JMLR: Workshop and Conference Proceedings 20 (2011) 1–16 Asian Conference on Machine Learning

Learning Rules from Incomplete Examples via Implicit Mention Models

Janardhan Rao Doppa DOPPA@EECS.OREGONSTATE.EDU
Mohammad Shahed Sorower SOROWER@EECS.OREGONSTATE.EDU
Mohammad Nasresfahani NASRESFM@EECS.OREGONSTATE.EDU
Jed Irvine IRVINE@EECS.OREGONSTATE.EDU
Walker Orr ORR@EECS.OREGONSTATE.EDU
Thomas G. Dietterich TGD@EECS.OREGONSTATE.EDU
Xiaoli Fern XFERN@EECS.OREGONSTATE.EDU
Prasad Tadepalli TADEPALL@EECS.OREGONSTATE.EDU
School of Electrical Engineering and Computer Science
Oregon State University

Editor: Chun-Nan Hsu and Wee Sun Lee

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 this paper, we propose an approach that combines simultaneous learning of multiple predictive rules with differential scoring of evidence which adapts to a presumed model of data generation. Learning multiple predicates simultaneously mitigates the problem of radical incompleteness, while the differential scoring would help reduce the effects of systematic bias. We evaluate our approach empirically on both textual and non-textual sources. We further present a theoretical analysis that elucidates our approach and explains the empirical results.

Keywords: missing data, rule learning, structured and relational data

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 as has been demonstrated in (Nahm and Mooney, 2000) and extended to web-scale in (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 to as the "man bites dog phenomenon."[1] 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. If Khadr was instead born in Canada, the phrase "Canadian-born" would most likely have been left out of the text because it is too obvious to mention given that he is a Canadian.

Learning from incomplete or missing data is studied in statistics under different models of missingness (Little and Rubin, 1987). Data is Missing Completely At Random (MCAR), when the missingness mechanism does not depend on any data at all, i.e., some data is omitted randomly with no attention to the content. A more general case is when data is Missing At Random (MAR). Here the missingness mechanism only depends on the data that has been observed. For example, a doctor might choose not to do certain tests if the observed data makes the test irrelevant or unnecessary. The most general case is when the data is Missing Not At Random (MNAR), where data is omitted based on the values of the missing data themselves. Among other things, this represents the case when a reporter omits the facts that are too obvious to mention in the news story given what has already been said. While there are statistical tests to determine if a given situation is MCAR or MAR from data alone, there are no such tests to distinguish MNAR from MAR in general (Little and Rubin, 1987).

A widely used approach in both MCAR and MAR is based on expectation maximization (EM), which is a two step iterative process. In the Expectation step or the E-step, the missing data is imputed based on their expected values given the observed data. In the Maximization step or the M-step, parameters are found that maximize the likelihood of the data including the imputed missing data. EM usually converges in a small number of iterations to a locally optimal set of parameters (Dempster et al., 1977). In the more sophisticated Multiple Imputation (MI) framework, results of multiple imputations are combined in order to reduce the variance due to single imputation (Rubin, 1987; Schafer, 1999). However, these statistical approaches are mostly confined to parameter estimation and do not address the structure learning or rule learning. A notable exception is Structural EM (SEM), which learns the structure and parameters of a Bayesian network in the presence of incomplete data (Friedman, 1998). However, SEM does not take into account the systematic bias and gives poor results when data is generated by an MNAR process.

Learning from incomplete examples or partial assignments has also been studied under noise-free deterministic rule learning setting in the probably approximately correct (PAC) 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 specified examples. With an appropriate interpretation of the meaning of incompleteness, 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 target rule is deterministic, and sensing does not corrupt the data, the problem of learning from incomplete examples polynomially reduces to that of learning from complete examples (Michael, 2009).

1. "When a dog bites a man, that is not news, because it happens so often. But if a man bites a dog, that is news," attributed to John Bogart of New York Sun among others.

In fact, any algorithm that learns from complete data can be used after the missing data is completed in a way that guarantees consistency with the target function. Interestingly, this result applies independent of the missingness process, which means that it is applicable to the general case of MNAR. This approach is validated in an extensive study of sentence completion tasks on a natural dataset (Michael and Valiant, 2008). At a high level, our work shares many of the features of (Michael, 2009) in scoring the evidence and learning multiple rules simultaneously to compensate for incomplete data. Our empirical results are based on completing the missing data in multiple relational domains some of which are extracted from text. We also support our results using a different kind of probabilistic analysis from the PAC analysis of (Michael and Valiant, 2008).

Our main solution to dealing with systematic bias is to differentially score the evidence for rules based on a presumed model of observation. In the MAR case, where data is omitted based only on information that is already mentioned, 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*, which is a special case of MNAR model, 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. We evaluate our approach in multiple textual and non-textual domains and show that it compares favorably to SEM.

2. Multiple-Predicate Bootstrapping

Algorithm 1 Multiple-Predicate Bootstrapping (MPB)

Input: $\mathcal{D}_I$ = Incomplete training examples, $\mathcal{M}$ = Implicit mention model, τ = support threshold, θ = confidence threshold
Output: set of learned rules $\mathcal{R}$

repeat
  Learn Rules: $\mathcal{R} = FARMER^*(M, \mathcal{D}_I, \tau, \theta)$
  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 for 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*. The confidence of a rule is defined to be the ratio of the number of records that satisfy both body and head of the rule to the number that satisfy the body, and represents an estimate of the conditional probability of the head of the rule given its body.

Within each iteration of MPB, we adapt a relational data mining algorithm called FARMER (Nijssen and Kok, 2003) for learning rules from incomplete data. FARMER systematically searches for all possible rules up to a fixed depth d (candidate rules) whose support and confidence exceed the given thresholds using depth first search. It has two main advantages over other rule learning systems such as FOIL (Quinlan, 1990). First, it can learn redundant rules. This is important in our setting where many of the predictive features may be missing. Learning many redundant rules allows the inference to proceed as much as possible. The second advantage is that it has the flexibility to vary the depth of the search which controls the efficiency of search and the complexity of rules. To handle missing data, our adapted version $FARMER^*$ measures the confidence of a rule either conservatively or aggressively according to the assumed implicit mention model.

At the end of this process we select all rules of a desired complexity that pass a support threshold and a confidence threshold. Given multiple learned rules that are applicable to a given instance, we only use the most confident one to make predictions. This would avoid making multiple conflicting predictions of the same attribute. The overall algorithm is summarized in Algorithm 1.

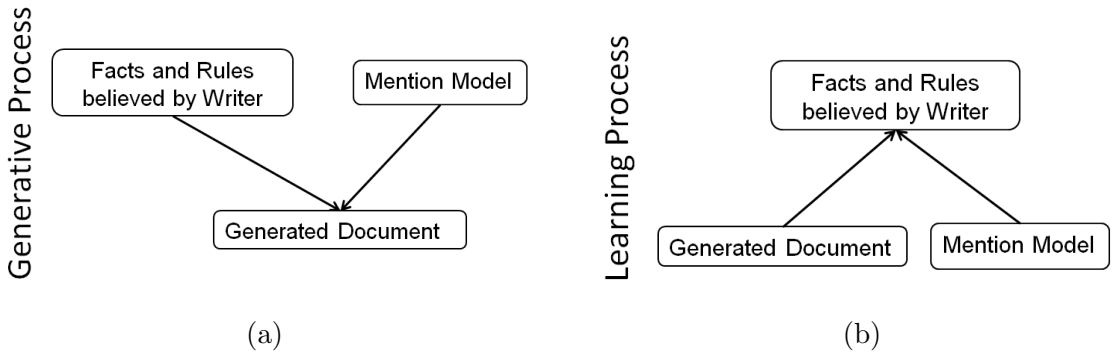


Figure 1: (a) Generative model for document generation process (b) Inverted model for learning process

Explicit Mention Model. It is instructive to consider a generative approach to solving the problem of learning from incomplete examples. In this approach, one would construct a probabilistic generative model we call "mention model," that captures what facts are mentioned and extracted by the programs from the text given the true facts about the world (see Figure 1(a)). Given some extracted facts, the learning agent would invert the

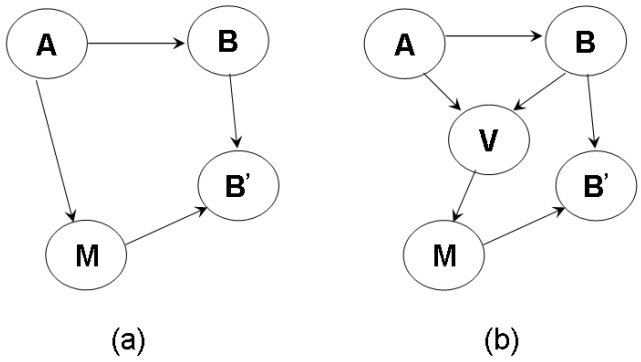


Figure 2: Bayes nets for data generation using (a) Random Mention Model and (b) Novelty Mention Model. $A \Rightarrow B$ is a rule, M is a random variable that represents the fact that B is mentioned, B' indicates the observed value of B and random variable V denotes violation of rule.

mention model to infer a distribution over sets of true facts (see Figure 1(b)). An inductive program could then be used to infer general rules from distributions over true facts. This approach has the advantages of being general and elegant. It is also flexible because the mention model is an explicit input to the learning algorithm. However, the flexibility comes with a price. The inference is highly intractable due to the need to marginalize over all possible mentioned and true sets of facts. Approximate inference will have to be used with its own errors and pitfalls. Instead of this computationally demanding approach, we describe a simpler method of adapting the learning algorithms directly to score and learn rules using an *implicit* mention model. This approach is similar in spirit to the work of (Michael, 2009) and extends it to the case of noisy data, multiple relations and probabilistic target rules.

Implicit Mention Models. Unlike in the maximum likelihood approach discussed above, our learners do not employ an explicit mention model. We address the problem of systematic bias by adapting the scoring function for the hypothesized rules according to a presumed *implicit* mention model. We now discuss two specific mention models and two methods for scoring evidence for rules which are inspired by them.

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 2(a). B' is equal to B if M is true. Here the node labeled M denotes a random variable that represents the fact that B is mentioned. B' indicates the observed value of B and is equal to B if M is true.

Novelty Mention Model (NMM): This model is a special case of the Missing Not At Random (MNAR) statistical model, where a fact is more likely to be omitted (mentioned)

if it is considered predictable (not predictable) based on other known facts and common knowledge. Specifically, we consider a fact predictable if its value can be correctly predicted by a highly confident rule (i.e., confidence $\geq \alpha$). We refer to such rules as being α-*general*. Given an α-general rule $A \Rightarrow B$, in NMM B will be mentioned with a higher probability when the rule is violated, i.e., $P(M|V) > P(M|\neg V)$. This is illustrated in Figure 2(b). For rules that are less than α-general, the facts entailed by these rules are not considered predictable, thus will not be missing under the novelty mention model. 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:

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

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. In contrast, it does not support the rule `killedIn(Y)` $\Rightarrow$ `citizen(Y)` because the rule directly contradicts the evidence, if it is known that `citizen` is a functional relationship.

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 2 shows the Bayes nets that explain 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.

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

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 confidence rank 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. ■

Under the novelty mention model, consider a α-general rule r: $A \Rightarrow B$. Let V stands for a random variable that represents a violation of the rule r. If V is true, according to the novelty model, B has a higher probability of being mentioned. Hence $P(M|V) > P(M|\neg V)$, where M denotes the random variable representing the fact that B is mentioned. Note that facts entailed by rules that are less than α-general will not be missed under the novelty mention model because they are not considered predictable.

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

Proof Under the novelty mention model, we have:

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

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 $P(M|\neg V) \ll P(M|V)$.

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

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

■

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 confidence rank order of 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 less than 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 consider two rules that are both not α-general. Because the conclusion for such rules is always mentioned according to the NMM, 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 less than α-general. This is because conservative scoring correctly estimates the confidence of rules that are less than α-general but underestimates the confidence of the α-general rules.

4. Experimental Results

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

4.1. Synthetic Experiments

To test our analysis of implicit mention models, we perform experiments on synthetic data generated using different missing data 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. We first learn a set of rules that have confidence of 80% or more from the complete training data. These rules are then used to create training and testing data with varying levels of missingness. For NMM, if a rule is violated, then its consequent is always mentioned (i.e., $p(M|V) = 1$). If a rule is not violated, then its consequent is omitted based on the missingness level ($p(M|\neg V)$. 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).

Baseline. We compare the results of our Multiple-Predicate Bootstrapping (MPB) approach with Structural EM (SEM) (Friedman, 1998). SEM directly learns from incomplete data by searching over the joint space of structures and parameters. At each search step, SEM either finds better parameters for the current structure (*parametric* EM step) or selects a new structure (*structural* EM step). It is shown that SEM converges to a local maximum (Friedman, 1998). Since the choice of initial model determines the convergence point (and

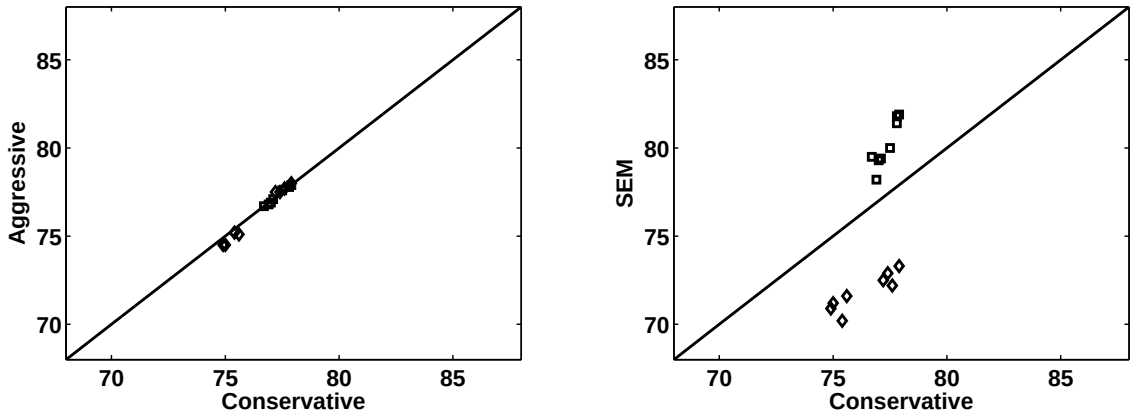
2. http://archive.ics.uci.edu/ml/datasets/SPECT+Heart

the quality of the learned model), we do 20 random restarts and compute the averaged results over these multiple runs of SEM.

Analysis of Results. For the RMM data, both conservative and aggressive scoring perform equally well (see Figure 3(a)), which was expected based on Theorem 2. Since the rank 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 scoring when missingness in the training data is small, i.e., 0.2 and 0.4 (see □'s in Figure 3(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 3(b)). Performance of all the algorithms decreases as the percentage of missingness in training data increases.

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

		Testing											
	Missing%	0.2			0.4			0.6			0.8		
		Con	Agg	Sem	Con	Agg	Sem	Con	Agg	Sem	Con	Agg	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



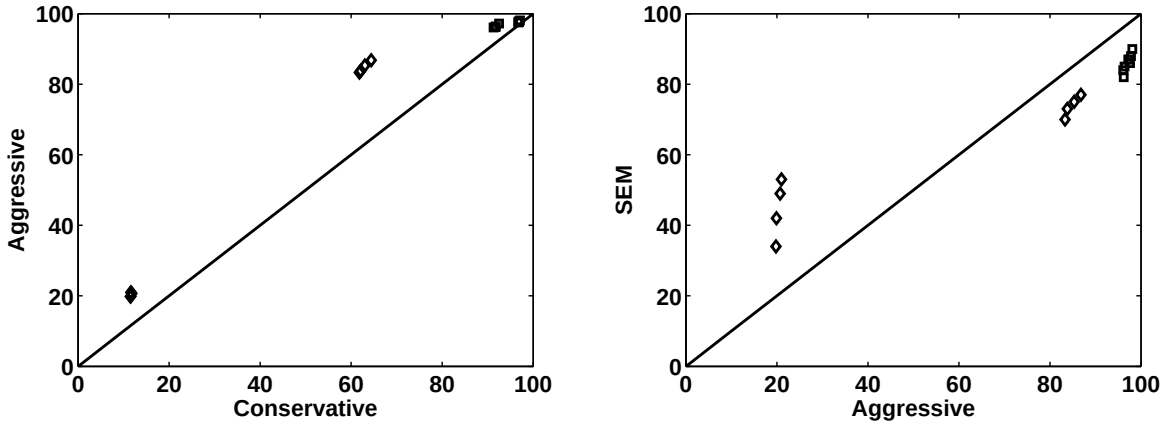
(a) (b)

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

For the NMM data, aggressive scoring significantly outperforms conservative scoring (see Figure 4(a)) which is consistent with our analysis in Theorem 4. Since the novelty mention model was strongly at play, i.e., $P(M|V) = 1$, aggressive scoring provides a very good

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

		Testing											
	Missing%	0.2			0.4			0.6			0.8		
		Con	Agg	Sem	Con	Agg	Sem	Con	Agg	Sem	Con	Agg	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**



(a) (b)

Figure 4: Accuracy for Novelty Mention Model data: (a) Aggressive vs. Conservative (b) SEM vs. Aggressive

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 very 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.

4.2. Experiments with Real data

We also performed experiments on three datasets extracted from news stories: **(a)** NFL games, **(b)** Birthplace-Citizenship data of people, **(c)** Somali ship-hijackings. We used data extracted by a state-of-the-art information extraction system from BBN technologies (Ramshaw et al., 2011; Freedman et al., 2010) applied to LDC (Linguistic Data Consortium) corpus of 110 NFL sports articles for **(a)**, 248 news stories related to the topics of

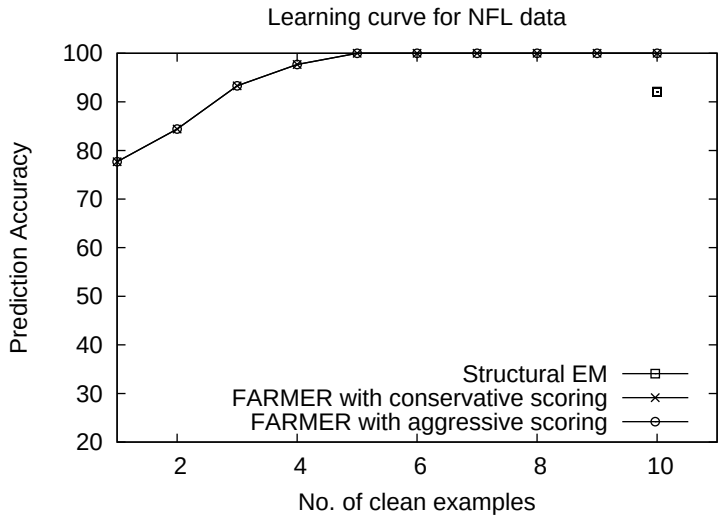


Figure 5: Results of NFL domain: no. of clean examples vs. prediction accuracy

people, organizations and relationships from ACE08 Evaluation corpus for **(b)**, and hand extractions from 101 news stories concerning ship hijacking incidents mentioned on the web site `coordination-maree-noire.eu` for **(c)**.

NFL domain. For the NFL domain, the following predicates were provided for each game with natural interpretations: `gameWinner`, `gameLoser`, `homeTeam`, `awayTeam`, `gameTeamScore`, and `teamInGame`. We manually extracted 55 games from news stories about the 2010 NFL season (which do not overlap with our training corpus) to serve as a test set. We observed that most of the input extractions are noisy and inconsistent, which makes the problem of rule learning even harder. These inconsistent examples are due to co-reference errors, e.g., the extractor does not realize that two mentions of the same team in a football article are in fact the same. To allow correcting the inconsistencies in the noisy data, we learned integrity constraints from a small number of complete and noise-free examples. We then applied the learned integrity constraints to generate consistent versions of the inconsistent examples (e.g., by deleting a literal) in all possible ways. Note that many of these examples are going to be factually incorrect, as we did not use the ground truth in correcting the examples. 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 5). The results of this approach are compared with Structural EM (SEM). For SEM, we used the ground-truth (instead of the learned) integrity constraints to correct the noisy examples and hence, there is only one point in Figure 5.

As we can see in Figure 5, both conservative and aggressive scoring significantly outperform SEM. 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 well. 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 performance. Similar to the results on the synthetic data, SEM does not perform very well when the data are radically incomplete.

Birthplace-Citizenship data. In this corpus, the birth place of a person is only mentioned 23 times in the 248 documents. In 14 of the 23 mentions, the mentioned information violates the default rule `citizen(Y)` $\Rightarrow$ `bornIn(Y)`. Since the data matches the assumption of aggressive scoring, we expect aggressive scoring to learn the correct rule. The extracted data had 583 examples where citizenship of a person was mentioned and 25 examples where birthplace was mentioned, 6 examples where both birthplace and citizenship was mentioned out of which 2 examples violated the default rule. The confidence of the rule `citizen(Y)` $\Rightarrow$ `bornIn(Y)` is 0.675 based on conservative scoring and 0.9931 based on aggressive scoring. According to Wikipedia the true confidence of this rule is > 0.97, which means that aggressive scoring achieves better probability estimate compared to conservative 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 SEM and found that its performance is similar to aggressive scoring.

Somali Ship-hijackings data. We manually extracted information about ship hijackings by Somali pirates from natural language documents on the web. From the 101 stories collected, we merged the information for each ship, resulting in a total of 35 summarized stories. We used 25 of them for training and the other 10 for testing. The data extracted consisted of the occurrences of 13 different kinds of events, e.g., attacked, captured, held, released, negotiations started, and so on. Our goal was to predict missing events from the mentioned events. For example, given that an article mentions a ship was captured, learned rules should infer that it was attacked. Rules may also predict events before they occur, such as predicting that after ransom negotiations have begun, a ransom will be paid. We provided some background knowledge in the form of hard-coded integrity constraints that allowed us to fill in some of the missing facts. For example, if a ship is released it is no longer held and vice versa. We experimented with both aggressive and conservative scoring and evaluated their performance on the test data. Our prediction accuracy with respect to the gold standard test set is 96.15% with aggressive scoring and 71.79% with conservative scoring. Both methods learned many good rules. Aggressive scoring did significantly better and extracted very few bad rules. The prediction performance of SEM (57.7%) was inferior to both scoring methods. We also experimented with the mentions of ownership country and flag country of hijacked ships from the manually-extracted stories. Similar to birthplace/citizenship case, the default rule is `ownership(Y)` $\Rightarrow$ `flag(Y)`. However, many ships fly a different flag than the country of ownership, which violates the default rule. There were 16 stories that mentioned both nationality of owner and nationality of flag, of which 14 violated the default rule. Since the data matches the assumptions of aggressive scoring, it learns the correct rule in this case. SEM performs similar to aggressive 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. Our theoretical analysis gives insights into why our approach works, and points 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 research is to explicitly represent the mention models and reason about them while learning from incomplete and biased examples. In (Sorower et al., 2011), we used Markov Logic Networks to explicitly represent a mention model in the form of weighted rules (Richardson and Domingos, 2006). The weights of the rules are also learned from the incomplete data showing the advantage of this approach. This approach is also more flexible in the sense that it is easy to consider the consequences of incorporating different mention models in the document generation process. More ambitiously, we could also consider a space of possible models and search them to fit the data. Understanding the flexibility vs. efficiency tradeoff between the explicit and the implicit approaches to learning from missing data appears to be a productive research direction.

6. Acknowledgements

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. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DARPA, the Air Force Research Laboratory (AFRL), or the US government. We would like to thank Linguistic Data Consortium (LDC) for providing the raw text and annotations, and BBN team for letting us use their extractions from raw text in our experiments.

References

Avrim Blum and Tom Mitchell. Combining Labeled and Unlabeled Data with Co-Training. In *Proceedings of International Conference on Learning Theory (COLT)*, pages 92–100, 1998.

Andrew Carlson, Justin Betteridge, Richard C. Wang, Estevam R. Hruschka Jr., and Tom M. Mitchell. Coupled Semi-Supervised Learning for Information Extraction. In *Proceedings of International Conference on Web Search and Data Mining (WSDM)*, pages 101–110, 2010.

R. Caruana. Multitask Learning: A Knowledge-Based Source of Inductive Bias. *Machine Learning Journal (MLJ)*, 28:41–75, 1997.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

M. Freedman, E. Loper, E. Boschee, and R. Weischedel. Empirical Studies in Learning to Read. In *Proceedings of Workshop on Formalisms and Methodology for Learning by Reading (NAACL-2010)*, pages 61–69, 2010.

Nir Friedman. The Bayesian Structural EM Algorithm. In *Proceedings of International Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 129–138, 1998.

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

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

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

Loizos Michael. Reading Between the Lines. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1525–1530, 2009.

Loizos Michael and Leslie G. Valiant. A First Experimental Demonstration of Massive Knowledge Infusion. In *Proceedings of International Conference on the Principles of Knowledge Representation and Reasoning (KR)*, pages 378–389, 2008.

Un Yong Nahm and Raymond J. Mooney. A Mutually Beneficial Integration of Data Mining and Information Extraction. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, pages 627–632, July 2000.

Siegfried Nijssen and Joost N. Kok. Efficient Frequent Query Discovery in FARMER. In *Proceedings of Principles on Knowledge Discovery from Databases (PKDD)*, pages 350–362, 2003.

J. Ross Quinlan. Learning Logical Definitions from Relations. *Machine Learning Journal (MLJ)*, 5:239–266, 1990.

L. Ramshaw, E. Boschee, M. Freedman, J. MacBride, R. Weischedel, and A.Zamanian. Serif language processing effective trainable language understanding. In Joseph Olive, Caitlin Christianson, and John McCary, editors, *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*. Springer, 2011.

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

D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley Series, NY, 1987.

J. L. Schafer. Multiple Imputation: A Primer. *Statistical Methods in Medical Research*, 8 (1):3, 1999.

Stefan Schoenmackers, Jesse Davis, Oren Etzioni, and Daniel S. Weld. Learning First-Order Horn Clauses from Web Text. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1088–1098, 2010.

Mohammad S. Sorower, Thomas G. Dietterich, Janardhan Rao Doppa, Xiaoli Fern, and Prasad Tadepalli. Inverting Grice's Maxims to Learn Rules from Natural Language Extractions. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2011.

D. Yarowsky. Unsupervised Word Sense Disambiguation rivaling Supervised Methods. In *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 189–196, 1995.