Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, Daniel S. Weld
Computer Science & Engineering
University of Washington
Seattle, WA 98195, USA
{raphaelh,clzhang,xiaoling,lsz,weld}@cs.washington.edu

Abstract

Information extraction (IE) holds the promise of generating a large-scale knowledge base from the Web's natural language text. Knowledge-based weak supervision, using structured data to heuristically label a training corpus, works towards this goal by enabling the automated learning of a potentially unbounded number of relation extractors. Recently, researchers have developed multi-instance learning algorithms to combat the noisy training data that can come from heuristic labeling, but their models assume relations are *disjoint* — for example they cannot extract the pair `Founded(Jobs, Apple)` and `CEO-of(Jobs, Apple)`.

This paper presents a novel approach for multi-instance learning with overlapping relations that combines a sentence-level extraction model with a simple, corpus-level component for aggregating the individual facts. We apply our model to learn extractors for NY Times text using weak supervision from Freebase. Experiments show that the approach runs quickly and yields surprising gains in accuracy, at both the aggregate and sentence level.

1 Introduction

Information-extraction (IE), the process of generating relational data from natural-language text, continues to gain attention. Many researchers dream of creating a large repository of high-quality extracted tuples, arguing that such a knowledge base could benefit many important tasks such as question answering and summarization. Most approaches to IE use supervised learning of relation-specific examples, which can achieve high precision and recall. Unfortunately, however, fully supervised methods are limited by the availability of training data and are unlikely to scale to the thousands of relations found on the Web.

A more promising approach, often called "weak" or "distant" supervision, creates its own training data by heuristically matching the contents of a database to corresponding text (Craven and Kumlien, 1999). For example, suppose that $r(e_1, e_2) =$ `Founded(Jobs,Apple)` is a ground tuple in the database and $s =$"Steve Jobs founded Apple, Inc." is a sentence containing synonyms for both $e_1 =$ `Jobs` and $e_2 =$ `Apple`, then s may be a natural language expression of the fact that $r(e_1, e_2)$ holds and could be a useful training example.

While weak supervision works well when the textual corpus is tightly aligned to the database contents (*e.g.*, matching Wikipedia infoboxes to associated articles (Hoffmann et al., 2010)), Riedel *et al.* (2010) observe that the heuristic leads to noisy data and poor extraction performance when the method is applied more broadly (*e.g.*, matching Freebase records to NY Times articles). To fix this problem they cast weak supervision as a form of multi-instance learning, assuming only that *at least one* of the sentences containing e_1 and e_2 are expressing $r(e_1, e_2)$, and their method yields a substantial improvement in extraction performance.

However, Riedel *et al.*'s model (like that of previous systems (Mintz et al., 2009)) assumes that *relations do not overlap* — there cannot exist two facts $r(e_1, e_2)$ and $q(e_1, e_2)$ that are both true for any pair of entities, e_1 and e_2. Unfortunately, this assumption is often violated;

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 541–550,
Portland, Oregon, June 19-24, 2011.

for example both `Founded(Jobs, Apple)` and `CEO-of(Jobs, Apple)` are clearly true. Indeed, 18.3% of the weak supervision facts in Freebase that match sentences in the NY Times 2007 corpus have overlapping relations.

This paper presents MULTIR, a novel model of weak supervision that makes the following contributions:

- MULTIR introduces a probabilistic, graphical model of multi-instance learning which handles overlapping relations.
- MULTIR also produces accurate sentence-level predictions, decoding individual sentences as well as making corpus-level extractions.
- MULTIR is computationally tractable. Inference reduces to weighted set cover, for which it uses a greedy approximation with worst case running time $O(|\mathrm{R}| \cdot |\mathrm{S}|)$ where R is the set of possible relations and S is largest set of sentences for any entity pair. In practice, MULTIR runs very quickly.
- We present experiments showing that MULTIR outperforms a reimplementation of Riedel *et al.* (2010)'s approach on both aggregate (corpus as a whole) and sentential extractions. Additional experiments characterize aspects of MULTIR's performance.

2 Weak Supervision from a Database

Given a corpus of text, we seek to extract facts about *entities*, such as the company `Apple` or the city `Boston`. A *ground fact* (or *relation instance*), is an expression $r(\mathbf{e})$ where r is a relation name, for example `Founded` or `CEO-of`, and $\mathbf{e} = e_1, \ldots, e_n$ is a list of entities.

An *entity mention* is a contiguous sequence of textual tokens denoting an entity. In this paper we assume that there is an *oracle* which can identify all entity mentions in a corpus, but the oracle doesn't normalize or disambiguate these mentions. We use $e_i \in E$ to denote both an entity and its name (*i.e.*, the tokens in its mention).

A *relation mention* is a sequence of text (including one or more entity mentions) which states that some ground fact $r(\mathbf{e})$ is true. For example, "Steve Ballmer, CEO of Microsoft, spoke recently at CES." contains three entity mentions as well as a relation mention for `CEO-of(Steve Ballmer, Microsoft)`. In this paper we restrict our attention to binary relations. Furthermore, we assume that both entity mentions appear as noun phrases in a single sentence.

The task of *aggregate extraction* takes two inputs, Σ, a set of sentences comprising the corpus, and an extraction model; as output it should produce a set of ground facts, I, such that each fact $r(\mathbf{e}) \in I$ is expressed somewhere in the corpus.

Sentential extraction takes the same input and likewise produces I, but in addition it also produces a function, $\Gamma : I \rightarrow \mathcal{P}(\Sigma)$, which identifies, for each $r(\mathbf{e}) \in I$, the set of sentences in Σ that contain a mention describing $r(\mathbf{e})$. In general, the corpus-level extraction problem is easier, since it need only make aggregate predictions, perhaps using corpus-wide statistics. In contrast, sentence-level extraction must justify each extraction with *every* sentence which expresses the fact.

The *knowledge-based weakly supervised learning* problem takes as input (1) Σ, a training corpus, (2) E, a set of entities mentioned in that corpus, (3) R, a set of relation names, and (4), Δ, a set of ground facts of relations in R. As output the learner produces an extraction model.

3 Modeling Overlapping Relations

We define an undirected graphical model that allows joint reasoning about aggregate (corpus-level) and sentence-level extraction decisions. Figure 1(a) shows the model in plate form.

3.1 Random Variables

There exists a connected component for each pair of entities $\mathbf{e} = (e_1, e_2) \in \mathrm{E} \times \mathrm{E}$ that models all of the extraction decisions for this pair. There is one Boolean output variable Y^r for each relation name $r \in \mathrm{R}$, which represents whether the ground fact $r(\mathbf{e})$ is true. Including this set of binary random variables enables our model to extract overlapping relations.

Let $S_{(e_1,e_2)} \subset \Sigma$ be the set of sentences which contain mentions of both of the entities. For each sentence $x_i \in S_{(e_1,e_2)}$ there exists a latent variable Z_i which ranges over the relation names $r \in R$ and,

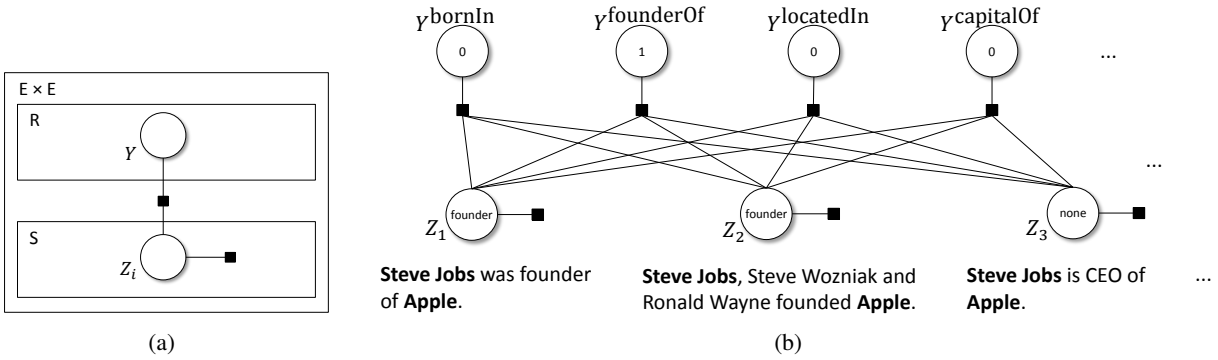

Figure 1: (a) Network structure depicted as plate model and (b) an example network instantiation for the pair of entities `Steve Jobs`, `Apple`.

importantly, also the distinct value `none`. Z_i should be assigned a value $r \in R$ only when x_i expresses the ground fact $r(\mathbf{e})$, thereby modeling sentence-level extraction.

Figure 1(b) shows an example instantiation of the model with four relation names and three sentences.

3.2 A Joint, Conditional Extraction Model

We use a conditional probability model that defines a joint distribution over all of the extraction random variables defined above. The model is undirected and includes repeated factors for making sentence level predictions as well as globals factors for aggregating these choices.

For each entity pair $\mathbf{e} = (e_1, e_2)$, define $\mathbf{x}$ to be a vector concatenating the individual sentences $x_i \in S_{(e_1,e_2)}$, $\mathbf{Y}$ to be vector of binary Y^r random variables, one for each $r \in R$, and $\mathbf{Z}$ to be the vector of Z_i variables, one for each sentence x_i. Our conditional extraction model is defined as follows:

$$p(\mathbf{Y} = \mathbf{y}, \mathbf{Z} = \mathbf{z}|\mathbf{x}; \theta) \stackrel{\text{def}}{=} \frac{1}{Z_x} \prod_r \Phi^{\text{join}}(y^r, \mathbf{z}) \prod_i \Phi^{\text{extract}}(z_i, x_i)$$

where the parameter vector θ is used, below, to define the factor Φ^{extract}.

The factors Φ^{join} are deterministic OR operators

$$\Phi^{\text{join}}(y^r, \mathbf{z}) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } y^r = true \wedge \exists i : z_i = r \\ 0 & \text{otherwise} \end{cases}$$

which are included to ensure that the ground fact $r(\mathbf{e})$ is predicted at the aggregate level for the assignment $Y^r = y^r$ only if at least one of the sentence level assignments $Z_i = z_i$ signals a mention of $r(\mathbf{e})$.

The extraction factors Φ^{extract} are given by

$$\Phi^{\text{extract}}(z_i, x_i) \stackrel{\text{def}}{=} \exp\left(\sum_j \theta_j \phi_j(z_i, x_i)\right)$$

where the features ϕ_j are sensitive to the relation name assigned to extraction variable z_i, if any, and cues from the sentence x_i. We will make use of the Mintz *et al.* (2009) sentence-level features in the expeiments, as described in Section 7.

3.3 Discussion

This model was designed to provide a joint approach where extraction decisions are almost entirely driven by sentence-level reasoning. However, defining the Y^r random variables and tying them to the sentence-level variables, Z_i, provides a direct method for modeling weak supervision. We can simply train the model so that the Y variables match the facts in the database, treating the Z_i as hidden variables that can take any value, as long as they produce the correct aggregate predictions.

This approach is related to the multi-instance learning approach of Riedel *et al.* (2010), in that both models include sentence-level and aggregate random variables. However, their sentence level variables are binary and they only have a single aggregate variable that takes values $r \in R \cup \{\texttt{none}\}$, thereby ruling out overlapping relations. Additionally, their aggregate decisions make use of Mintz-style aggregate features (Mintz et al., 2009), that collect evidence from multiple sentences, while we use

Inputs:
(1) Σ, a set of sentences,
(2) E, a set of entities mentioned in the sentences,
(3) R, a set of relation names, and
(4) Δ, a database of atomic facts of the form $r(e_1, e_2)$ for $r \in R$ and $e_i \in E$.

Definitions:
We define the training set $\{(\mathbf{x}_i, \mathbf{y}_i)|i = 1 \ldots n\}$, where i is an index corresponding to a particular entity pair (e_j, e_k) in Δ, $\mathbf{x}_i$ contains all of the sentences in Σ with mentions of this pair, and $\mathbf{y}_i =$ **relVector**(e_j, e_k).

Computation:
initialize parameter vector $\mathbf{\Theta} \leftarrow \mathbf{0}$
for $t = 1...T$ **do**
  for $i = 1...n$ **do**
    $(\mathbf{y}', \mathbf{z}') \leftarrow \arg\max_{\mathbf{y},\mathbf{z}} p(\mathbf{y}, \mathbf{z}|\mathbf{x_i}; \theta)$
    if $\mathbf{y}' \neq \mathbf{y_i}$ **then**
      $\mathbf{z}^* \leftarrow \arg\max_{\mathbf{z}} p(\mathbf{z}|\mathbf{x_i}, \mathbf{y_i}; \theta)$
      $\mathbf{\Theta} \leftarrow \mathbf{\Theta} + \phi(\mathbf{x_i}, \mathbf{z}^*) - \phi(\mathbf{x_i}, \mathbf{z}')$
    end if
  end for
end for
Return Θ

Figure 2: The MULTIR Learning Algorithm

only the deterministic OR nodes. Perhaps surprising, we are still able to improve performance at both the sentential and aggregate extraction tasks.

4 Learning

We now present a multi-instance learning algorithm for our weak-supervision model that treats the sentence-level extraction random variables Z_i as latent, and uses facts from a database (*e.g.*, Freebase) as supervision for the aggregate-level variables Y^r.

As input we have (1) Σ, a set of sentences, (2) E, a set of entities mentioned in the sentences, (3) R, a set of relation names, and (4) Δ, a database of atomic facts of the form $r(e_1, e_2)$ for $r \in R$ and $e_i \in E$. Since we are using weak learning, the Y^r variables in $\mathbf{Y}$ are not directly observed, but can be approximated from the database Δ. We use a procedure, **relVector**(e_1, e_2) to return a bit vector whose j^{th} bit is one if $r_j(e_1, e_2) \in \Delta$. The vector does *not* have a bit for the special *none* relation; if there is no relation between the two entities, all bits are zero.

Finally, we can now define the training set to be pairs $\{(\mathbf{x}_i, \mathbf{y}_i)|i = 1 \ldots n\}$, where i is an index corresponding to a particular entity pair (e_j, e_k), $\mathbf{x}_i$ contains all of the sentences with mentions of this pair, and $\mathbf{y}_i =$ **relVector**(e_j, e_k).

Given this form of supervision, we would like to find the setting for θ with the highest likelihood:

$$O(\theta) = \prod_i p(\mathbf{y}_i|\mathbf{x}_i; \theta) = \prod_i \sum_{\mathbf{z}} p(\mathbf{y}_i, \mathbf{z}|\mathbf{x}_i; \theta)$$

However, this objective would be difficult to optimize exactly, and algorithms for doing so would be unlikely to scale to data sets of the size we consider. Instead, we make two approximations, described below, leading to a Perceptron-style additive (Collins, 2002) parameter update scheme which has been modified to reason about hidden variables, similar in style to the approaches of (Liang et al., 2006; Zettlemoyer and Collins, 2007), but adapted for our specific model. This approximate algorithm is computationally efficient and, as we will see, works well in practice.

Our first modification is to do online learning instead of optimizing the full objective. Define the feature sums $\phi(\mathbf{x}, \mathbf{z}) = \sum_j \phi(x_j, z_j)$ which range over the sentences, as indexed by j. Now, we can define an update based on the gradient of the local log likelihood for example i:

$$\frac{\partial \log O_i(\theta)}{\partial \theta_j} = E_{p(\mathbf{z}|\mathbf{x}_i, \mathbf{y}_i; \theta)}[\phi_j(\mathbf{x}_i, \mathbf{z})] - E_{p(\mathbf{y}, \mathbf{z}|\mathbf{x}_i; \theta)}[\phi_j(\mathbf{x}_i, \mathbf{z})]$$

where the deterministic OR Φ^{join} factors ensure that the first expectation assigns positive probability only to assignments that produce the labeled facts $\mathbf{y}_i$ but that the second considers all valid sets of extractions.

Of course, these expectations themselves, especially the second one, would be difficult to compute exactly. Our second modification is to do a Viterbi approximation, by replacing the expectations with maximizations. Specifically, we compute the most likely sentence extractions for the label facts $\arg\max_{\mathbf{z}} p(\mathbf{z}|\mathbf{x}_i, \mathbf{y}_i; \theta)$ and the most likely extraction for the input, without regard to the labels, $\arg\max_{\mathbf{y},\mathbf{z}} p(\mathbf{y}, \mathbf{z}|\mathbf{x}_i; \theta)$. We then compute the features for these assignments and do a simple additive update. The final algorithm is detailed in Figure 2.

5 Inference

To support learning, as described above, we need to compute assignments $\arg\max_{\mathbf{z}} p(\mathbf{z}|\mathbf{x}, \mathbf{y}; \theta)$ and $\arg\max_{\mathbf{y},\mathbf{z}} p(\mathbf{y}, \mathbf{z}|\mathbf{x}; \theta)$. In this section, we describe algorithms for both cases that use the deterministic OR nodes to simplify the required computations.

Predicting the most likely joint extraction $\arg\max_{\mathbf{y},\mathbf{z}} p(\mathbf{y}, \mathbf{z}|\mathbf{x}; \theta)$ can be done efficiently given the structure of our model. In particular, we note that the factors Φ^{join} represent deterministic dependencies between $\mathbf{Z}$ and $\mathbf{Y}$, which when satisfied do not affect the probability of the solution. It is thus sufficient to independently compute an assignment for each sentence-level extraction variable Z_i, ignoring the deterministic dependencies. The optimal setting for the aggregate variables $\mathbf{Y}$ is then simply the assignment that is consistent with these extractions. The time complexity is $O(|\mathrm{R}| \cdot |\mathrm{S}|)$.

Predicting sentence level extractions given weak supervision facts, $\arg\max_{\mathbf{z}} p(\mathbf{z}|\mathbf{x}, \mathbf{y}; \theta)$, is more challenging. We start by computing extraction scores $\Phi^{\text{extract}}(x_i, z_i)$ for each possible extraction assignment $Z_i = z_i$ at each sentence $x_i \in \mathrm{S}$, and storing the values in a dynamic programming table. Next, we must find the most likely assignment $\mathbf{z}$ that respects our output variables $\mathbf{y}$. It turns out that this problem is a variant of the weighted, edge-cover problem, for which there exist polynomial time optimal solutions.

Let $G = (\mathcal{E}, \mathcal{V} = \mathcal{V}^{\mathrm{S}} \cup \mathcal{V}^{\mathbf{y}})$ be a complete weighted bipartite graph with one node $v_i^{\mathrm{S}} \in \mathcal{V}^{\mathrm{S}}$ for each sentence $x_i \in \mathrm{S}$ and one node $v_r^{\mathbf{y}} \in \mathcal{V}^{\mathbf{y}}$ for each relation $r \in \mathrm{R}$ where $y^r = 1$. The edge weights are given by $c((v_i^{\mathrm{S}}, v_r^{\mathbf{y}})) \stackrel{\text{def}}{=} \Phi^{\text{extract}}(\mathbf{x}_i, z_i)$. Our goal is to select a subset of the edges which maximizes the sum of their weights, subject to each node $v_i^{\mathrm{S}} \in \mathcal{V}^{\mathrm{S}}$ being incident to exactly one edge, and each node $v_r^{\mathbf{y}} \in \mathcal{V}^{\mathbf{y}}$ being incident to at least one edge.

Exact Solution An exact solution can be obtained by first computing the maximum weighted bipartite matching, and adding edges to nodes which are not incident to an edge. This can be computed in time $O(|\mathcal{V}|(|\mathcal{E}| + |\mathcal{V}| \log |\mathcal{V}|))$, which we can rewrite as $O((|\mathrm{R}| + |\mathrm{S}|)(|\mathrm{R}||\mathrm{S}| + (|\mathrm{R}| + |\mathrm{S}|) \log(|\mathrm{R}| + |\mathrm{S}|)))$.

Approximate Solution An approximate solution can be obtained by iterating over the nodes in $\mathcal{V}^{\mathbf{y}}$, and each time adding the highest weight incident edge whose addition doesn't violate a constraint. The running time is $O(|\mathrm{R}||\mathrm{S}|)$. This greedy search guarantees each fact is extracted at least once and allows any additional extractions that increase the overall probability of the assignment. Given the computational advantage, we use it in all of the experimental evaluations.

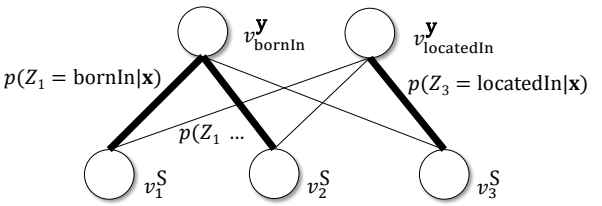

Figure 3: Inference of $\arg\max_{\mathbf{z}} p(\mathbf{Z} = \mathbf{z}|\mathbf{x}, \mathbf{y})$ requires solving a weighted, edge-cover problem.

6 Experimental Setup

We follow the approach of Riedel *et al.* (2010) for generating weak supervision data, computing features, and evaluating aggregate extraction. We also introduce new metrics for measuring sentential extraction performance, both relation-independent and relation-specific.

6.1 Data Generation

We used the same data sets as Riedel *et al.* (2010) for weak supervision. The data was first tagged with the Stanford NER system (Finkel et al., 2005) and then entity mentions were found by collecting each continuous phrase where words were tagged identically (*i.e.*, as a person, location, or organization). Finally, these phrases were matched to the names of Freebase entities.

Given the set of matches, define Σ to be set of NY Times sentences with two matched phrases, E to be the set of Freebase entities which were mentioned in one or more sentences, Δ to be the set of Freebase facts whose arguments, e_1 and e_2 were mentioned in a sentence in Σ, and R to be set of relations names used in the facts of Δ. These sets define the weak supervision data.

6.2 Features and Initialization

We use the set of sentence-level features described by Riedel *et al.* (2010), which were originally de-

veloped by Mintz *et al.* (2009). These include indicators for various lexical, part of speech, named entity, and dependency tree path properties of entity mentions in specific sentences, as computed with the Malt dependency parser (Nivre and Nilsson, 2004) and OpenNLP POS tagger[1]. However, unlike the previous work, we did not make use of any features that explicitly aggregate these properties across multiple mention instances.

The MULTIR algorithm has a single parameter T, the number of training iterations, that must be specified manually. We used $T = 50$ iterations, which performed best in development experiments.

6.3 Evaluation Metrics

Evaluation is challenging, since only a small percentage (approximately 3%) of sentences match facts in Freebase, and the number of matches is highly unbalanced across relations, as we will see in more detail later. We use the following metrics.

Aggregate Extraction Let Δ^e be the set of extracted relations for any of the systems; we compute aggregate precision and recall by comparing Δ^e with Δ. This metric is easily computed but underestimates extraction accuracy because Freebase is incomplete and some true relations in Δ^e will be marked wrong.

Sentential Extraction Let S^e be the sentences where some system extracted a relation and S^F be the sentences that match the arguments of a fact in Δ. We manually compute sentential extraction accuracy by sampling a set of 1000 sentences from $S^e \cup S^F$ and manually labeling the correct extraction decision, either a relation $r \in R$ or $none$. We then report precision and recall for each system on this set of sampled sentences. These results provide a good approximation to the true precision but can overestimate the actual recall, since we did not manually check the much larger set of sentences where no approach predicted extractions.

6.4 Precision / Recall Curves

To compute precision / recall curves for the tasks, we ranked the MULTIR extractions as follows. For sentence-level evaluations, we ordered according to the extraction factor score $\Phi^{\text{extract}}(z_i, x_i)$. For aggregate comparisons, we set the score for an extraction $Y^r = true$ to be the max of the extraction factor scores for the sentences where r was extracted.

[1] http://opennlp.sourceforge.net/

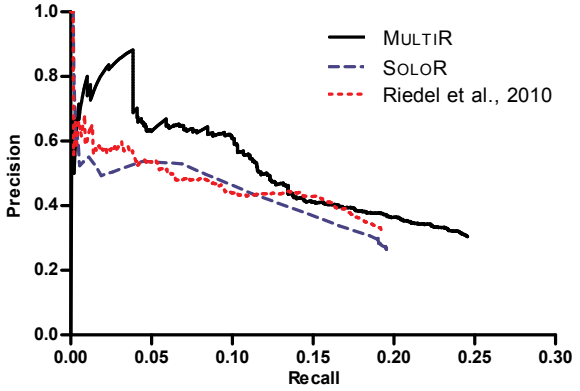

Figure 4: Aggregate extraction precision / recall curves for Riedel *et al.* (2010), a reimplementation of that approach (SOLOR), and our algorithm (MULTIR).

7 Experiments

To evaluate our algorithm, we first compare it to an existing approach for using multi-instance learning with weak supervision (Riedel et al., 2010), using the same data and features. We report both aggregate extraction and sentential extraction results. We then investigate relation-specific performance of our system. Finally, we report running time comparisons.

7.1 Aggregate Extraction

Figure 4 shows approximate precision / recall curves for three systems computed with aggregate metrics (Section 6.3) that test how closely the extractions match the facts in Freebase. The systems include the original results reported by Riedel *et al.* (2010) as well as our new model (MULTIR). We also compare with SOLOR, a reimplementation of their algorithm, which we built in Factorie (McCallum et al., 2009), and will use later to evaluate sentential extraction.

MULTIR achieves competitive or higher precision over all ranges of recall, with the exception of the very low recall range of approximately 0-1%. It also significantly extends the highest recall achieved, from 20% to 25%, with little loss in precision. To investigate the low precision in the 0-1% recall range, we manually checked the ten highest con-

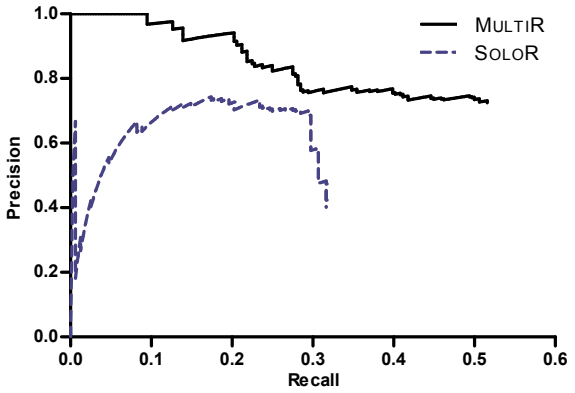

Figure 5: Sentential extraction precision / recall curves for MULTIR and SOLOR.

fidence extractions produced by MULTIR that were marked wrong. We found that all ten were true facts that were simply missing from Freebase. A manual evaluation, as we perform next for sentential extraction, would remove this dip.

7.2 Sentential Extraction

Although their model includes variables to model sentential extraction, Riedel *et al.* (2010) did not report sentence level performance. To generate the precision / recall curve we used the joint model assignment score for each of the sentences that contributed to the aggregate extraction decision.

Figure 4 shows approximate precision / recall curves for MULTIR and SOLOR computed against manually generated sentence labels, as defined in Section 6.3. MULTIR achieves significantly higher recall with a consistently high level of precision. At the highest recall point, MULTIR reaches 72.4% precision and 51.9% recall, for an F1 score of 60.5%.

7.3 Relation-Specific Performance

Since the data contains an unbalanced number of instances of each relation, we also report precision and recall for each of the ten most frequent relations. Let S_r^M be the sentences where MULTIR extracted an instance of relation $r \in R$, and let S_r^F be the sentences that match the arguments of a fact about relation r in Δ. For each r, we sample 100 sentences from both S_r^M and S_r^F and manually check accuracy. To estimate precision $\tilde{P}_r$ we compute the ratio of true relation mentions in S_r^M, and to estimate recall $\tilde{R}_r$ we take the ratio of true relation mentions in S_r^F which are returned by our system.

Table 1 presents this approximate precision and recall for MULTIR on each of the relations, along with statistics we computed to measure the quality of the weak supervision. Precision is high for the majority of relations but recall is consistently lower. We also see that the Freebase matches are highly skewed in quantity and can be low quality for some relations, with very few of them actually corresponding to true extractions. The approach generally performs best on the relations with a sufficiently large number of true matches, in many cases even achieving precision that outperforms the accuracy of the heuristic matches, at reasonable recall levels.

7.4 Overlapping Relations

Table 1 also highlights some of the effects of learning with overlapping relations. For example, in the data, almost all of the matches for the administrative_divisions relation overlap with the contains relation, because they both model relationships for a pair of locations. Since, in general, sentences are much more likely to describe a contains relation, this overlap leads to a situation were almost none of the administrate_division matches are true ones, and we cannot accurately learn an extractor. However, we can still learn to accurately extract the contains relation, despite the distracting matches. Similarly, the place_of_birth and place_of_death relations tend to overlap, since it is often the case that people are born and die in the same city. In both cases, the precision outperforms the labeling accuracy and the recall is relatively high.

To measure the impact of modeling overlapping relations, we also evaluated a simple, restricted baseline. Instead of labeling each entity pair with the set of all true Freebase facts, we created a dataset where each true relation was used to create a different training example. Training MULTIR on this data simulates effects of conflicting supervision that can come from not modeling overlaps. On average across relations, precision increases 12 points but recall drops 26 points, for an overall reduction in F1 score from 60.5% to 40.3%.

7.5 Running Time

One final advantage of our model is the modest running time. Our implementation of the

Relation	Freebase Matches		MULTIR	
	#sents	% true	$\tilde{P}$	$\tilde{R}$
/business/person/company	302	89.0	100.0	25.8
/people/person/place_lived	450	60.0	80.0	6.7
/location/location/contains	2793	51.0	100.0	56.0
/business/company/founders	95	48.4	71.4	10.9
/people/person/nationality	723	41.0	85.7	15.0
/location/neighborhood/neighborhood_of	68	39.7	100.0	11.1
/people/person/children	30	80.0	100.0	8.3
/people/deceased_person/place_of_death	68	22.1	100.0	20.0
/people/person/place_of_birth	162	12.0	100.0	33.0
/location/country/administrative_divisions	424	0.2	N/A	0.0

Table 1: Estimated precision and recall by relation, as well as the number of matched sentences (#sents) and accuracy (% true) of matches between sentences and facts in Freebase.

Riedel *et al.* (2010) approach required approximately 6 hours to train on NY Times 05-06 and 4 hours to test on the NY Times 07, each without preprocessing. Although they do sampling for inference, the global aggregation variables require reasoning about an exponentially large (in the number of sentences) sample space.

In contrast, our approach required approximately one minute to train and less than one second to test, on the same data. This advantage comes from the decomposition that is possible with the deterministic OR aggregation variables. For test, we simply consider each sentence in isolation and during training our approximation to the weighted assignment problem is linear in the number of sentences.

7.6 Discussion

The sentential extraction results demonstrates the advantages of learning a model that is primarily driven by sentence-level features. Although previous approaches have used more sophisticated features for aggregating the evidence from individual sentences, we demonstrate that aggregating strong sentence-level evidence with a simple deterministic OR that models overlapping relations is more effective, and also enables training of a sentence extractor that runs with no aggregate information.

While the Riedel *et al.* approach does include a model of which sentences express relations, it makes significant use of aggregate features that are primarily designed to do entity-level relation predictions and has a less detailed model of extractions at the individual sentence level. Perhaps surprisingly, our model is able to do better at both the sentential and aggregate levels.

8 Related Work

Supervised-learning approaches to IE were introduced in (Soderland et al., 1995) and are too numerous to summarize here. While they offer high precision and recall, these methods are unlikely to scale to the thousands of relations found in text on the Web. Open IE systems, which perform self-supervised learning of relation-independent extractors (*e.g.*, Preemptive IE (Shinyama and Sekine, 2006), TEXTRUNNER (Banko et al., 2007; Banko and Etzioni, 2008) and WOE (Wu and Weld, 2010)) can scale to millions of documents, but don't output canonicalized relations.

8.1 Weak Supervision

Weak supervision (also known as distant- or self supervision) refers to a broad class of methods, but we focus on the increasingly-popular idea of using a store of structured data to heuristicaly label a textual corpus. Craven and Kumlien (1999) introduced the idea by matching the Yeast Protein Database (YPD) to the abstracts of papers in PubMed and training a naive-Bayes extractor. Bellare and McCallum (2007) used a database of BibTex records to train a CRF extractor on 12 bibliographic relations. The KYLIN system aplied weak supervision to learn relations from Wikipedia, treating infoboxes as the associated database (Wu and Weld, 2007); Wu *et al.* (2008) extended the system to use smoothing over an automatically generated infobox taxon-

omy. Mintz *et al.* (2009) used Freebase facts to train 100 relational extractors on Wikipedia. Hoffmann *et al.* (2010) describe a system similar to KYLIN, but which dynamically generates lexicons in order to handle sparse data, learning over 5000 Infobox relations with an average F1 score of 61%. Yao *et al.* (2010) perform weak supervision, while using selectional preference constraints to a jointly reason about entity types.

The NELL system (Carlson et al., 2010) can also be viewed as performing weak supervision. Its initial knowledge consists of a selectional preference constraint and 20 ground fact seeds. NELL then matches entity pairs from the seeds to a Web corpus, but instead of learning a probabilistic model, it bootstraps a set of extraction patterns using semi-supervised methods for multitask learning.

8.2 Multi-Instance Learning

Multi-instance learning was introduced in order to combat the problem of ambiguously-labeled training data when predicting the activity of different drugs (Dietterich et al., 1997). Bunescu and Mooney (2007) connect weak supervision with multi-instance learning and extend their relational extraction kernel to this context.

Riedel *et al.* (2010), combine weak supervision and multi-instance learning in a more sophisticated manner, training a graphical model, which assumes only that *at least one* of the matches between the arguments of a Freebase fact and sentences in the corpus is a true relational mention. Our model may be seen as an extension of theirs, since both models include sentence-level and aggregate random variables. However, Riedel *et al.* have only a single aggregate variable that takes values $r \in R \cup \{\texttt{none}\}$, thereby ruling out overlapping relations. We have discussed the comparison in more detail throughout the paper, including in the model formulation section and experiments.

9 Conclusion

We argue that weak supervision is promising method for scaling information extraction to the level where it can handle the myriad, different relations on the Web. By using the contents of a database to heuristically label a training corpus, we may be able to automatically learn a nearly unbounded number of relational extractors. Since the processs of matching database tuples to sentences is inherently heuristic, researchers have proposed multi-instance learning algorithms as a means for coping with the resulting noisy data. Unfortunately, previous approaches assume that all relations are *disjoint* — for example they cannot extract the pair `Founded(Jobs, Apple)` and `CEO-of(Jobs, Apple)`, because two relations are not allowed to have the same arguments.

This paper presents a novel approach for multi-instance learning with overlapping relations that combines a sentence-level extraction model with a simple, corpus-level component for aggregating the individual facts. We apply our model to learn extractors for NY Times text using weak supervision from Freebase. Experiments show improvements for both sentential and aggregate (corpus level) extraction, and demonstrate that the approach is computationally efficient.

Our early progress suggests many interesting directions. By joining two or more Freebase tables, we can generate many more matches and learn more relations. We also wish to refine our model in order to improve precision. For example, we would like to add type reasoning about entities and selectional preference constraints for relations. Finally, we are also interested in applying the overall learning approaches to other tasks that could be modeled with weak supervision, such as coreference and named entity classification.

The source code of our system, its output, and all data annotations are available at `http://cs.uw.edu/homes/raphaelh/mr`.

Acknowledgments

We thank Sebastian Riedel and Limin Yao for sharing their data and providing valuable advice. This material is based upon work supported by a WRF / TJ Cable Professorship, a gift from Google and by the Air Force Research Laboratory (AFRL) under prime contract no. FA8750-09-C-0181. Any opinions, findings, and conclusion or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of the Air Force Research Laboratory (AFRL).

References

Michele Banko and Oren Etzioni. 2008. The tradeoffs between open and traditional relation extraction. In *Proceedings of 46th Annual Meeting of the Association for Computational Linguistics (ACL-08)*, pages 28–36.

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-07)*, pages 2670–2676.

Kedar Bellare and Andrew McCallum. 2007. Learning extractors from unlabeled text using relevant databases. In *Sixth International Workshop on Information Integration on the Web*.

Razvan Bunescu and Raymond Mooney. 2007. Learning to extract relations from the web using minimal supervision. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-07)*.

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 the AAAI Conference on Artificial Intelligence (AAAI-10)*.

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*.

Mark Craven and Johan Kumlien. 1999. Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pages 77–86.

Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89:31–71, January.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*, pages 363–370.

Raphael Hoffmann, Congle Zhang, and Daniel S. Weld. 2010. Learning 5000 relational extractors. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-10)*, pages 286–295.

Percy Liang, A. Bouchard-Côté, Dan Klein, and Ben Taskar. 2006. An end-to-end discriminative approach to machine translation. In *International Conference on Computational Linguistics and Association for Computational Linguistics (COLING/ACL)*.

Andrew McCallum, Karl Schultz, and Sameer Singh. 2009. Factorie: Probabilistic programming via imperatively defined factor graphs. In *Neural Information Processing Systems Conference (NIPS)*.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL-2009)*, pages 1003–1011.

Joakim Nivre and Jens Nilsson. 2004. Memory-based dependency parsing. In *Proceedings of the Conference on Natural Language Learning (CoNLL-04)*, pages 49–56.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of the Sixteenth European Conference on Machine Learning (ECML-2010)*, pages 148–163.

Yusuke Shinyama and Satoshi Sekine. 2006. Preemptive information extraction using unrestricted relation discovery. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computation Linguistics (HLT-NAACL-06)*.

Stephen Soderland, David Fisher, Jonathan Aseltine, and Wendy G. Lehnert. 1995. Crystal: Inducing a conceptual dictionary. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-1995)*, pages 1314–1321.

Fei Wu and Daniel S. Weld. 2007. Autonomously semantifying wikipedia. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM-2007)*, pages 41–50.

Fei Wu and Daniel S. Weld. 2008. Automatically refining the wikipedia infobox ontology. In *Proceedings of the 17th International Conference on World Wide Web (WWW-2008)*, pages 635–644.

Fei Wu and Daniel S. Weld. 2010. Open information extraction using wikipedia. In *The Annual Meeting of the Association for Computational Linguistics (ACL-2010)*, pages 118–127.

Limin Yao, Sebastian Riedel, and Andrew McCallum. 2010. Collective cross-document relation extraction without labelled data. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP-2010)*, pages 1013–1023.

Luke Zettlemoyer and Michael Collins. 2007. Online learning of relaxed CCG grammars for parsing to logical form. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL-2007)*.