Modeling Missing Data in Distant Supervision for Information Extraction

Alan Ritter
Machine Learning Department
Carnegie Mellon University
rittera@cs.cmu.edu

Luke Zettlemoyer, Mausam
Computer Sci. & Eng.
University of Washington
{lsz,mausam}@cs.washington.edu

Oren Etzioni
Vulcan Inc.
Seattle, WA
orene@vulcan.com

Abstract

Distant supervision algorithms learn information extraction models given only large readily available databases and text collections. Most previous work has used heuristics for generating labeled data, for example assuming that facts not contained in the database are not mentioned in the text, and facts in the database must be mentioned at least once. In this paper, we propose a new latent-variable approach that models missing data. This provides a natural way to incorporate side information, for instance modeling the intuition that text will often mention rare entities which are likely to be missing in the database. Despite the added complexity introduced by reasoning about missing data, we demonstrate that a carefully designed local search approach to inference is very accurate and scales to large datasets. Experiments demonstrate improved performance for binary and unary relation extraction when compared to learning with heuristic labels, including on average a 27% increase in area under the precision recall curve in the binary case.

1 Introduction

This paper addresses the issue of missing data (Little and Rubin, 1986) in the context of distant supervision. The goal of distant supervision is to learn to process unstructured data, for instance to extract binary or unary relations from text (Bunescu and Mooney, 2007; Snyder and Barzilay, 2007; Wu and Weld, 2007; Mintz et al., 2009; Collins and Singer, 1999), using a large database of propositions as a distant source of supervision. In the case of binary relations, the intuition is that any sentence which mentions a pair of entities (e_1 and e_2) that participate in a relation, r, is likely to express the proposition $r(e_1, e_2)$, so we can treat it as a positive training example of r. Figure 1 presents an example of this process.

Person	EMPLOYER
Bibb Latané	UNC Chapel Hill
Tim Cook	Apple
Susan Wojcicki	Google

True Positive	*"***Bibb Latané***, a professor at the* **University of North Carolina at Chapel Hill***, published the theory in 1981."*
False Positive	*"***Tim Cook** *praised* **Apple***'s record revenue..."*
False Negative	*"***John P. McNamara***, a professor at* **Washington State University***'s Department of Animal Sciences..."*

Figure 1: A small hypothetical database and heuristically labeled training data for the EMPLOYER relation.

One question which has received little attention in previous work is how to handle the situation where information is missing, either from the text corpus, or the database. As an example, suppose the pair of entities (*John P. McNamara*, *Washington State University*) is absent from the EMPLOYER relation. In this case, the sentence in Figure 1 (and others which mention the entity pair) is effectively treated as a negative example of the relation. This is an issue

Transactions of the Association for Computational Linguistics, 1 (2013) 367–378. Action Editor: Kristina Toutanova.
Submitted 7/2013; Revised 8/2013; Published 10/2013.

of practical concern, as most databases of interest are highly incomplete - this is the reason we need to extend them by extracting information from text in the first place.

We need to be cautious in how we handle missing data in distant supervision, because this is a case where data is *not missing at random* (NMAR). Whether a proposition is observed or missing in the text or database depends heavily on its truth value: given that it is true we have some chance to observe it, however we do not observe those which are false. To address this challenge, we propose a joint model of extraction from text and the process by which propositions are observed or missing in both the database and text. Our approach provides a natural way to incorporate *side information* in the form of a *missing data model*. For instance, popular entities such as Barack Obama already have good coverage in Freebase, so new extractions are more likely to be errors than those involving rare entities with poor coverage.

Our approach to missing data is general and can be combined with various IE solutions. As a proof of concept, we extend MultiR (Hoffmann et al., 2011), a recent model for distantly supervised information extraction, to explicitly model missing data. These extensions complicate the MAP inference problem which is used as a subroutine in learning. This motivated us to explore a variety of approaches to inference in the joint extraction and missing data model. We explore both exact inference based on A* search and efficient approximate inference using local search. Our experiments demonstrate that with a carefully designed set of search operators, local search produces optimal solutions in most cases.

Experimental results demonstrate large performance gains over the heuristic labeling strategy on both binary relation extraction and weakly supervised named entity categorization. For example our model obtains a 27% increase in area under the precision recall curve on the sentence-level relation extraction task.

2 Related Work

There has been much interest in distantly supervised[1] training of relation extractors using databases. For example, Craven and Kumlien (1999) build a heuristically labeled dataset, using the Yeast Protein Database to label Pubmed abstracts with the *subcellular-localization* relation. Wu and Weld (2007) heuristically annotate Wikipedia articles with facts mentioned in the infoboxes, enabling automated infobox generation for articles which do not yet contain them. Benson et. al. (2011) use a database of music events taking place in New York City as a source of distant supervision to train event extractors from Twitter. Mintz et. al. (2009) used a set of relations from Freebase as a distant source of supervision to learn to extract information from Wikipedia. Ridel et. al. (2010), Hoffmann et. al. (2011), and Surdeanu et. al. (2012) presented a series of models casting distant supervision as a multiple-instance learning problem (Dietterich et al., 1997).

Recent work has begun to address the challenge of noise in heuristically labeled training data generated by distant supervision, and proposed a variety of strategies for correcting erroneous labels. Takamatsu et al. (2012) present a generative model of the labeling process, which is used as a preprocessing step for improving the quality of labels before training relation extractors. Independently, Xu et. al. (2013) analyze a random sample of 1834 sentences from the New York Times, demonstrating that most entity pairs expressing a Freebase relation correspond to false negatives. They apply pseudo-relevance feedback to add missing entries in the knowledge base before applying the MultiR model (Hoffmann et al., 2011). Min et al. (2013) extend the MIML model of Surdeanu et. al. (2012) using a semi-supervised approach assuming a fixed proportion of true positives for each entity pair.

The Min et al. (2013) approach is perhaps the most closely related of the recent approaches for distant supervision. However, there are a number of key differences: (1) They impose a hard constraint on the proportion of true positive examples for each entity pair, whereas we jointly model relation extraction and missing data in the text and KB. (2) They only handle the case of missing information in the database and not in the text. (3) Their model, based on Surdeanu (2012), uses hard discriminative EM to tune parameters, whereas we use perceptron-style updates. (4) We evaluate various inference strategies

[1] also referred to as weakly supervised

for exact and approximate inference.

The issue of missing data has been extensively studied in the statistical literature (Little and Rubin, 1986; Gelman et al., 2003). Most methods for handling missing data assume that variables are *missing at random* (MAR): whether a variable is observed does not depend on its value. In situations where the MAR assumption is violated (for example distantly supervised information extraction), ignoring the missing data mechanism will introduce bias. In this case it is necessary to jointly model the process of interest (e.g. information extraction) in addition to the missing data mechanism.

Another line of related work is iterative semantic bootstrapping (Brin, 1999; Agichtein and Gravano, 2000). Carlson et. al. (2010) exploit constraints between relations to reduce semantic drift in the bootstrapping process; such constraints are potentially complementary to our approach of modeling missing data.

3 A Latent Variable Model for Distantly Supervised Relation Extraction

In this section we review the MultiR model (due to Hoffmann et. al. (2011)) for distant supervision in the context of extracting binary relations. This model is extended to handle missing data in Section 4. We focus on binary relations to keep discussions concrete; unary relation extraction is also possible.

Given a set of sentences, $\mathbf{s} = s_1, s_2, \ldots, s_n$, which mention a specific pair of entities (e_1 and e_2) our goal is to correctly predict which relation is mentioned in each sentence, or "NA" if none of the relations under consideration are mentioned. Unlike the standard supervised learning setup, we do not observe the latent sentence-level relation mention variables, $\mathbf{z} = z_1, z_2, \ldots, z_n$.[2] Instead we only observe *aggregate binary variables* for each relation, $\mathbf{d} = d_1, d_2, \ldots, d_k$, which indicate whether the proposition $r_j(e_1, e_2)$ is present in the database (Freebase). Of course the question which arises is: how do we relate the aggregate-level variables, d_j, to the sentence-level relation mentions, z_i? A sensible answer to this question is a simple deterministic-OR function. The deterministic-OR states that if there exists at least one i such that $z_i = j$, then $d_j = 1$. For example, if at least one sentence mentions that *"Barack Obama was born in Honolulu"*, then that fact is true in aggregate, if none of the sentences mentions the relation, then the fact is assumed false. The model also makes the converse assumption: if Freebase contains the relation BIRTHLOCATION(*Barack Obama*, *Honolulu*), then we must extract it from at least one sentence. A summary of this model, which is due to Hoffmann et. al. (2011), is presented in Figure 2.

3.1 Learning

To learn the parameters of the sentence-level relation mention classifier, θ, we maximize the likelihood of the facts observed in Freebase conditioned on the sentences in our text corpus:

$$\begin{aligned}\theta^* &= \arg\max_\theta P(\mathbf{d}|\mathbf{s};\theta) \\ &= \arg\max_\theta \prod_{e_1,e_2} \sum_{\mathbf{z}} P(\mathbf{z},\mathbf{d}|\mathbf{s};\theta)\end{aligned}$$

Here the conditional likelihood of a given entity pair is defined as follows:

$$\begin{aligned}P(\mathbf{z},\mathbf{d}|\mathbf{s};\theta) &= \prod_{i=1}^{n} \phi(z_i, s_i;\theta) \times \prod_{j=1}^{k} \omega(\mathbf{z}, d_j) \\ &= \prod_{i=1}^{n} e^{\theta \cdot f(z_i, s_i)} \times \prod_{j=1}^{k} \mathbf{1}_{\neg d_j \oplus \exists i: j = z_i}\end{aligned}$$

Where $\mathbf{1}_x$ is an indicator variable which takes the value 1 if x is true and 0 otherwise, the $\omega(\mathbf{z}, d_j)$ factors are hard constraints corresponding to the deterministic-OR function, and $f(z_i, s_i)$ is a vector of features extracted from sentence s_i and relation z_i.

An iterative gradient-ascent based approach is used to tune θ using a latent-variable perceptron-style additive update scheme (Collins, 2002; Liang et al., 2006; Zettlemoyer and Collins, 2007). The gradient of the conditional log likelihood, for a single pair of entities, e_1 and e_2, is as follows:[3]

$$\begin{aligned}\frac{\partial \log P(\mathbf{d}|\mathbf{s};\theta)}{\partial \theta} &= \mathbf{E}_{P(\mathbf{z}|s,d;\theta)} \left(\sum_j f(s_j, z_j) \right) \\ &\quad - \mathbf{E}_{P(\mathbf{z},\mathbf{d}|\mathbf{s};\theta)} \left(\sum_j f(s_j, z_j) \right)\end{aligned}$$

[2] These variables indicate which relation is mentioned between e_1 and e_2 in each sentence.

[3] For details see Koller and Friedman (2009), Chapter 20.

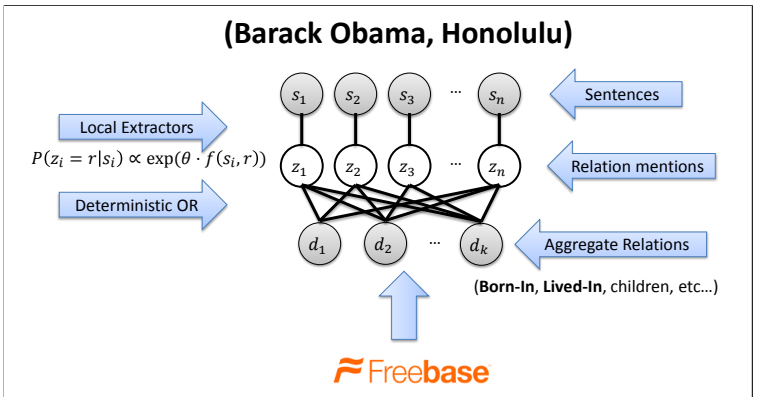

Figure 2: MultiR (Hoffmann et. al. 2011)

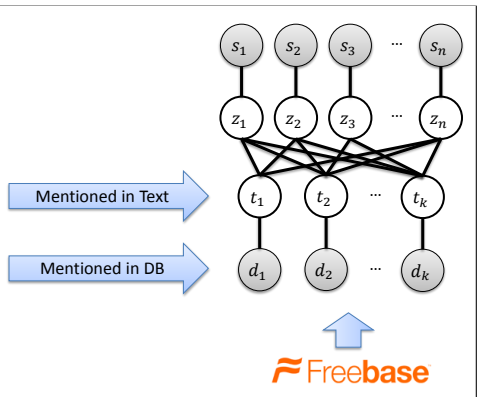

Figure 3: DNMAR

These expectations are too difficult to compute in practice, so instead they are approximated as maximizations. Computing this approximation to the gradient requires solving two inference problems corresponding to the two maximizations:

$$\mathbf{z}^{*\text{DB}} = \arg\max_{\mathbf{z}} P(\mathbf{z}|\mathbf{s}, \mathbf{d}; \theta)$$
$$\mathbf{z}^{*} = \arg\max_{\mathbf{z}} P(\mathbf{z}, \mathbf{d}|\mathbf{s}; \theta)$$

The MAP solution for the second term is easy to compute: because $\mathbf{d}$ and $\mathbf{z}$ are deterministically related, we can simply find the highest scoring relation, r, for each sentence, s_i, according to the sentence-level factors, ϕ, independently. The first term, is more difficult, however, as this requires finding the best assignment to the sentence-level hidden variables $\mathbf{z} = z_1 \ldots z_n$ conditioned on the observed sentences and facts in the database. Hoffmann et. al. (2011) show how this reduces to a well-known weighted edge cover problem which can be solved exactly in polynomial time.

4 Modeling Missing Data

The model presented in Section 3 makes two assumptions which correspond to hard constraints:

1. If a fact is not found in the database it cannot be mentioned in the text.

2. If a fact is in the database, it must be mentioned in at least one sentence.

These assumptions drive the learning, however if there is information missing from either the text or the database this leads to errors in the training data (false positives, and false negatives respectively).

In order to gracefully handle the problem of missing data, we propose to extend the model presented in Section 3 by splitting the aggregate level variables, $\mathbf{d}$, into two parts: $\mathbf{t}$ which represents whether a fact is mentioned in the text (in at least one sentence), and $\mathbf{d}'$ which represents whether the fact is mentioned in the database. We introduce pairwise potentials $\psi(t_j, d_j)$ which penalize disagreement between t_j and d_j, that is:

$$\psi(t_j, d_j) = \begin{cases} -\alpha_{\text{MIT}} & \text{if } t_j = 0 \text{ and } d_j = 1 \\ -\alpha_{\text{MID}} & \text{if } t_j = 1 \text{ and } d_j = 0 \\ 0 & \text{otherwise} \end{cases}$$

Where α_{MIT} (**M**issing **I**n **T**ext) and α_{MID} (**M**issing **I**n **D**atabase) are parameters of the model which can be understood as penalties for missing information in the text and database respectively. We refer to this model as DNMAR (for **D**istant Supervision with Data **N**ot **M**issing **A**t **R**andom). A graphical model representation is presented in Figure 3.

This model can be understood as relaxing the two hard constraints mentioned above into soft constraints. As we show in Section 7, simply relaxing these hard constraints into soft constraints and setting the two parameters α_{MIT}, and α_{MID} by hand on development data results in a large improvement to precision at comparable recall over MultiR on two different applications of distant supervision: binary relation extraction and named entity categorization.

Inference in this model becomes more challenging however, because the constrained inference problem no longer reduces to a weighted edge cover problem as before. In Section 5, we present an inference technique for the new model which is time and

memory efficient and almost always finds an exact MAP solution.

The learning proceeds analogously to what was described in section 3.1, with the exception that we now maximize over the additional aggregate-level hidden variables $\mathbf{t}$, which have been introduced. As before, MAP inference is a subroutine in learning, both for the unconstrained case corresponding to the second term (which is again trivial to compute), and for the constrained case which is more challenging as it no longer reduces to a weighted edge cover problem as before.

5 MAP Inference

The only difference in the new inference problem is the addition of $\mathbf{t}$; $\mathbf{z}$ and $\mathbf{t}$ are deterministically related, so we can simply find a MAP assignment to $\mathbf{z}$, from which $\mathbf{t}$ follows. The resulting inference problem can be viewed as optimization under soft constraints, where the objective includes terms for each fact *not* in Freebase which is extracted from the text: $-\alpha_{\text{MID}}$, and an effective reward for extracting a fact which *is* contained in Freebase: α_{MIT}.

The solution to the MAP inference problem is the value of $\mathbf{z}$ which maximizes the following objective:

$$\begin{aligned}\mathbf{z}^{*\text{DB}} &= \arg\max_{\mathbf{z}} P(\mathbf{z}|\mathbf{d};\theta,\alpha) \\ &= \arg\max_{\mathbf{z}} \sum_{i=1}^{n} \theta \cdot f(z_i, s_i) \qquad (1) \\ &\quad + \sum_{j=1}^{k} \left(\alpha_{\text{MIT}} \mathbf{1}_{d_j \wedge \exists i: j=z_i} - \alpha_{\text{MID}} \mathbf{1}_{\neg d_j \wedge \exists i: j=z_i}\right)\end{aligned}$$

Whether we choose to set the parameters α_{MIT} and α_{MID} to fixed values (Section 4), or incorporate side information through a missing data model (Section 6), inference becomes more challenging than in the model where facts observed in Freebase are treated as hard constraints (Section 3); the hard constraints are equivalent to setting $\alpha_{\text{MID}} = \alpha_{\text{MIT}} = \infty$.

We now present exact and approximate approaches to inference. Standard search methods such as A* and branch and bound have high computation and memory requirements and are therefore only feasible on problems with few variables; they are, however, guaranteed to find an optimal solution.[4] Approximate methods scale to large problem sizes, but we loose the guarantee of finding an optimal solution. After showing how to find guaranteed exact solutions for small problem sizes (e.g. up to 200 variables), we present an inference algorithm based on local search which is empirically shown to find optimal solutions in almost every case by comparing its solutions to those found by A*.

5.1 A* Search

We cast exact MAP inference in the DNMAR model as an application of A* search. Each partial hypothesis, h, in the search space corresponds to a partial assignment of the first m variables in $\mathbf{z}$; to expand a hypothesis, we generate k new hypotheses, where k is the total number of relations. Each new hypothesis h' contains the same partial assignment to $z_1, \ldots, z_m$ as h, with each h' having a different value of $z_{m+1} = r$.

A* operates by maintaining a priority queue of hypotheses to expand, with each hypothesis' priority determined by an admissible heuristic. The heuristic represents an upper bound on the score of the best solution with h's partial variable assignment under the objective from Equation 1. In general, a tighter upper bound corresponds to a better heuristic and faster solutions. To upper bound our objective, we start with the $\phi(z_i, s_i)$ factors from the partial assignment. Unassigned variables ($i > k$), are set to their maximum possible value, $z_i = \max_r \phi(r, s_i)$ independently. Next to account for the effect the aggregate $\psi(t_j, d_j)$ factors on the unassigned variables, we consider independently changing each unassigned z_i variable for each $\psi(t_j, d_j)$ factor to improve the overall score. This approach can lead to inconsistencies, but provides us with a good upper bound for the best possible solution with a partial assignment to $z_1, \ldots, z_k$.

5.2 Local Search

While A* is guaranteed to find an exact solution, its time and memory requirements prohibit use on large problems involving many variables. As a more scalable alternative we propose a greedy hill climbing method (Russell et al., 1996), which starts with a full assignment to $\mathbf{z}$, and repeatedly moves to the best neighboring solution $\mathbf{z}'$ according to the objective in

[4] Each entity pair defines an inference problem where the number of variables is equal to the number of sentences which mention the pair.

Equation 1. The neighborhood of $\mathbf{z}$ is defined by a set of *search operators*. If none of the neighboring solutions has a higher score, then we have reached a (local) maximum at which point the algorithm terminates with the current solution which may or may not correspond to a global maximum. This process is repeated using a number of *random restarts*, and the best local maximum is returned as the solution.

Search Operators: We start with a standard search operator, which considers changing each relation-mention variable, z_i, individually to maximize the overall score. At each iteration, all z_is are considered, and the one which produces the largest improvement to the overall score is changed to form the neighboring solution, $\mathbf{z}'$. Unfortunately, this definition of the solution neighborhood is prone to poor local optima because it is often required to traverse many low scoring states before changing one of the aggregate variables, t_j, and achieving a higher score from the associated aggregate factor, $\psi(t_j, d_j)$. For example, consider a case where the proposition $r(e_1, e_2)$ is not in Freebase, but is mentioned many times in the text, and imagine the current solution contains no mention $z_i = r$. Any neighboring solution which assigns a mention to r will include the penalty α_{MID}, which could outweigh the benefit from changing any individual z_i to r: $\phi(r, s_i) - \phi(z_i, s_i)$. If multiple mentions were changed to r however, together they could outweigh the penalty for extracting a fact not in Freebase, and produce an overall higher score.

To avoid the problem of getting stuck in local optima, we propose an additional search operator which considers changing *all* variables, z_i, which are currently assigned to a specific relation r, to a new relation r', resulting in an additional $(k-1)^2$ possible neighbors, in addition to the $n \times (k-1)$ neighbors which come from the standard search operator. This aggregate-level search operator allows for more global moves which help to avoid local optima, similar to the type-level sampling approach for MCMC (Liang et al., 2010).

At each iteration, we consider all $n \times (k-1) + (k-1)^2$ possible neighboring solutions generated by both search operators, and pick the one with biggest overall improvement, or terminate the algorithm if no improvements can be made over the current solution. 20 random restarts were used for each inference problem. We found this approach to almost always find an optimal solution. In over 100,000 problems with 200 or fewer variables from the New York Times dataset used in Section 7, an optimal solution was missed in only 3 cases which was verified by comparing against optimal solutions found using A*. Without including the aggregate-level search operator, local search almost always gets stuck in a local maximum.

6 Incorporating Side Information

In Section 4, we relaxed the hard constraints made by MultiR, which allows for missing information in either the text or database, enabling errors in the distantly supervised training data to be naturally corrected as a side-effect of learning. We made the simplifying assumption, however, that all facts are equally likely to be missing from the text or database, which is encoded in the choice of 2 fixed parameters α_{MIT}, and α_{MID}. Is it possible to improve performance by incorporating side information in the form of a missing data model (Little and Rubin, 1986), taking into account how likely each fact is to be observed in the text and the database conditioned on its truth value? In our setting, the missing data model corresponds to choosing the values of α_{MIT} and α_{MID} dynamically based on the entities and relations involved.

Popular Entities: Consider two entities: Barack Obama, the 44th president of the United States, and Donald Parry, a professional rugby league footballer of the 1980s.[5] Since Obama is much more well-known than Parry, we wouldn't be very surprised to see information missing from Freebase about Parry, but it would seem odd if true propositions were missing about Obama.

We can encode these intuitions by choosing entity-specific values of α_{MID}:

$$\alpha_{\text{MID}}^{(e_1,e_2)} = -\gamma \min\left(c(e_1), c(e_2)\right)$$

where $c(e_i)$ is the number of times e_i appears in Freebase, which is used as an estimate of its coverage.

Well Aligned Relations: Given that a pair of entities, e_1 and e_2, participating in a Freebase relation,

[5] http://en.wikipedia.org/wiki/Donald_Parry

r, appear together in a sentence s_i, the chance that s_i expresses r varies greatly depending on r. For example, if a sentence mentions a pair of entities which participate in both the COUNTRYCAPITOL relation and the LOCATIONCONTAINS relation (for example Moscow and Russia), it is more likely that the a random sentence will express LOCATIONCONTAINS than COUNTRYCAPITOL.

We can encode this preference for matching certain relations over others by setting α^r_{MIT} on a per-relation basis. We choose a different value of α^r_{MIT} for each relation based on quick inspection of the data, and estimating the number of true positives. Relations such as *contains*, *place_lived*, and *nationality* which contain a large number of true positive matches are assigned a large value of $\alpha^r_{\text{MIT}} = \gamma_{\text{large}}$, those with a medium number such as *capitol*, *place_of_death* and *administrative_divisions* were assigned a medium value γ_{medium}, and those relations with few matches were assigned a small value γ_{small}. These 3 parameters were tuned on held out development data.

7 Experiments

In Section 5, we presented a scalable approach to inference in the DNMAR model which almost always finds an optimal solution. Of course the real question is: does modeling missing data improve performance at extracting information from text? In this section we present experimental results showing large improvements in both precision and recall on two distantly supervised learning tasks: binary relation extraction and named entity categorization.

7.1 Binary Relation Extraction

For the sake of comparison to previous work we evaluate performance on the New York Times text, features and Freebase relations developed by Riedel et. al. (2010) which was also used by Hoffmann et. al. (2011). This dataset is constructed by extracting named entities from 1.8 million New York Times articles, which are then match against entities in Freebase. Sentences which contain pairs of entities participating in one or more relations are then used as training examples for those relations. The sentence-level features include word sequences appearing in context with the pair of entities, in addition to part of speech sequences, and dependency paths from the Malt parser (Nivre et al., 2004).

7.1.1 Baseline

To evaluate the effect of modeling missing data in distant supervision, we compare against the MultiR model for distant supervision (Hoffmann et al., 2011), a state of the art approach for binary relation extraction which is the most similar previous work, and models facts in Freebase as hard constraints disallowing the possibility of missing information in either the text or the database. To make our experiment as controlled as possible and rule-out the possibility of differences in performance due to implementation details, we compare against our own re-implementation of MultiR which reproduces Hoffmann et. al.'s performance, and shares as much code as possible with the DNMAR model.

7.1.2 Experimental Setup

We evaluate binary relation extraction using two evaluations. We first evaluate on a sentence-level extraction task using a manually annotated dataset provided by Hoffmann et. al. (2011).[6] This dataset consists of sentences paired with human judgments on whether each expresses a specific relation. Secondly, we perform an automatic evaluation which compares propositions extracted from text against held-out data from Freebase.

7.1.3 Results

Sentential Extraction: Figure 4 presents precision and recall curves for the sentence-level relation extraction task on the same manually annotated data presented by Hoffmann et. al. (2011). By explicitly modeling the possibility of missing information in both the text and the database we achieve a 17% increase in area under the precision recall curve. Incorporating additional side information in the form of a missing data model, as described in Section 6, produces even better performance, yielding a 27% increase over the baseline in area under the curve. We also compare against the system described by Xu et. al. (2013) (hereinafter called Xu13). To do this, we trained our implementation of MultiR on

[6] http://raphaelhoffmann.com/mr/

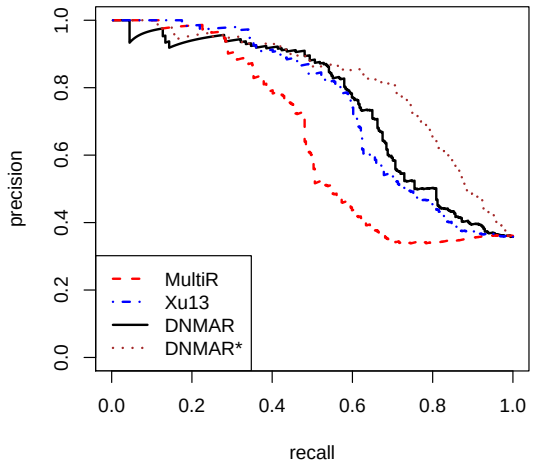

Figure 4: Overall precision and Recall at the sentence-level extraction task comparing against human judgments. DNMAR* incorporates side-information as discussed in Section 6.

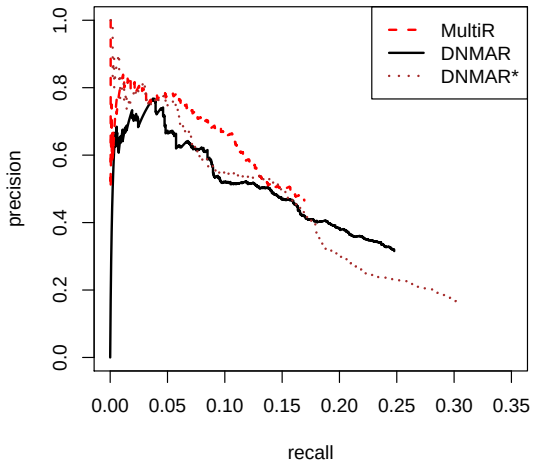

Figure 5: Aggregate-level automatic evaluation comparing against held-out data from Freebase. DNMAR* incorporates side-information as discussed in Section 6.

the labels predicted by their Pseudo-relevance Feedback model. [7] The differences between each pair of systems, except DNMAR and Xu13[8], is significant with p-value less than 0.05 according to a paired t-test assuming a normal distribution.

Per-relation precision and recall curves are presented in Figure 6. For certain relations, for instance */location/us_state/capital*, there simply isn't enough overlap between the information contained in Freebase and facts mentioned in the text to learn anything useful. For these relations, entity pair matches are unlikely to actually express the relation; for instance, in the following sentence from the data:

NHPF , which has its **Louisiana** office in **Baton Rouge** , gets the funds ...

although Baton Rouge is the capital of Louisiana, the */location/us_state/capital* relation is not expressed in this sentence. Another interesting observation which we can make from Figure 6, is that the benefit from modeling missing data varies from one relation to another. Some relations, for instance */people/person/place_of_birth*, have relatively good coverage in both Freebase and the text, and therefore we do not see as much gain from modeling missing data. Other relations, such as */location/location/contains*, and */people/person/place_lived* have poorer coverage making our missing data model very beneficial.

Aggregate Extraction: Following previous work, we evaluate precision and recall against held-out data from Freebase in Figure 5. As mentioned by Mintz et. al. (2009), this automatic evaluation underestimates precision because many facts correctly extracted from the text are missing in the database and therefore judged as incorrect. Riedel et. al. (2013) further argues that this evaluation is biased because frequent entity pairs are more likely to contain facts in Freebase, so systems which rank extractions involving popular entities higher will achieve better performance independently of how accurate their predictions are. Indeed in Figure 5 we see that the precision of our system which models missing data is generally lower than the system which assumes no data is missing from Freebase, although we do roughly double the recall. By better modeling

[7] We thank Wei Xu for making this data available.

[8] DNMAR has a 1.3% increase in AUC over Xu13, though this difference is not significant according to a paired t-test. DNMAR* achieves a 10% increase in AUC over Xu13 which is significant.

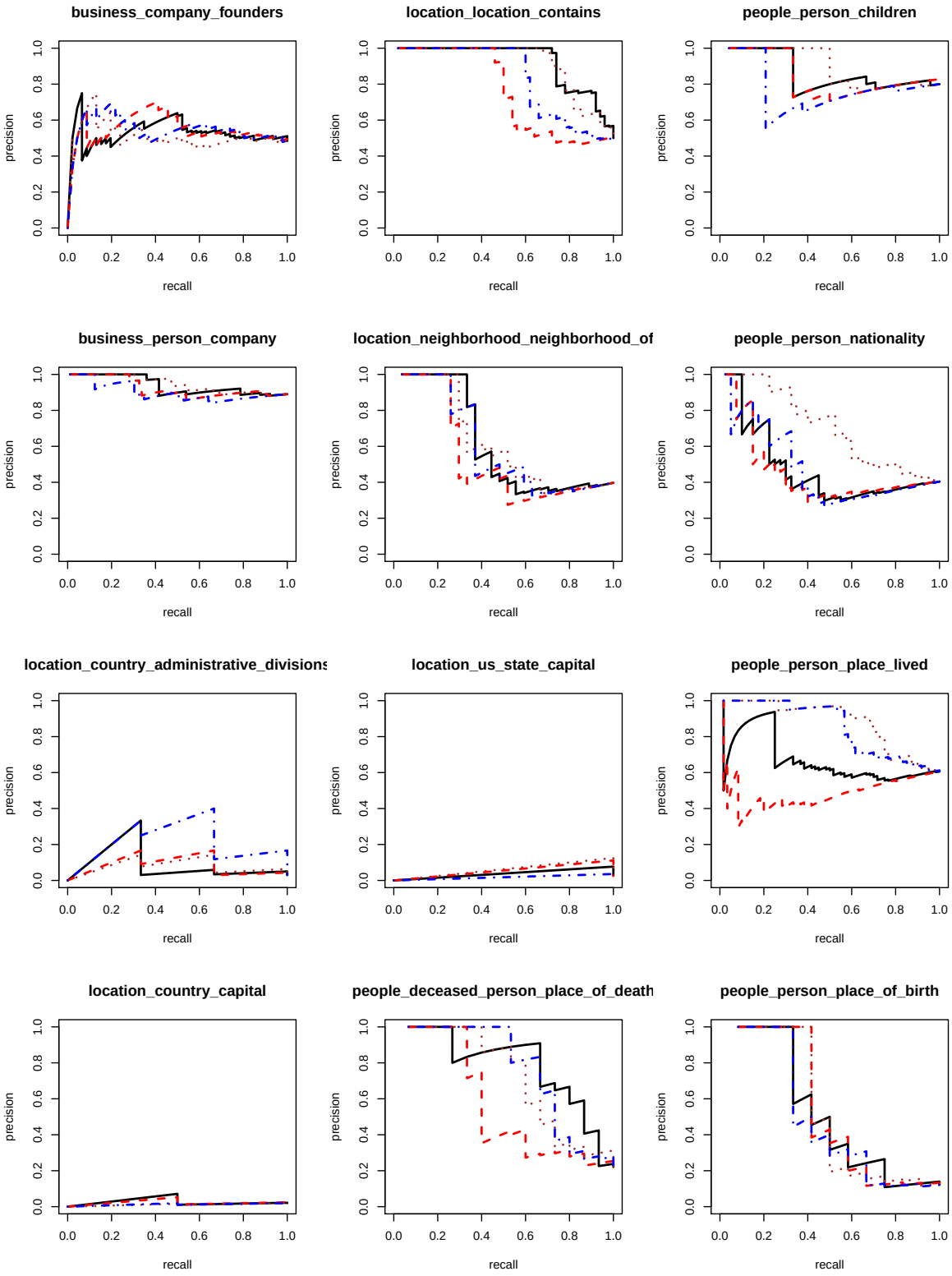

Figure 6: Per-relation precision and recall on the sentence-level relation extraction task. The dashed line corresponds to MultiR, DNMAR is the solid line, and DNMAR*, which incorporates side-information, is represented by the dotted line.

missing data we achieve lower precision on this automatic held-out evaluation as the system using hard constraints is explicitly trained to predict facts which occur in Freebase (not those which are mentioned in the text but unlikely to appear in the database).

7.2 Named Entity Categorization

As mentioned previously, the problem of missing data in distant (weak) supervision is a very general issue; so far we have investigated this problem in the context of extracting binary relations using distant supervision. We now turn to the problem of weakly supervised named entity recognition (Collins and Singer, 1999; Talukdar and Pereira, 2010).

7.2.1 Experimental Setup

To demonstrate the effect of modeling missing data in the distantly supervised named entity categorization task, we adapt the MultiR and DNMAR models to the Twitter named entity categorization dataset which was presented by Ritter et. al. (2011). The models described so far are applied unchanged: rather than modeling a set of relations in Freebase between a pair of entities, e_1 and e_2, we now model a set of possible Freebase categories associated with a single entity e. This is a natural extension of distant supervision from binary to unary relations. The unlabeled data and features described by Ritter et. al. (2011) are used for training the model, and their manually annotated Twitter named entity dataset is used for evaluation.

7.2.2 Results

Precision and recall at weakly supervised named entity categorization comparing MultiR against DNMAR is presented in Figure 7. We observe substantial improvement in precision at comparable recall by explicitly modeling the possibility of missing information in the text and database. The missing data model leads to a 107% increase in area under the precision-recall curve (from 0.16 to 0.34), but still falls short of the results presented by Ritter et. al. (2011). Intuitively this makes sense, because the model used by Ritter et. al. is based on latent Dirichlet allocation which is better suited to this highly ambiguous unary relation data.

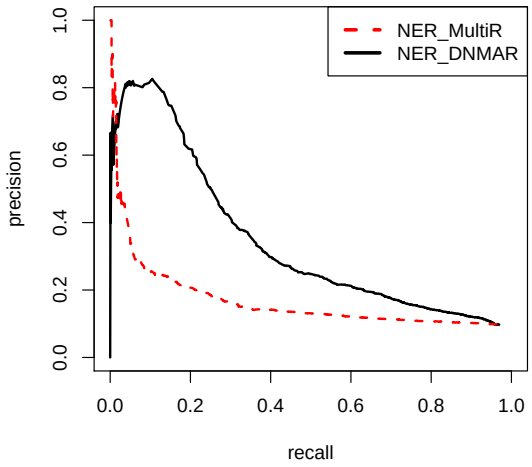

Figure 7: Precision and Recall at the named entity categorization task

8 Conclusions

In this paper we have investigated the problem of missing data in distant supervision; we introduced a joint model of information extraction and missing data which relaxes the hard constraints used in previous work to generate heuristic labels, and provides a natural way to incorporate side information through a missing data model. Efficient inference breaks in the new model, so we presented an approach based on A* search which is guaranteed to find exact solutions, however exact inference is not computationally tractable for large problems. To address the challenge of large problem sizes, we proposed a scalable inference algorithm based on local search, which includes a set of aggregate search operators allowing for long-distance jumps in the solution space to avoid local maxima; this approach was experimentally demonstrated to find exact solutions in almost every case. Finally we evaluated the performance of our model on the tasks of binary relation extraction and named entity categorization showing large performance gains in each case.

In future work we would like to apply our approach to modeling missing data to additional models, for instance the model of Surdeanu et. al. (2012) and Ritter et. al. (2011), and also explore new missing data models.

Acknowledgements

The authors would like to thank Dan Weld, Chris Quirk, Raphael Hoffmann and the anonymous reviewers for helpful comments. Thanks to Wei Xu for providing data. This research was supported in part by ONR grant N00014-11-1-0294, DARPA contract FA8750-09- C-0179, a gift from Google, a gift from Vulcan Inc., and carried out at the University of Washington's Turing Center.

References

Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*.

Edward Benson, Aria Haghighi, and Regina Barzilay. 2011. Event discovery in social media feeds. In *Proceedings of ACL*.

Sergey Brin. 1999. Extracting patterns and relations from the world wide web. In *The World Wide Web and Databases*.

Razvan Bunescu and Raymond Mooney. 2007. Learning to extract relations from the web using minimal supervision. In *ACL*.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka Jr, and Tom M Mitchell. 2010. Toward an architecture for never-ending language learning. In *Proceedings of AAAI*.

Michael Collins and Yoram Singer. 1999. Unsupervised models for named entity classification. In *EMNLP*.

Michael Collins. 2002. Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithms. In *Proceedings of EMNLP*.

Mark Craven, Johan Kumlien, et al. 1999. Constructing biological knowledge bases by extracting information from text sources. In *ISMB*.

Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. 2003. *Bayesian data analysis*. CRC press.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of ACL-HLT*.

D. Koller and N. Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

Percy Liang, Alexandre Bouchard-Côté, Dan Klein, and Ben Taskar. 2006. An end-to-end discriminative approach to machine translation. In *Proceedings of ACL*.

Percy Liang, Michael I Jordan, and Dan Klein. 2010. Type-based mcmc. In *Proceedings of ACL*.

Roderick J A Little and Donald B Rubin. 1986. *Statistical analysis with missing data*. John Wiley & Sons, Inc., New York, NY, USA.

Bonan Min, Ralph Grishman, Li Wan, Chang Wang, and David Gondek. 2013. Distant supervision for relation extraction with an incomplete knowledge base. In *Proceedings of NAACL-HLT*.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of ACL-IJCNLP*.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2004. Memory-based dependency parsing. In *Proceedings of CoNLL*.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of ECML/PKDD*.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M Marlin. 2013. Relation extraction with matrix factorization and universal schemas. In *Proceedings of NAACL-HLT*.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. *Proceedings of EMNLP*.

Stuart J. Russell, Peter Norvig, John F. Candy, Jitendra M. Malik, and Douglas D. Edwards. 1996. *Artificial intelligence: a modern approach*.

Benjamin Snyder and Regina Barzilay. 2007. Database-text alignment via structured multilabel classification. In *Proceedings of IJCAI*.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of EMNLP-Conll*.

Shingo Takamatsu, Issei Sato, and Hiroshi Nakagawa. 2012. Reducing wrong labels in distant supervision for relation extraction. In *Proceedings ACL*.

Partha Pratim Talukdar and Fernando Pereira. 2010. Experiments in graph-based semi-supervised learning methods for class-instance acquisition. In *Proceedings of ACL*.

Fei Wu and Daniel S. Weld. 2007. Autonomously semantifying wikipedia. In *Proceedings of CIKM*.

Wei Xu, Raphael Hoffmann Le Zhao, and Ralph Grishman. 2013. Filling knowledge base gaps for distant supervision of relation extraction. In *Proceedings of ACL*.

Luke S. Zettlemoyer and Michael Collins. 2007. Online learning of relaxed ccg grammars for parsing to logical form. In *EMNLP-CoNLL*.