Joint Deduplication of Multiple Record Types in Relational Data

Aron Culotta
University of Massachusetts
140 Governor’s Drive Amherst, MA USA
culotta@cs.umass.edu

Andrew McCallum
University of Massachusetts
140 Governor’s Drive Amherst, MA USA
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 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.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—clustering

General Terms
algorithms, performance

Keywords
record linkage, deduplication, conditional random fields

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 field value variants (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 [3]. The decisions are combined by taking the transitive closure of the resulting adjacency matrix.

More recently, McCallum and Wellner [6] and Parag and Domingos[7] have demonstrated that making multiple deduplication decisions collectively can provide better results than historical approaches. These models are types of conditional random fields (CRFs) [4], 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 to 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 publications database domain that provides a conditional probabilistic model of deduplication decisions for both paper and venue records given observed record mentions and the relations among
them. We propose a novel, constrained 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 is approximated by maximizing the product of local marginals for paper-venue pairs. That is, we paramterize the CRF to learn weights over 4-tuples consisting of paper record pairs and their corresponding venue record pairs. 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 [6] 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 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” [3], record deduplication was later formalized by Fellegi and Sunter [3] as the computation over features between pairs of records, and further extended by Winkler [7]. This work calculates a similarity score for record pairs, collapses those above a similarity threshold, then performs transitive closure. This is not relational in the sense that one deduplication decision does not directly affect another.

Methods of learning a better similarity score have been investigated recently in the database community [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 in thresholding techniques analogous to those found in the database community [7, 8].

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 [8]. 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 [6, 7]. McCallum and Wellner [6] have presented CRFs which perform collective coreference by equating inference in the CRF as a graph partitioning problem and shown their superiority over classical approaches. However, these models do not model multiple types of coreferent objects, but instead collectively deduplicate many records of the same type.

Parag and Domingos [7] present a CRF model similar to that in McCallum and Wellner[6] 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.

In this 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[7], 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 treats attributes as first-class objects, performing full deduplication on them as well as their related records.

Other recent work in knowledge discovery has leveraged relational information to perform coreference [1, 8]. 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.

Our model can also be described as a type of relational Markov network (RMN) [7], 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 more common loopy belief propagation 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 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 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) [4] are undirected graphical models encoding the conditional probability of a set of output variables $Y$ given a set of evidence variables $X$. The set of distributions expressible by a CRF is specified by an undirected graph $G$, where each vertex corresponds to a random variable. If $C = \{ \{ y_c, x_c \} \}$ is the set of cliques in $G$, then the conditional probability of $y$ given $x$ is

$$p_y(y|x) = \frac{1}{Z_x} \prod_{c \in C} \phi_c(y_c, x_c)$$

where $\phi$ is a potential function parameterized by $\Lambda$ and $Z_x = \sum_y \prod_{c \in C} \phi(y_c, x_c)$ is a normalization factor. We assume $\phi$ factorizes as a log-linear combination of arbitrary features computed over clique $c$, therefore

$$\phi_c(y_c, x_c) = \exp \left( \sum_k \lambda_k f_k(y_c, x_c) \right)$$

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 $X$ 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, $X = (X^a, X^b)$, where $X^a = (X^a_1, \ldots, X^a_n)$, $X^b = (X^b_1, \ldots, X^b_m)$. The goal of deduplication is to cluster $X$ into clusters of records that all refer to the same abstract entity.

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

For example, in a research paper database, $X^a$ represents the set of paper records, $X^b$ represents the set venue records, $Y^a_{ij}$ indicates whether or not records $X^a_i$ and $X^a_j$ are duplicates, and $R_{ij}$ indicates whether paper $X^b_i$ was published at venue $X^b_j$.

In the general case where $R$ is unobserved, one could construct the conditional distribution $P(Y^a, Y^b, R|X)$. 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.
Figure 1: Three increasingly complex models of paper and venue deduplication. $X^a$ and $X^b$ are paper and venue records, respectively. $Y$ is a binary random variable indicating duplicate records, and $R^i_j$ denotes whether paper $X^a_i$ was published in venue $X^b_j$.

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

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 [6], 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 we propose. 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 $R$ which are true.

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

Let $x_{ij}^{ab} = (x^a_i, x^b_j, x^b_j, x^b_j)$ 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:

$$p(y^a, y^b|x, r) = \frac{1}{Z_x} \exp \left( \sum_{i,j,l} \lambda_{ij} f_i(x_{ij}^{ab}, y_{ij}^a, y_{ij}^b, r_{ij}) + \sum_{i,j,k} \lambda_{jk} f_k(y_{ij}^a, y_{ij}^b, y_{ij}^a, y_{ij}^b, r_{ij}) \right)$$

where the features $f_k$ are consistency checking functions used to enforce transitivity among deduplication decisions. For example, if papers $x^a_i$ and $x^a_j$ are coreferent, and $x^b_j$ and $x^b_j$ are coreferent, then not only must papers $x^a_i$ and $x^a_j$ be coreferent, but venues $x^b_j$ and $x^b_j$ must also be coreferent. (Note $f_k$ is of notational use only — in practice, the inference algorithm simply avoids these impossible configurations.)

Because both $y_{ij}^a$ and $y_{ij}^a$ are arguments to the feature functions $f_k$, 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

$$y^* = (y^*^a, y^*^b) = \arg \max_y p(y^a, y^b|x^a, x^b, r)$$
that is, finding the most probable deduplication decisions \( y^* \) given \( x^a, x^b, r \) and the learned parameters \( \Lambda \).

Exact inference in this model is intractable because the space of possible \( y \) is exponential in the number of records \( x \), 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 [8], 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 [2]. We first transform our graph to a weighted, undirected graph that only contains vertices for variables \( x \) 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
\[
 w_{ij}^a = \sum_l \lambda_l f_i(x_{ij}^a, y_{ij}^a = 1, r_{ij}) - \sum_l \lambda_l f_i(x_{ij}^a, y_{ij}^a = 0, r_{ij})
\]
and similarly for venue edges:
\[
 w_{ij}^b = \sum_l \lambda_l f_i(x_{ij}^b, y_{ij}^b = 1, r_{ij}) - \sum_l \lambda_l f_i(x_{ij}^b, y_{ij}^b = 0, r_{ij})
\]

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 \( x_i^a, x_j^a \), summed over the possible deduplication decisions for papers \( x_i^b, x_j^b \).

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 \( y^* \) 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 negative edge weights is NP-hard, there exist several good approximations, including recent work in correlation clustering [?].

However, a traditional partitioning algorithms would not account for the known dependencies between clusters that exist in our data. Therefore, we develop a novel, constrained greedy 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 \( C_i^a, C_i^b \) be two paper clusters that are candidates to be merged, and let \( C_j, C_j' \) be the venue clusters corresponding to these papers. The first enhancement requires that if we merge \( C_i^a, C_i^b \), we must also merge \( C_j, C_j' \). However, the current distance metric between \( C_i^a, C_i^b \) does not reflect this fact.

To remedy this, we redefine the distance between two paper clusters \( d(C_i^a, C_j^a) \) to be the average of (1) the traditional distance between the paper clusters \( (C_i^a, C_j^a) \) and (2) the traditional distance between their corresponding venue clusters \( (C_i^b, C_j^b) \). This metric is likely to better approximate the effect merging \( C_i^a, C_i^b \) will have on the objective function \( p_\Lambda(y|x,r) \), 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. “KDD” and “Int’l Conference on Knowledge Discovery”). The second enhancement addresses this problem by using the evidence from the venue clusterings to inform paper deduplication.

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 \( \Lambda \) which maximize the log-likelihood of the labeled training data. Exact estimation is intractable here because it requires calculating the normalization term \( Z_\Lambda \), a sum over all possible values of \( y \), 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 approx-
We perform gradient ascent on \( L \). Whereas in \([?]\) 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

\[
P_{\lambda}(y^a_{ij}, y^b_{ij} | x^a, x^b, r) = \frac{1}{Z_{xy}} \exp \left( \sum_{i,j,l} \lambda_l f_l(x^a_{ij}, y^a_{ij}, y^b_{ij}, r_{ij}) \right)
\]

where \( Z_{xy} \) is a normalization constant summing over possible values for the pair \( (y^a_{ij}, y^b_{ij}) \).

For a labeled dataset \( D \), the log-likelihood is defined as

\[
\mathcal{L}_{\lambda}(D) = \log \left( \prod_{(y^a_{ij}, y^b_{ij}) \in D} P_{\lambda}(y^a_{ij}, y^b_{ij} | x^a, x^b, r) \right)
\]

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

\[
\frac{\partial \mathcal{L}}{\partial \lambda_l} = \sum_{(x,y,r) \in D} \left( \sum_{i,j,l} \lambda_l f_l(x^a_{ij}, y^a_{ij}, y^b_{ij}, r_{ij}) - \sum_{(y^a_{ij}, y^b_{ij})} P_{\lambda}(y^a_{ij}, y^b_{ij} | x^a, x^b, r) \sum_{i,j,l} \lambda_l f_l(x^a_{ij}, y^a_{ij}, y^b_{ij}, r_{ij}) \right)
\]

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 \([?]\).

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 clustered at inference time.

5. EXPERIMENTS

We evaluate our model on two datasets of research paper citations. The first is from Citeseer \([5]\), 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 200 unique papers and 600 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.

\(^1\)http://www.cs.umass.edu/~mccallum/data/cora-reft.tar.gz

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 presented, 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 \([6]\).

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. 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.

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, modest recall improvements suggest that papers that would not have been otherwise merged were merged because their venues were determined to be coreferent.

5.1 Scalability

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

Because parameter estimation maximizes the product of lo-
Table 2: Pairwise F1 deduplication performance on Citeseer data.

<table>
<thead>
<tr>
<th>Paper</th>
<th>indep</th>
<th>joint</th>
<th>Venue</th>
<th>indep</th>
<th>joint</th>
</tr>
</thead>
<tbody>
<tr>
<td>constraint</td>
<td>88.9</td>
<td>91.0</td>
<td>79.4</td>
<td>94.1</td>
<td></td>
</tr>
<tr>
<td>reinforce</td>
<td>92.2</td>
<td>92.2</td>
<td>56.5</td>
<td>60.1</td>
<td></td>
</tr>
<tr>
<td>face</td>
<td>88.2</td>
<td>93.7</td>
<td>80.9</td>
<td>82.8</td>
<td></td>
</tr>
<tr>
<td>reason</td>
<td>97.4</td>
<td>97.0</td>
<td>75.6</td>
<td>79.5</td>
<td></td>
</tr>
<tr>
<td>Micro Avg.</td>
<td>91.7</td>
<td>93.4</td>
<td>73.1</td>
<td>79.1</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Pairwise F1 deduplication performance on Cora data.

<table>
<thead>
<tr>
<th>Paper</th>
<th>indep</th>
<th>joint</th>
<th>Venue</th>
<th>indep</th>
<th>joint</th>
</tr>
</thead>
<tbody>
<tr>
<td>kibl</td>
<td>92.9</td>
<td>93.3</td>
<td>93.6</td>
<td>99.3</td>
<td></td>
</tr>
<tr>
<td>fahl</td>
<td>95.5</td>
<td>95.0</td>
<td>87.3</td>
<td>99.7</td>
<td></td>
</tr>
<tr>
<td>utgo</td>
<td>79.9</td>
<td>84.0</td>
<td>51.7</td>
<td>60.4</td>
<td></td>
</tr>
<tr>
<td>Micro Avg.</td>
<td>89.4</td>
<td>90.8</td>
<td>77.5</td>
<td>84.5</td>
<td></td>
</tr>
</tbody>
</table>

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 $R$. 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-related 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 subcontract 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.

8. REFERENCES