Learning Rules from Incomplete Examples via Observation Models

Janardhan Rao Doppa, Mohammad NasrEsfahani, Mohammad S. Sorower, Jed Irvine
Thomas G. Dietterich, **Xiaoli Fern**, and **Prasad Tadepalli**
School of EECS, Oregon State University
Corvallis, OR 97330, USA
{doppa,nasresfm,sorower,irvine,tgd,xfern,tadepall}@cs.orst.edu

Abstract

We study the problem of learning general rules from concrete facts extracted from natural data sources such as the newspaper stories and medical histories. Natural data sources present two challenges to automated learning, namely, *radical incompleteness* and *systematic bias*. In previous work we proposed an approach that combines simultaneous learning of multiple predictive rules with differential scoring of evidence based on implicit observation models to address the above problems. In this paper, we further evaluate our approach empirically on natural datasets based on both textual and non-textual sources. We present a theoretical analysis that elucidates our approach and explains the empirical results. [1]

1 Introduction

Learning common sense knowledge in the form of rules by reading from natural texts has long been a dream of AI [Guha and Lenat, 1990]. This problem presents an opportunity to exploit the long strides of research progress made in natural language processing and machine learning in recent years [Nahm and Mooney, 2000; Carlson *et al.*, 2010; Schoenmackers *et al.*, 2010].

Unfortunately there are two major obstacles to fully realizing the dream of robust learning of general rules from natural sources. First, natural data sources such as texts and medical histories are *radically incomplete* — only a tiny fraction of all true facts are ever mentioned. More importantly, natural sources are *systematically biased* in what is mentioned. In particular, news stories emphasize newsworthiness, which correlates with rarity or novelty, sometimes referred as "the man bites dog phenomenon." For example, consider the following sentence in a real news story:

"*Ahmed Said Khadr, an Egyptian-born Canadian, was killed last October in Pakistan.*"

Presumably, the phrase "Egyptian-born" was considered important by the reporter because it violates the expectation that most Canadians are born in Canada. The birth place would most likely have been omitted if it was Canada.

In previous work, learning from incomplete examples or partial assignments has been studied under noise-free settings in the probably approximately correct learning framework. The goal is to learn an approximation of a function that has a small error with respect to the training distribution from incompletely described examples. It has been shown that the sample complexity of finite hypothesis spaces remains the same under incomplete examples as under complete examples [Khardon and Roth, 1999]. Further, when the hypothesis space obeys certain conditions such as "shallow monotonicity," the problem of learning from incomplete examples polynomially reduces to that of learning from complete examples [Michael, 2009]. In fact, the same learning algorithm can be used after the missing data is completed in a way that guarantees consistency with the target function. This approach is validated on an extensive study of sentence completion tasks on a natural dataset [Michael and Valiant, 2008].

Our approach to learning from incomplete examples extends the above work in multiple directions. First, we learn multiple rules for several predicates simultaneously and use the rules to complete the missing data to improve learning. We call this approach "multiple predicate bootstrapping (MPB)" [Doppa *et al.*, 2010]. Second, we adopt the above approaches to noisy observations and systematic bias.

Our main solution to deal with systematic bias is to differentially score the evidence for rules based on a presumed model of observation. In the "missing at random" (MAR) model [Little and Rubin, 1987; Jaeger, 2006], data is omitted based only on information that is already mentioned. In this case, conservative scoring of evidence, where rules are only evaluated when all relevant data is present, gives an unbiased estimate of the rule correctness. In the *novelty mention model*, a special case of "missing not at random" (MNAR) model and illustrated by the above Egyptian-born Canadian example, data is mentioned with a higher probability if it cannot be inferred from the remaining data. We show that under this model, aggressive scoring of rules, where we count evidence against a rule only if it contradicts the rule regardless of how the missing information transpires, gives a better approximation to the accuracy of the rule. Our empirical results compare favorably to baselines such as EM and Structural EM and are consistent with the theoretical predictions.

[1] This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL) under Contract No. FA8750-09-C-0179.

2 Multiple-Predicate Bootstrapping

Algorithm 1 Multiple-Predicate Bootstrapping (MPB)

```
Input: $\mathcal{D}_I$ = Incomplete training examples, $\mathcal{M}$ = Implicit mention model, $\tau$ = support threshold, $\theta$ = confidence threshold
Output: set of learned rules $\mathcal{R}$
repeat
    LEARN RULES: $\mathcal{R} = \phi$
    for each hypothesized rule $r$ do
        compute support $\tau_r$ and confidence $\theta_r$ of the rule $r$
          using $\mathcal{D}_I$ and implicit mention model $\mathcal{M}$
        if $\tau_r > \tau$ and $\theta_r > \theta$ then $\mathcal{R} = \mathcal{R} \cup \{r\}$
    end for
    IMPUTE MISSING DATA:
    for each missing fact $f_m \in \mathcal{D}_I$ do
        Predict $f_m$ using the most-confident applicable rule
          $r \in \mathcal{R}$
        if $f_m$ is predicted then $\mathcal{D}_I = \mathcal{D}_I - \{f_m\}$
    end for
until convergence
return the set of learned rules $\mathcal{R}$
```

Our algorithmic approach, called “Multiple-Predicate Bootstrapping,” (MPB) is inspired by several lines of work including co-training [Blum and Mitchell, 1998], multitask learning [Caruana, 1997], coupled semi-supervised learning [Carlson *et al.*, 2010], and self-training [Yarowsky, 1995]. It simultaneously learns a set of rules for each predicate in the domain given other predicates and then applies the learned rules to impute missing facts in the data. This is repeated until no new fact can be added. Following the data mining literature, we evaluate each rule using two measures: support and confidence. The support of a rule is measured by the number of examples that satisfy the body of the rule. The higher the support, the more statistical evidence we have on the predictive accuracy of the rule. In order to use a rule to impute facts, we require its support to be greater than a *support threshold*. We measure the confidence of a rule as the ratio of the number of records that satisfy both body and head of the rule to the number that satisfy the body, which estimates the conditional probability of the head of the rule given its body.

We use a relational data mining algorithm called FARMER [Nijssen and Kok, 2003] for learning rules. FARMER systematically searches the space of possible rules up to a fixed depth d (candidate rules) whose support and confidence exceed the given thresholds using depth first search. Its advantages over other rule learning systems such as FOIL are that a) it can learn redundant rules, which is useful with incomplete data; and b) it has the flexibility to vary the depth for search efficiency. Given multiple learned rules that are applicable to a given instance, we use the “most confident” one to make predictions. The overall algorithm is summarized in Algorithm 1.

Implicit Mention Models. We address the problem of systematic bias by adapting the scoring function for the hypothesized rules according to a presumed *implicit* mention model.

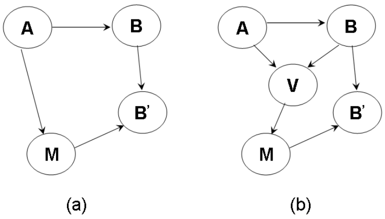


Figure 1: Bayes nets for data generation using (a) Random Mention Model (b) Novelty Mention Model. $A \Rightarrow B$ is a rule, M is a random variable that represents the fact that B is mentioned, B' indicates the observed value of B and random variable V denotes the violation of rule.

We now discuss two specific mention models and two methods for scoring evidence for rules.

Random Mention Model (RMM): This is equivalent to the Missing At Random (MAR) statistical model. In this model, it is assumed that facts are mentioned based on other known facts but not based on missing facts. For example, a doctor might omit a test if some other tests come out negative. A Bayesian network that illustrates this case is shown in Figure 1(a). B' is equal to B if M is true.

Novelty Mention Model (NMM): In this model, facts that are not entailed by the previously mentioned facts and prior knowledge are more likely to be mentioned. This is a special case of Missing Not At Random (MNAR) statistical model, since whether a fact is missing depends on the value of the fact itself as illustrated in Figure 1(b). In the novelty mention model, consider a α-general rule (i.e., a rule with confidence $\geq \alpha$) $A \Rightarrow B$, B will be mentioned with higher probability when the rule is violated, i.e., $P(M|V) > P(M|\neg V)$. Note that for rules that are not α-general, the facts entailed by these rules will not be missing because they are not considered α-general predictable due to the lack of generality of the rules. This model more closely captures the citizenship-birth place example, since whether or not the birth place of a person is mentioned depends on the birth place and other mentioned facts of the person such as the citizenship.

Inspired by the two types of mention models, we propose two different ways of scoring rules. We use the following notation to define our rule scoring functions. Each literal may be either true, false or unknown. We write $n(A = t, B = f, C = u)$ to be the count of examples where A is true, B is false and C is unknown. For brevity we write A for $A = t$. The `Support` of a rule $A \Rightarrow B$ is defined as the number of examples in which A is known to be true i.e., $n(A)$, for both conservative and aggressive scoring.

In *conservative scoring*, evidence is counted in favor of a rule only when all facts relevant to determining the truth value of the rule are actually known. The confidence of the rule in this case is defined as follows:

$$p_c(A \Rightarrow B) = \frac{n(A, B)}{n(A, B \neq u)} \qquad (1)$$

In *aggressive scoring*, a fact is counted as evidence for a rule if the rule's premise is satisfied and the conclusion is not contradicted. The confidence of a rule is defined as follows:

$$p_a(A \Rightarrow B) = \frac{n(A,B) + n(A,B=u)}{n(A)} \tag{2}$$

For example, consider the text "Khadr, a Canadian citizen, was killed in Pakistan". For conservative scoring, it is counted as neither supporting nor contradicting the rule `citizen(X,Y)` $\Rightarrow$ `bornIn(X,Y)`, as we are not told that `bornIn(Khadr,Canada)`. In contrast, it is counted as supporting the rule `citizen(Y)` $\Rightarrow$ `bornIn(Y)` for aggressive scoring because, adding `bornIn(Canada)` supports the rule without contradicting the available evidence.

3 Analysis of Implicit Mention Models

This section analyzes aggressive and conservative scoring of data generated using different mention models.

Consider a rule $A \Rightarrow B$. Figure 1 shows the Bayes nets that explains the data generation process in the random and novelty mention models. Let S be the support set of the rule $A \Rightarrow B$, i.e., the set of examples where A is true. Let $p(r)$ be the true confidence of the rule r, i.e., the conditional probability of B given A. Let $\hat{p}_c(r)$ and $\hat{p}_a(r)$ denote the conservative and aggressive estimates of confidence of the rule r.

Theorem 1. *If the data is generated by the random mention model then $\hat{p}_c(r)$ is an unbiased estimate and $\hat{p}_a(r)$ is an overestimate of the true confidence of rule $p(r)$.*

Proof. Conservative scoring estimates the confidence of the rule from only a subset of S where B is not missing.

$$\hat{p}_c(r) = \frac{|S|\,p(r)P(M|A)}{|S|\,P(M|A)} = p(r) \tag{3}$$

Therefore, $\hat{p}_c$ is a unbiased estimate of the true confidence.

Aggressive scoring deterministically imputes the missing value of B such that it satisfies the hypothesized rule.

$$\begin{aligned}
\hat{p}_a(r) &= \frac{|S|\,p(r)P(M|A) + |S|\,(1-P(M|A))}{|S|} \\
&= p(r)P(M|A) + (1-P(M|A)) \qquad (4) \\
&= p(r) + (1-P(M|A))\,(1-p(r)) \\
&\geq p(r)
\end{aligned}$$

Therefore, $\hat{p}_a(r)$ overestimates the confidence of the rule. The bias of $\hat{p}_a(r)$ increases with decreased $P(M|A)$. □

Theorem 2. *If the data is generated by the random mention model, then the true ranking order of rules is preserved with both conservative and aggressive scoring.*

Proof. It is enough to show that the ordering is preserved for any two rules r_1 and r_2 that predict the value of the same variable. Without loss of generality, let $p(r_1) > p(r_2)$. From (3), $\hat{p}_c(r_1) > \hat{p}_c(r_2)$. Therefore, order is preserved with conservative scoring.

$$\begin{aligned}
&p(r_1) > p(r_2) \\
&\Rightarrow p(r_1)P(M|A) + (1-P(M|A)) \\
&> p(r_2)P(M|A) + (1-P(M|A)) \\
&\Rightarrow \hat{p}_a(r_1) > \hat{p}_a(r_2) \quad (From \;\; (4)\;)
\end{aligned}$$

Thus, aggressive scoring also preserves the ordering. □

Theorem 3. *If the data is generated by the novelty mention model, then $\hat{p}_c(r)$ is an underestimate and $\hat{p}_a(r)$ is an overestimate of true confidence of the rule $p(r)$.*

Proof. In what follows V stands for a random variable that represents a violation of a confident rule in predicting B. If V is true, according to the novelty model, B has a higher probability of being mentioned. Hence $P(M|V) > P(M|\neg V)$.

$$\begin{aligned}
\hat{p}_c(r) &= \frac{|S|\,p(r)P(M|\neg V)}{|S|\,p(r)P(M|\neg V) + |S|\,(1-p(r))P(M|V)} \\
&= \frac{p(r)P(M|\neg V)}{p(r)P(M|\neg V) + (1-p(r))P(M|V)}
\end{aligned}$$

To compare this with $p(r)$, we estimate the odds:

$$\begin{aligned}
&\frac{\hat{p}_c(r)}{1-\hat{p}_c(r)} \\
&= \frac{p(r)P(M|\neg V)}{(1-p(r))P(M|V)} \\
&= true\ odds \times \frac{P(M|\neg V)}{P(M|V)} \\
&< true\ odds
\end{aligned}$$

Since in the novelty mention model $P(M|\neg V) < P(M|V)$, $\hat{p}_c(r)$ underestimates $p(r)$ and significantly so if $P(M|\neg V) << P(M|V)$.

It is easy to show that for aggressive scoring, we have:

$$\begin{aligned}
\hat{p}_a(r) &= \frac{|S|\,p(r) + |S|\,(1-p(r))\,(1-P(M|V))}{|S|} \\
&= p(r) + (1-p(r))\,(1-P(M|V)) \qquad (5) \\
&\geq p(r)
\end{aligned}$$

□

Therefore, similar to the random mention model, $\hat{p}_a(r)$ overestimates the true confidence of the rule $p(r)$. However, when the novelty mention model is strongly at play, i.e., $P(M|V) \approx 1$, it provides a good estimate of $p(r)$.

Theorem 4. *If the data is generated by the novelty mention model, then the true ranking order of the rules is preserved with aggressive scoring.*

Proof. We first show that the ordering is preserved for any α-general rules r_1 and r_2 where $p(r_1) > p(r_2)$.

$$\begin{aligned}
&p(r_1) > p(r_2) \\
&\Rightarrow p(r_1)P(M|V) > p(r_2)P(M|V) \\
&\Rightarrow p(r_1)P(M|V) + (1-P(M|V)) \\
&> p(r_2)P(M|V) + (1-P(M|V)) \\
&\Rightarrow \hat{p}_a(r_1) > \hat{p}_a(r_2) \quad (From \;\; (5)\;)
\end{aligned}$$

We then compare an α-general rule r_1 with a rule r_2 that is not α-general. For r_1, $\hat{p}_a(r_1) \geq p(r_1)$ over-estimates the confidence based on Theorem 3. For r_2, because no data is missing, $\hat{p}_a(r_2) = p(r_2)$ is an unbiased estimate of $p(r_2)$. $\hat{p}_a(r_1) \geq p(r_1) > p(r_2) = \hat{p}_a(r2)$. Thus, r_1 will be correctly ranked higher than r_2 by aggressive scoring. Finally

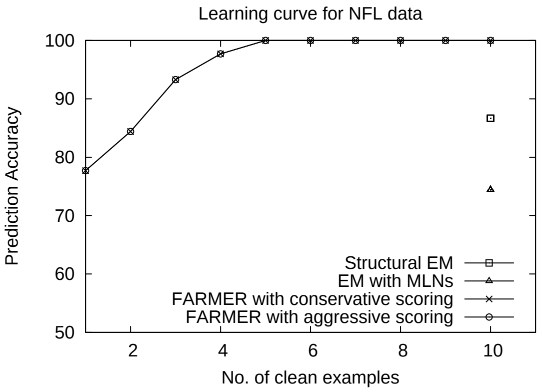


Figure 4: Results of NFL domain: no. of clean examples vs. prediction accuracy

consider two rules that are both not α-general, because there is no missing data, aggressive scoring provides unbiased estimate of the confidences and preserves their rank order. □

It is interesting to note that while conservative scoring preserves the ranking order of α-general rules, it can potentially reverse the order of an α-general rule with a rule that is not α-general. This is because conservative scoring correctly estimates the confidence of rules that are not α-general but underestimates the confidence of the α-general rules.

4 Experimental Results

In this section, we describe our experimental results with both synthetic and natural datasets and carefully analyze them.

Synthetic Experiments. To test our analysis of implicit mention models, we perform experiments on synthetic data generated using different missing mechanisms, i.e., RMM and NMM. We use the UCI database *SPECT Heart*, which describes diagnosing of Single Proton Emission Computed Tomography (SPECT) images[2]. This database contains 267 examples with 23 binary features extracted from the SPECT image sets (patients). A 70% / 30% split of the data is created for training and testing respectively. We generate two different synthetic versions based on RMM and NMM missing mechanisms (see Figure 1). We first learn a set of deterministic rules from the training data, and then retain those that have a confidence of 80% or more. These rules are then used to create training and testing data with varying levels of missingness. For NMM, if some rule is violated, then the consequent is always mentioned. If no rule is violated, then the consequent is omitted based on the missingness level. We evaluate the learning algorithms on the test data generated by the same mention model that generates its training data. Experiments are performed with different levels of missingness in both training and testing data and we report the accuracies (averaged over all attributes) with which the missing data is predicted correctly w.r.t the gold standard data. The averaged results over 10 different versions of the generated data are reported (see Table 1 and Table 2).

Baselines: We compare the results of our Multiple-Predicate Bootstrapping (MPB) approach using the relational data mining algorithm FARMER [Nijssen and Kok, 2003] with Structural EM (SEM) [Friedman, 1998] and EM implemented using Markov Logic Networks(MLNs). Structural EM learns both structure and parameters of the Bayes net from incomplete data, and then the learned Bayes net is used to make predictions on the test dataset. To run EM, we initialize an MLN with high support rules learned from incomplete data and learn its weights generatively [Richardson and Domingos, 2006].

Analysis of Results: For the RMM data, both conservative and aggressive scoring perform equally well (see Figure 2(a)), which was expected based on Theorem 2. Since the ranking order of rules is the same in both scoring methods, they make the same prediction by picking the same rule that is applicable. SEM performs better than both conservative and aggressive when missingness in training data is small, i.e., 0.2 and 0.4 (see □'s in Figure 2(b)), but conservative/aggressive scoring significantly outperforms SEM when missingness in training data is large, i.e., 0.6 and 0.8 (see ◊'s in Figure 2(b)). Performance of all the algorithms decreases as the percentage of missingness in training data increases.

For the NMM data, aggressive scoring significantly outperforms conservative scoring (see Figure 3(a)) which is consistent with our analysis in Theorem 4. Since the novelty mention model was strongly at play, i.e., $P(M|V) \approx 1$, aggressive scoring provides a very good estimate of the true confidence of the rules, resulting in excellent performance. Aggressive scoring significantly outperforms SEM when the missingness in data is tolerable, i.e., 0.2, 0.4 and 0.6. However, all algorithms including SEM perform poorly with exceedingly high missingness, i.e., 0.8. Note that, although our analysis of implicit mention models is for the simple case where only the head of the rule can be missing, our synthetic data were generated for a more difficult problem where the body of the rule could be missing as well.

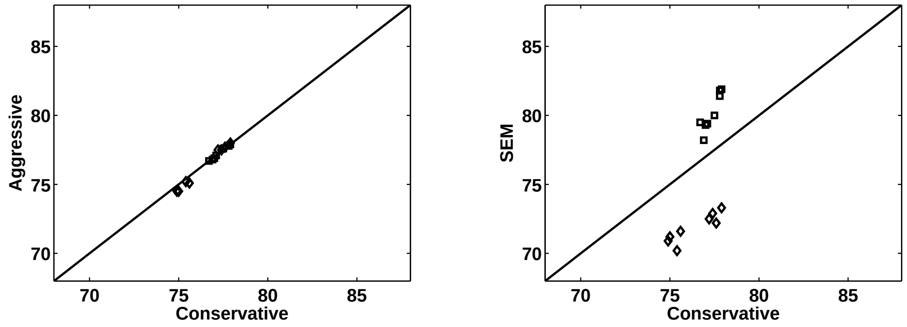
Experiments with Real data. We also performed experiments on two real world domains: 1. NFL data, 2. Birthplace-Citizenship data. We used data extracted by BBN's information extraction system on a LDC (Linguistic Data Consortium) training corpus of 110 NFL sports articles and 248 news stories related to the topics of people, organizations and relationships respectively.

For the NFL domain, the following predicates were provided for each game with natural interpretations: `gameWinner`, `gameLoser`, `homeTeam`, `awayTeam`, `gameTeamScore`, and `teamInGame`. A test set of 100 examples was used, to evaluate the performance of the learned rules based on the accuracy of predicting the missing facts w.r.t the ground truth. We observed that most of the input extractions are noisy and inconsistent, that makes the problem of rule learning even harder. These inconsistent examples are due to co-reference errors, e.g., the extractor may not realize that two mentions of the same team in a football article are in fact the same. We want to use this insight while scoring each rule. To address this we learned integrity constraints from a

[2] http://archive.ics.uci.edu/ml/datasets/SPECT+Heart

		Testing											
	Missing %	0.2			0.4			0.6			0.8		
		CONS	AGGR	SEM	CONS	AGGR	SEM	CONS	AGGR	SEM	CONS	AGGR	SEM
Training	0.2	77.8	77.8	**81.8**	77.9	77.9	**81.9**	77.8	77.8	**81.4**	77.5	77.6	**80.0**
	0.4	76.7	76.7	**79.5**	77.1	77.1	**79.4**	77.0	76.9	**79.3**	76.9	76.8	**78.2**
	0.6	77.6	**77.7**	72.2	77.9	**78.0**	73.3	77.4	**77.5**	72.9	77.2	**77.5**	72.5
	0.8	**75.4**	75.2	70.2	**75.6**	75.1	71.6	**75.0**	74.5	71.2	**74.9**	74.5	70.9

Table 1: Accuracy results of synthetic experiments with Random Mention Model (RMM) data



(a)

(b)

Figure 2: Accuracy for Random Mention Model (RMM) data : (a) Conservative vs. Aggressive (b) Conservative vs. SEM

small number of complete examples, and applied the learned integrity constraints to generate consistent versions of the inconsistent examples (e.g., by deleting a literal) in all possible ways. Finally, we scored the rules against these ‘corrected’ examples with a lower weight $\gamma(< 1)$. The prediction accuracy of the learned rules is reported as a function of number of clean examples (see Figure 4). The results of this approach are compared with Structural EM (SEM) and EM algorithm using Markov Logic Networks (MLNs) initialized with the highest support rules learned from incomplete data. We then perform generative weight learning of MLNs via EM. We use Lazy MCSAT to do MAP inference for predicting the missing facts in the test set. For both SEM and EM, we use the ground-truth (instead of the learned) integrity constraints to correct the noisy examples and hence, there is only one point for each of them in Figure 4.

Analysis of Results: As we can see in Figure 4, both conservative and aggressive scoring significantly outperform both SEM and EM. Since the NFL domain is deterministic, i.e., $\forall\ r,\ p(r) = 1$, and similar to RMM data, both conservative and aggressive scoring perform equally. We observed that once we learn the true integrity constraints from the clean examples, conservative scoring exactly learns the ground truth rules and aggressive scoring learns a few other spurious rules as well. However, the ground-truth rules are ranked higher than the spurious rules based on the estimated confidences and therefore, the spurious rules do not degrade the performance. Similar to the results on the synthetic data, SEM does not perform very well when the data is radically incomplete.

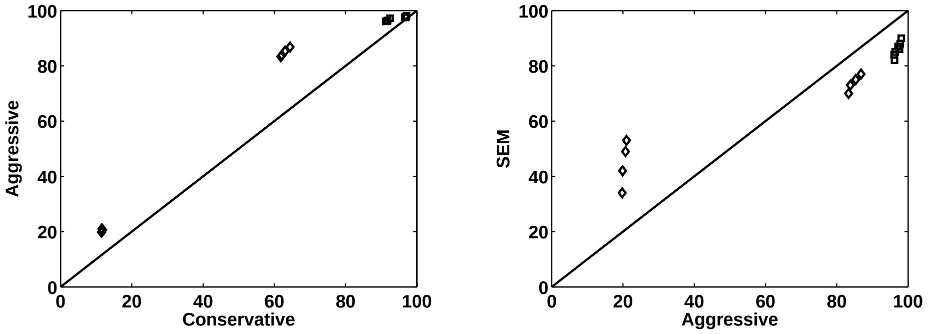
Birthplace-Citizenship data: Manual analysis of this training corpus revealed that the birth place of some person is only mentioned 23 times in the 248 documents. Moreover, in 14 of the 23 mentions, the information violates the default rule `citizen(Y)` $\Rightarrow$ `bornIn(Y)`. Since the data matches the assumption of aggressive scoring, it is expected to learn the correct rule. However, our extracted data was highly noisy and inaccurate. More specifically, the extracted data had 479 examples which mentioned only the citizenship of a person, 4 examples where both birth place and citizenship were mentioned out of which only 1 example violated the default rule. Therefore, confidence of the rule `citizen(Y)` $\Rightarrow$ `bornIn(Y)` was 0.75 based on conservative scoring and 0.9967 based on aggressive scoring. Since we used a confidence threshold of 0.8 for all our experiments, only aggressive scoring learned the correct rule. We also did this experiment with EM using MLNs and found that its performance was similar to conservative scoring.

5 Conclusions and Future Work

We motivated and studied the problem of learning from natural data sources which presents the dual challenges of radical incompleteness and systematic bias. Our solutions to these problems consist of bootstrapping from learning of multiple relations and scoring the rules or hypotheses differently based on an assumed mention model. Our experimental results validate the usefulness of differential scoring of rules and show that our approach can outperform other state-of-the-art methods such as Structural EM and EM. Our theoretical analysis gives insights into why our approach works, and point to some future directions. One of the open questions is the analysis of multiple-predicate bootstrapping and the conditions under which it works. Another avenue of future research is

		Testing											
	Missing %	0.2			0.4			0.6			0.8		
		Cons	Aggr	Sem	Cons	Aggr	Sem	Cons	Aggr	Sem	Cons	Aggr	Sem
Training	0.2	97.1	**98.1**	90.0	96.8	**97.8**	88.0	96.7	**97.5**	87.0	96.9	**97.6**	86.0
	0.4	92.5	**97.2**	87.0	91.8	**96.4**	85.0	91.3	**96.1**	84.0	91.7	**96.2**	82.0
	0.6	64.4	**86.8**	77.0	63.0	**85.3**	75.0	62.1	**83.8**	73.0	61.8	**83.3**	70.0
	0.8	11.6	21.0	**53.0**	11.8	20.7	**49.0**	11.6	19.9	**42.0**	11.5	19.8	**34.0**

Table 2: Accuracy results of synthetic experiments with Novelty Mention Model (NMM) data



(a)

(b)

Figure 3: Accuracy for Novelty Mention Model data: (a) Aggressive vs. Conservative (b) SEM vs. Aggressive

the use of explicit mention models and their use in learning from radically incomplete and biased examples. Exploration of the relationship of this work to the statistical models of missing data is another fruitful direction.

References

[Blum and Mitchell, 1998] Avrim Blum and Tom Mitchell. Combining Labeled and Unlabeled Data with Co-Training. In *COLT*, pages 92–100, 1998.

[Carlson *et al.*, 2010] Andrew Carlson, Justin Betteridge, Richard C. Wang, Estevam R. Hruschka Jr., and Tom M. Mitchell. Coupled Semi-Supervised Learning for Information Extraction. In *WSDM*, pages 101–110, 2010.

[Caruana, 1997] R. Caruana. Multitask Learning: A Knowledge-Based Source of Inductive Bias. *MLJ*, 28:41–75, 1997.

[Doppa *et al.*, 2010] Janardhan Rao Doppa, Mohammad NasrEsfahani, Mohammad S. Sorower, Thomas G. Dietterich, Xiaoli Fern, and Prasad Tadepalli. Towards Learning Rules from Natural Texts. In *Workshop on Formalisms and Methodology for Learning by Reading (NAACL-2010)*, pages 70–77, 2010.

[Friedman, 1998] Nir Friedman. The Bayesian Structural EM Algorithm. In *UAI*, pages 129–138, 1998.

[Guha and Lenat, 1990] R. V. Guha and D. B. Lenat. Cyc: A Medterm Report. *AI Magazine*, 11(3), 1990.

[Jaeger, 2006] M. Jaeger. The AI & M Procedure for Learning from Incomplete data. In *UAI*, pages 225–232, 2006.

[Khardon and Roth, 1999] Roni Khardon and Dan Roth. Learning to Reason with a Restricted View. *Machine Learning*, 35(2):95–116, 1999.

[Little and Rubin, 1987] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley-Interscience, NY, 1987.

[Michael and Valiant, 2008] Loizos Michael and Leslie G. Valiant. A First Experimental Demonstration of Massive Knowledge Infusion. In *KR*, pages 378–389, 2008.

[Michael, 2009] Loizos Michael. Reading Between the Lines. In *IJCAI*, pages 1525–1530, 2009.

[Nahm and Mooney, 2000] Un Yong Nahm and Raymond J. Mooney. A Mutually Beneficial Integration of Data Mining and Information Extraction. In *AAAI*, pages 627–632, July 2000.

[Nijssen and Kok, 2003] Siegfried Nijssen and Joost N. Kok. Efficient Frequent Query Discovery in FARMER. In *PKDD*, pages 350–362, 2003.

[Richardson and Domingos, 2006] Matthew Richardson and Pedro Domingos. Markov Logic Networks. *Machine Learning*, 62(1-2):107–136, 2006.

[Schoenmackers *et al.*, 2010] Stefan Schoenmackers, Jesse Davis, Oren Etzioni, and Daniel S. Weld. Learning First-Order Horn Clauses from Web Text. In *EMNLP*, pages 1088–1098, 2010.

[Yarowsky, 1995] D. Yarowsky. Unsupervised Word Sense Disambiguation rivaling Supervised Methods. In *ACL*, pages 189–196, 1995.