam, and Lyle Ungar. Efficient clustering of high-dimensional data sets with application to reference matching. In *Proceedings of the Sixth International Conference On Knowledge Discovery and Data Mining (KDD-2000)*, Boston, MA, 2000.

[15] Brian Milch, Bhaskara Marthi, and Stuart Russell. Blog: Relational modeling with unknown objects. In *ICML 2004 Workshop on Statistical Relational Learning and Its Connections to Other Fields*, 2004.

[16] Thomas Morton. Coreference for NLP applications. In *ACL*, 1997.

[17] H. B. Newcombe, J. M. Kennedy, S. J. Axford, and A.P. James. Automatic linkage of vital records. *Science*, 130:954–9, 1959.

[18] Vincent Ng and Claire Cardie. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002.

[19] J. Nocedal. Updating quasi-newton matrices with limited storage. *Mathematics of Computation*, 95:339–353, 1980.

[20] Parag and Pedro Domingos. Multi-relational record linkage. In *Proceedings of the KDD-2004 Workshop on Multi-Relational Data Mining*, pages 31–48, August 2004.

[21] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufman, 1988.

[22] Pradeep Ravikumar and William W. Cohen. A hierarchical graphical model for record linkage. In *UAI 2004*, 2004.

[23] Ben Taskar, Abbeel Pieter, and Daphne Koller. Discriminative probabilistic models for relational data. In *Uncertainty in Artificial Intelligence: Proceedings of the Eighteenth Conference (UAI-2002)*, pages 485–492, San Francisco, CA, 2002. Morgan Kaufmann Publishers.

[24] Ben Wellner, Andrew McCallum, Fuchun Peng, and Michael Hay. An integrated, conditional model of information extraction and coreference with application to citation matching. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2004.

[25] William E. Winkler. Improved decision rules in the fellegi-sunter model of record linkage. Technical report, Statistical Research Division, U.S. Census Bureau, Washington, DC, 1993.

[26] William E. Winkler. Methods for record linkage and Bayesian networks. Technical report, Statistical Research Division, U.S. Census Bureau, Washington, DC, 2002.