Modeling Relations and Their Mentions without Labeled Text

Sebastian Riedel, Limin Yao, and Andrew McCallum

University of Massachusetts, Amherst,
Amherst, MA 01002, U.S.
{riedel,lmyao,mccallum,lncs}@cs.umass.edu

Abstract. Several recent works on relation extraction have been applying the distant supervision paradigm: instead of relying on annotated text to learn how to predict relations, they employ existing knowledge bases (KBs) as source of supervision. Crucially, these approaches are trained based on the assumption that each sentence which mentions the two related entities is an expression of the given relation. Here we argue that this leads to noisy patterns that hurt precision, in particular if the knowledge base is not directly related to the text we are working with. We present a novel approach to distant supervision that can alleviate this problem based on the following two ideas: First, we use a factor graph to explicitly model the decision whether two entities are related, and the decision whether this relation is mentioned in a given sentence; second, we apply constraint-driven semi-supervision to train this model without any knowledge about which sentences express the relations in our training KB. We apply our approach to extract relations from the New York Times corpus and use Freebase as knowledge base. When compared to a state-of-the-art approach for relation extraction under distant supervision, we achieve 31% error reduction.

1 Introduction

In relation extraction we often encounter a lack of explicitly annotated text, but an abundance of structured data sources such as company databases or large scale public knowledge bases like Freebase [2]. In the spirit of *distant supervision*[1] [8, 19], recent work [18, 3] has shown how to exploit such knowledge: they heuristically align the given knowledge base to text and use this alignment to learn a relation extractor. Their approach is based on the following *distant supervision assumption*:

> If two entities participate in a relation, **all sentences** that mention these two entities express that relation.

In practice, this allows them to extract features from all the sentence to feed a relation classifier. This approach has helped [18] to extract several thousand relations from Wikipedia at a precision of about 70% using Freebase as supervision source, a knowledge base derived in large parts from Wikipedia info-boxes.

[1] Also referred to as *weak* or *self supervision*.

In this work we argue that the distant supervision assumption is too strong and needs to be relaxed, in particular when the training knowledge base is an *external* source of information and not primarily derived from the training text. This is the case, for example, when we want to extract new relations from newswire instead of Wikipedia. This scenario is very relevant in practice—after all, many structured data sources are not derived from the textual data we want to extract new relations from.

When the knowledge base is external, entities may just appear in the same sentence because they are related to the same topic, not necessarily because the sentence is expressing their relations in our training knowledge base. In fact, by manual inspection (see section 2) we find that the distant supervision assumption is violated approximately 13% of the time when aligning Freebase to Wikipedia, but 31% when aligning to the New York Times Corpus [22].[2]

In this paper we employ the following *expressed-at-least-once* assumption and show that it leads to more accurate results:

> If two entities participate in a relation, **at least one sentence** that mentions these two entities might express that relation.

Intuitively this statement holds with more certainty, but it also complicates our prediction task. Previously, we could simply take all sentences, aggregate features, and then solve a simple classification task. Now we do not know which sentences express relations, both during testing and training.

To tackle this problem we make two contributions. First, we introduce a novel undirected graphical model that captures both the task of predicting relations between entities, and the task of predicting which sentences express these relations. Our model connects a *relation variable* for two entities with a set of binary *relation mention* variables that indicate whether certain candidate sentences are expressing this relation. Crucially, the relation mention variables are unobserved at training time: we only know that a relation is expressed at least once, but not in which sentences.

Second, we propose to train this graphical model by framing distant supervision as an instance of *constraint-driven semi-supervision* [5, 4, 1, 16]. This type of supervision is applied to settings where some variables are latent. Roughly speaking, here model parameters are optimized to ensure that predictions will satisfy a set of user-defined constraints, as opposed to a set of target labels. In this framework our approach of distant supervision can be implemented by using the expressed-at-least-once constraint at training time.

As learner we choose SampleRank [27], a very efficient method to train parameters of large scale factor graphs. Recent work has shown that it can be naturally extended to the case of constraint-driven learning [24]. We believe that this choice will also be crucial for future work, where we expect our models to make joint relation, entity and coreference decisions across a whole corpus.

[2] This is the average over three relation types: *nationality*, *contains* and *place_of_birth*.

SampleRank supports this setup because it makes parameter updates early and *within* inference.

We apply our model to extract relations from the New York Times corpus using Freebase as the external supervision source. We observe that our model with expressed-at-least-once assumption leads to 91% precision for our top 1000 predictions. When compared to 87% precision for a model based on the distant supervision assumption, this amounts to 31% error reduction.

In the following we will first give some background on relation extraction and distant supervision, then present our factor graph for joint relation type and relation mention identification. We then explain how to use SampleRank to incorporate the expressed-at-least-once assumption, present related work, show our empirical results, and conclude.

2 Relation Extraction under Distant Supervision

Relation Extraction is understood here as the task of predicting the relations expressed in natural language text. Consider, for example, the following sentence

> **Elevation Partners**, the $ 1.9 billion private equity group that was *founded* by **Roger McNamee** ...

Here the pair of entity mentions "Elevation Partners" and "Roger McNamee" is a *relation mention candidate* because its context might express a semantic relation between the corresponding pair of entities. A relation extractor takes such a candidate and determines the semantic relation that it might express, if any. In the above case a good extractor predicts the *founded* relation; this implies that the relation mention candidate is indeed a *relation mention*.[3]

In the works of [18, 3, 8], relation extraction is understood somewhat differently. Their primary goal is to determine whether a relation between a given pair of entities is expressed *somewhere* in the text, not necessarily where it is expressed. In other words, they care for *relations*, not *relation mentions*.

Focusing on relations instead of relation mentions has several benefits. First, it is very relevant in practice because downstream applications often care for entities and their relations, not for every mention of these. Second, it allows us to aggregate evidence for a relation from several places in the corpus. Finally, it simplifies the machine learning task: while we usually only have a few annotated relation mentions, we often have many existing pairs of related entities of the type we want to extract.

To illustrate the final point, let us consider the work of [18]. Their task is to extract relations of Freebase, a large online and collaborative knowledge base, from Wikipedia text. They tackle it by using the existing relations in Freebase as training data: for each related pair of entities they collect all sentences that mention both entities as input observation $\mathbf{x}$, and use their relation type in

[3] Note that in this work we focus on *closed* relation extraction where the extractor predicts one of a finite and fixed set of relations.

Freebase as label y. Together with a set of unrelated pairs of entities as negative instances, they train a classifier to predict relations (but not relation mentions).

The approaches of [18, 3] assume that each relation mention candidate is indeed a relation mention. Clearly, this assumption can be violated. Let us again consider the *founded* relation between "Roger McNamee" and "Elevation Partners. In an 2007 article of the New York Times we find this relation mention candidate:

Roger McNamee, a managing director at **Elevation Partners**, ...

This sentence does not express that Roger McNamee is a founder of Elevation Partners. It may make it more likely, but there are certainly cases where managing directors are not founders. The problem with this observation is that at training time we may learn some positive weight for the feature "<Entity1>, a managing director at <Entity2>". When testing our model we may see this feature for a director A that is *not* a founder of a company B, and predict a false positive.

To get a sense of how frequently the distant supervision assumption is violated in practice, we test it for the case of Freebase and two different text corpora: Wikipedia articles and the New York Times corpus. To this end we consider three frequent relation types: *nationality*, *place_of_birth*, and *contains*. For each type we sample 100 relation mention candidates from both corpora, and evaluate whether these candidates are or are not expressing the relation in question. Table 1 presents the gathered statistics.

Table 1. Percentage of times a related pair of entities is mentioned in the same sentence, but where the sentence does not express the corresponding relation.

Relation Type	New York Times	Wikipedia
nationality	38%	20%
place_of_birth	35%	20%
contains	20%	10%

We can see that for NYT data, the probability that a candidate mention is a non-mention is quite high. This is not difficult to understand: Take *nationality* as an example. Since a citizen of a country usually lives in the country, he/she will be involved in events happening in the country. News articles will report such events, and hence naturally mention the country and the person together. However, there is usually no need to express the fact that the person is indeed a citizen of the country—most readers care about the events but not the nationality of their participants.

How does this compare to relation mentions in Wikipedia? Here articles are centered around entities, and express targeted information about them. For example, if the article concerns a person, we expect the article to mention the person's citizenship. However, unless the person holds a special role (say, a political role) in his country, we do not expect many additional sentences that

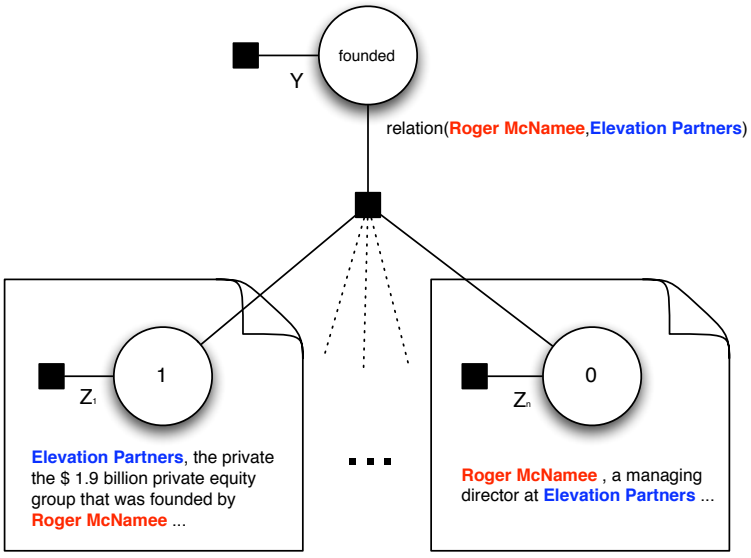

Fig. 1. Factor Graph for joint relation mention prediction and relation type identification. For each pair of entities that are mentioned together in at least one sentence we create one relation variable (the top variable here). For each of the pairs of entity mentions that appear in a sentence we create one relation mention variable, and connect it to the corresponding relation variable. Note that relation variables for different pairs of entities are considered to be independent.

mention both him and his country. Indeed, when comparing the percentage of non-mentions for *nationality*, we find about twice as many cases of non-mentions in NYT articles than in Wikipedia data.

3 Model

Our observations in the previous section suggest that we should model both relations and relation mentions in order to avoid the use of noisy patterns. We propose to achieve this using an undirected graphical model with two types of hidden variables. First, for a pair of entities S (source) and D (destination) that appears together in at least one sentence, a *relation variable* Y denotes the relation between them, or NA if there is no such relation. See an example relation variable in figure 1. Second, for each relation mention candidate i, we define a binary *relation mention variable* Z_i that is true if and only if mention i is indeed expressing the relation Y between the two entities. Two example mention variables can be seen in figure 1.

For each relation mention variable Z_i we will refer to its two argument entity mentions as (source) S_i and (destination) D_i. We will store additional information about the sentence Z_i appears in, such as the dependency path between

S_i and D_i, in an observed value $\mathbf{x}_i$. This information is aggregated across all mention candidates in the vector $\mathbf{x}$. Finally, we will use $\mathbf{Z}$ to denote the state of all mention candidates, and $\|\mathbf{z}\|$ to represent the number of active relation mentions for a given assignment $\mathbf{z}$ of $\mathbf{Z}$.

Our conditional probability distribution over these variables is defined as follows:

$$p\left(Y = y, \mathbf{Z} = \mathbf{z} | \mathbf{x}\right) \stackrel{\text{def}}{=} \frac{1}{Z_{\mathbf{x}}} \Phi^{\text{r}}\left(y\right) \Phi^{\text{join}}\left(y, \mathbf{z}, \mathbf{x}\right) \prod_i \Phi^{\text{m}}\left(z_i, \mathbf{x}_i\right) .$$

Here the factor (template) Φ^{r} assesses the general bias of the model towards a particular relation type y and is defined as a loglinear potential function $\Phi^{\text{r}}\left(y\right) = \exp\left(\theta^{\text{r}}_y\right)$. The factor Φ^{m} is defined as a function over a relation mention variable and its corresponding observation $\mathbf{x}_i$:

$$\Phi^{\text{m}}\left(z_i, \mathbf{x}_i\right) \stackrel{\text{def}}{=} \exp\left(\sum_j \theta^{\text{m}}_j \phi^{\text{m}}_j\left(z_i, \mathbf{x}_i\right)\right)$$

The feature functions ϕ^{m}_j are taken to be the binary features presented in [18]. For example, the feature

$$\phi^{\text{m}}_{101}\left(z_i, \mathbf{x}_i\right) \stackrel{\text{def}}{=} \begin{cases} 1 & z_i = 1 \wedge S_i, \text{ a managing director of } D_i \in \mathbf{x}_i \\ 0 & \text{otherwise} \end{cases}$$

returns 1 if Z_i is active and there is a sequence "S_i, a managing director of D_i", and 0 otherwise.

The factor Φ^{join} links relation variables to their mentions. It is defined as follows:

$$\Phi^{\text{join}}\left(y, \mathbf{z}, \mathbf{x}\right) \stackrel{\text{def}}{=} \exp\left(\sum_j \theta^{\text{join}}_{j,y} \ \phi^{\text{join}}_j \ \left(\mathbf{z}, \mathbf{x}\right)\right)$$

Here the feature functions ϕ^{join}_j are defined in terms of the mention feature functions ϕ^{m}_j:

$$\phi^{\text{join}}_j \ \left(\mathbf{z}, \mathbf{x}\right) \stackrel{\text{def}}{=} \begin{cases} 1 & \exists i : z_i = 1 \wedge \phi^{\text{m}}_j\left(z_i, \mathbf{x}_i\right) = 1 \\ 0 & \text{otherwise} \end{cases}$$

Hence, the feature ϕ^{join}_j indicates whether feature ϕ^{m}_j is active for any of the active relation mentions (as indicated by $z_i = 1$). For example, Φ^{join}_{101} fires if "a managing director" appears between the corresponding entity mentions of any the active relation mentions. This is precisely the type of feature used in [18] for the relation classifier. The crucial difference is that here we consider only *active* mention candidates, instead of *all* mention candidates.

To construct this factor graph for each pair of candidate entities, we use FACTORIE [17], a probabilistic programming language that simplifies the construction process, as well as inference and learning.

3.1 Inference

There are two types of inference in the model: sampling from the posterior during training (see section 4), and finding the most likely configuration (aka MAP inference). In both settings we employ a block Gibbs sampler [14, 13] that randomly picks a relation mention i, and jointly samples both Z_i and the corresponding relation variable$Y$. Here the sample is drawn conditioned on the state of all remaining relation mention variables. At test time we decrease the temperature of our sampler in order to find an approximation of the MAP solution.

We use block sampling instead of single-variable Gibbs sampling because of the strong correlation between mention variables Z_i and relation type Y. Assume, for example, that the current relation type Y is set to NA, all relation mentions Z_i are inactive, and we want to sample a new state for the first relation mention Z_1. In this case a model will give a near zero probability for $Z_1 = 1$ because it has learned that this assignment is inconsistent with $Y = \text{NA}$. Likewise, changing Y with all Z_i fixed to be 0 will also receive a very low probability. This may happen even if the model assigns a high probability to the combination of $Z_1 = 1$ and, say, $Y = \text{founder}$. Changing both relation and relation variable in concert overcomes this problem.

4 Rank-Based Learning and Distant Supervision

Most learning methods need to calculate the model expectations [15] or the MAP configuration [7] before making an update to the parameters. This step of inference is usually the bottleneck for learning, even when performed approximately.

SampleRank [21, 27] is a rank-based learning framework that alleviates this problem by performing parameter updates *within* MCMC inference. Every pair of consecutive samples in the MCMC chain is ranked according to the model and the ground truth, and the parameters are updated when the rankings disagree. This allows the learner to acquire more supervision per instance, and has led to efficient training for models in which inference is expensive and generally intractable [23].

SampleRank considers two ranking functions for an assignment $\mathbf{y}$: (1) the probability (model ranking) $p(\mathbf{y}) = \frac{1}{Z} \exp(\langle \Theta, \phi(\mathbf{y}) \rangle)$, where $\phi(\mathbf{y})$ is a feature representation of $\mathbf{y}$, and (2) a *truth function* $\mathcal{F}(\mathbf{y})$ (objective ranking). One such truth function could be a per-entity-pair accuracy with respect to some labeled relations, another could be the F1-measure.

The goal of applying SampleRank is to find parameters that make model ranking and objective ranking as consistent as possible. To achieve this, SampleRank performs the following update at each step of an MCMC chain (see section 3.1). Let $\mathbf{y}^{i-1}$ be the previous sample, and $\mathbf{y}^i$ the current sample of the chain, α be the learning rate, and $\Delta = \phi\left(\mathbf{y}^{i-1}\right) - \phi(\mathbf{y^i})$. Then the weights Θ

are updated as follows:

$$\Theta = \Theta + \begin{cases} \alpha\Delta & \text{if } \frac{p(\mathbf{y}^{i-1})}{p(\mathbf{y}^{\mathbf{i}})} < 1 \wedge \mathcal{F}(\mathbf{y}^{i-1}) > \mathcal{F}(\mathbf{y}^{\mathbf{i}}) \\ -\alpha\Delta & \text{if } \frac{p(\mathbf{y}^{i-1})}{p(\mathbf{y}^{\mathbf{i}})} > 1 \wedge \mathcal{F}(\mathbf{y}^{i-1}) < \mathcal{F}(\mathbf{y}^{\mathbf{i}}) \\ 0 & \text{otherwise} \end{cases}.$$

Note that to evaluate the model ratios we do not require to calculate the partition function Z.

To better illustrate SampleRank, let us consider the factor graph of figure 1, and assume we are in the shown state. Our Gibbs Sampler in section 3.1 now assigns a new state to relation variable and one of the relation mention variables. For example, it leaves the relation mention variable unchanged but sets the relation variable to *child-of*.

In early stages of training, this proposal may still have a higher probability than the previous state, hence $\frac{p(\mathbf{y^{i-1}})}{p(\mathbf{y^i})} < 1$. However, since we know that Roger McNamee is not a child of Elevation Partners, the previous state has higher truth score than the current state, and hence $\mathcal{F}(\mathbf{y^{i-1}}) > \mathcal{F}(\mathbf{y^i})$. This means that SampleRank will update weights into the direction Δ of the feature vector $\phi\left(\mathbf{y}^{i-1}\right)$ for the previous state.

In the following we will show how several distant supervision approaches can be incorporated into this framework. In all cases the truth function $\mathcal{F}(y, \mathbf{z})$ for an assignment of relation and relation mention variables is decomposed into

$$\mathcal{F}(y, \mathbf{z}) = \mathcal{F}_{\mathrm{r}}(y) + \mathcal{F}_m\left(\mathbf{z}, y_{\mathrm{truth}}\right)$$

where $\mathcal{F}_{\mathrm{r}}(y)$ only assesses the truth of the relation variable y and $\mathcal{F}_m\left(\mathbf{z}, y_{\mathrm{truth}}\right)$ assesses the truth of the relation mention variables. Here y_{truth} is set to be the relation associated with the entity pair of y in our training knowledge base, or NA if no such relation exists. We will see later why $\mathcal{F}_m$ needs knowledge of y_{truth}.

For all approaches $\mathcal{F}_{\mathrm{r}}(y)$ is fixed to be

$$\mathcal{F}_{\mathrm{r}}(y) = \begin{cases} 1 & y = y_{\mathrm{truth}} \\ -1 & \text{otherwise} \end{cases}.$$

That is, a match with the true relation increases the truth score by one, otherwise the score is decreased by one.

4.1 Distant Supervision

A distant supervision baseline akin to [18] and [3] can be easily implemented using SampleRank. In this case we simply consider all variables Z_i to be fixed: $Z_i = 0$ if the corresponding relation variable y is NA, and $Z_i = 1$ otherwise. Inference then only considers the Y variables.

4.2 Joint Supervision

We essentially propose two modifications to the original distant supervision approach. First, jointly infer mentions and relations, and second, relax the assumption that each candidate mention is indeed an actual mention. We can easily implement the first modification by using the distant supervision truth function

$$\mathcal{F}_m^{\text{distant}}(\mathbf{z}, y_{\text{truth}}) = \begin{cases} \|\mathbf{z}\| & y_{\text{truth}} \neq \text{NA} \\ -\|\mathbf{z}\| & \text{otherwise} \end{cases} . \tag{1}$$

That is, if the two entities in question are related in our training knowledge base (i.e.,$y_{\text{truth}} \neq \text{NA}$), every active relation mention is encouraged. Otherwise every relation mention is preferred to be inactive.

4.3 Expressed-at-least-once Supervision

SampleRank allows us to naturally incorporate the expressed-at-least-once assumption we presented in section 1. We simply use the following truth function for relation mentions:

$$\mathcal{F}_{\text{m}}^{\text{once}}(\mathbf{z}, y_{\text{truth}}) = \begin{cases} 1 & y_{\text{truth}} \neq \text{NA} \wedge \|\mathbf{z}\| \geq 1 \\ -1 & y_{\text{truth}} \neq \text{NA} \wedge \|\mathbf{z}\| = 0 \\ -\|\mathbf{z}\| & \text{otherwise} \end{cases} . \tag{2}$$

That is, if the true relation type y_{truth} is not NA, an assignment to the relation mention variables has maximal rank if at least one mention is active. In case the pair of entities is not related according to Freebase, an assignment is discouraged proportional to the amount of active relation mentions.

5 Related Work

While much work on relation extraction has focused on fully-supervised approaches [11, 9], we work in the framework of distant supervision. In this context existing work has primarily relied on the distant supervision assumption [18, 3, 8]. [28] use a more sophisticated heuristic to decide which candidates are relation mentions. This heuristic is tailored to extracting infobox attributes from Wikipedia articles. By contrast, our method is designed for relations between entity pairs mentioned in newswire.

Our work is based on constraint-driven semi-supervised learning [6, 16]. Generally, constraint-driven methods are used when (a) only a small set of labelled training instances are available, and (b) there are some (hard or soft) constraints that are known to hold across the unlabelled data. These constraints are then used as additional source of supervision. To our knowledge, constraint-driven learning has not been applied to information extraction under distant supervision.

There are many approaches to train undirected models with latent variables besides SampleRank. An alternative method is the latent perceptron [26]. Here in each iteration the MAP assignment for a set of latent variables is predicted. Then the non-latent label variable is deterministically inferred from the latent variables and compared to the gold label. If labels disagree, weights are updated in a manner similar to the regular perceptron. For us this approach would not directly be helpful for two reasons: First, it is not clear how we could incorporate prior knowledge such as the at-least-once assumption; second, in our case the relation between latent and non-latent label variables is not deterministic.

Contrastive estimation [25] is another related approach. Here parameters of a model with latent variables are learned by maximizing the ratio between the probability of the observed data x and the sum of probabilities over "sensible but wrong" x' in a local neighborhood of x. In some sense, what we do is similar if we consider the relation variables as observed x, and their negated state $1-x$ as local neighborhood. However, we do not maximize probability, but match rankings. Also, our at-least-once constraint would need to be formulated as a (complex) prior over the set of hidden mention variables.

Finally, our approach can be seen as a novel take on Multi-Instance Learning [10]. Here training instances are divided into bags, and we only know that some bags contain at least one positive example while the remaining ones contain only negative examples. Our constraint-driven approach works in the same setting. However, it also allows to include additional constraints that a user may have. For example, we could inject constraints corresponding to the heuristics used in[28]. Moreover, our approach is the first that can discriminatively train general factor graphs. This will be important for future work such as joint coreference, entity type and relation extraction. Here the graphical structure will get more involved.

6 Evaluation

Our experiments aim to provide evidence for the following hypothesis: explicitly relaxing the distant supervision assumption in a probabilistic model can lead to substantial increase in precision. To this end we follow [18] and compare it against both the distant and the supervised joint model using a held-out set of relations and documents. Since this type of evaluation suffers from false negatives (some negative relations we extracted may in fact be positive but not in the knowledge base), we also manually evaluate the predicted relations.

Note that for all our models we use exactly the same feature set. However, we choose the best number of training epochs for each model individually. Also note that for training we need negative instances. For this purpose we generally pick 10% of the entity pairs that appear in the same sentence but are not related according to Freebase.

6.1 Data

Following [18] we use Freebase as our distant supervision source. Freebase is an online database that stores facts about entities and their relations. We extract all relations from a December 2009 snapshot of Freebase. Four categories of Freebase relations are used: "people", "business", "person", and "location". These types of relations are chosen because we expect that they appear frequently in the newswire corpus. In total this provided over 3.2 million relation instances of 430 Freebase relation types, and over 1.8 million entities that participate in these relations.

For the choice of text corpus we divert from [18] and use the New York Times corpus [22]. This allows us to evaluate our approach when the distant supervision source is external. The New York Times data contains over 1.8 million articles written and published by the New York Times between January 1, 1987 and June 19, 2007. Generally, we find that Freebase entities are frequently mentioned in the NYT corpus. For example, for the year 2007 about 700,000 mentions of Freebase entities appear in the corpus. Naturally, we observe a smaller number of cases in which two related entities are mentioned in the same sentence: again for the year 2007 we find about 170,000 such cases.

6.2 Preprocessing

In order to find entity mentions in text we first used the Stanford named entity recognizer [12]. The NER tagger segments each document into sentences and classifies each token into four categories: PERSON, ORGANIZATION, LOCATION and NONE. We treat consecutive tokens which share the same category as single entity mention. Then we associate these mentions with Freebase entities. This is achieved by simply performing a string match between entity mention phrase and the canonical names of entities in Freebase.

Next, for each pair of entities participating in a relation of our training KB, we traverse the text corpus and find sentences in which the two entities co-occur. Each pair of entity mentions is considered to be a relation mention candidate. For each such candidate we extract a set of features (see section 3). The types of features are essentially corresponding to the ones used by [18]: we used lexical, Part-Of-Speech (POS), named entity and syntactic features (i.e. features obtained from the dependency parsing tree of a sentence). We applied the openNLP POS tagger[4] to obtain POS tags and used the MaltParser [20] for dependency parsing.

6.3 Held-out Evaluation

Following [18] we divide the Freebase relations into two parts, one for training and one for testing. The former is aligned to the years 2005-2006 of the NYT corpus, the latter to the year 2007. As candidate relation instances we use all

[4] available at http://opennlp.sourceforge.net/

pairs of Freebase entities that are at least once mentioned in the same sentence. Note that the amount of Freebase relations mentioned in the training set (4700) and test set (1950) is relatively low due to a smaller overlap between Freebase and the New York Times. Hence we cannot evaluate our models with the same quantity of data as [18].

In figure 2 we compare the precision and recall curve for the baseline distant-supervision model (distant), the supervised joint model (joint) and the distant model with expressed-at-least-once assumption (at-least-once). The curve is constructed by ranking the predicted relation instances using their loglinear score. For the distant supervision baseline this score is first normalized by the number of mentions.[5] We traverse this list from high score to low score, and measure precision and recall at each position.

We can see that the model with expressed-at-least-once assumption is consistently outperforming the distant supervision baseline and the supervised joint model. This suggests that the at-least-once model has the best sense of how relations that are already contained in Freebase are expressed in NYT data. However, it does not necessarily mean that it knows best how relations are expressed that are not yet in Freebase. We address this in the next section.

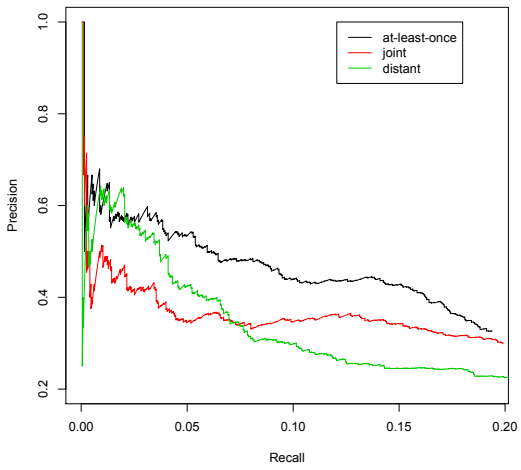

Fig. 2. Precision and recall for the held out data and three approaches: distant supervision, joint supervision, and at-least-once supervision.

[5] This yielded the best results for the baseline. We also tried to use conditional probabilities to rank. This lead to poor results because SampleRank training has no probabilistic interpretation.

6.4 Manual Evaluation

For manual evaluation all Freebase entities and relations are used as training instances. As candidate relation instances we choose those entity pairs which appear together in the NYT test set, but for which at least one participating entity is not in Freebase. This means that there is no overlap between the held-out and manual candidates. Then we apply our models to this test set, and asked two annotators to evaluate the top 1000 predicted relation instances.

We cannot calculate recall in this case, since we cannot provide all relation instances expressed in our corpus. Instead we use a "Precision at K" metric with respect to the ranked lists we extracted in section 6.3. Figure 3 shows the precisions for values of K between 0 and 1000.

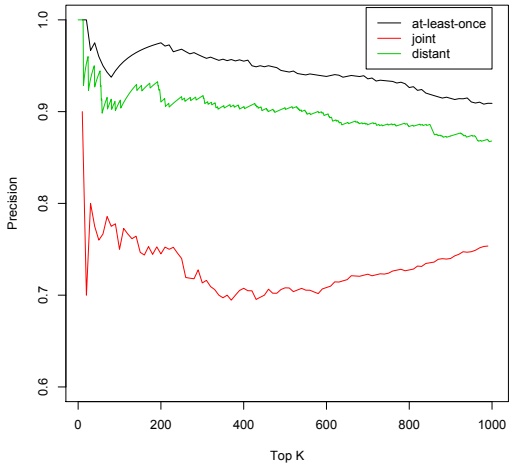

Fig. 3. Precision at K for manually evaluated predictions.

We first note that the precision is much higher for manual evaluation than for held-out evaluation. This shows that false negatives in Freebase are an issue when doing held-out evaluation. Many of the false positives we predict are in fact true relation instances and just do not appear in Freebase.

For manual evaluation the at-least-once model is still the winner. At $K = 1000$ we observe a precision of 91% for at-least-once supervision, 87% for distant supervision. This amounts to an error reduction rate of 31%. The sign test shows that the at-least-once model is significantly better than the distant supervision model, with $p \ll 0.05$. We also note that despite using the same assumption, the joint model performs much worse than the distant supervision approach in this scenario. Learning a model of relations and mentions is inherently more difficult. Using a wrong assumption will hence more likely hurt performance.

Does the at-least-once model help to fix the type of error discussed in section 2? To find out, we inspect the results of the *founded* relation. When we consider the top 100 instances of this relation for the distant supervision system, we observe a precision of 45%. Compare this to 72% precision for the at-least-once model.

On close inspection, most of the distant supervision errors for the *founded* relation stem from cases where patterns such as "director of" appear. They indicate that the person in question works for the given company. Because in the training set such patterns often appear when a person is a founder, they gain high weights and appear high up in the ranking.

The at-least-once model also makes this type of error, but to a much lesser extent. This is not surprising if we consider that for training instances with only one mention, the at-least-once and distant supervision assumptions are equivalent. Assume that according to Freebase, person A founded company B. If there is only one mention of A and B in the NYT training corpus, it has to be a mention of *founded*, even if the sentence says "director-of". This leads to a higher weight for "director-of" as *founded* pattern.

7 Conclusion

This paper presents a novel approach to extract relations from text without explicit training annotation. Recent approaches assume that every sentence that mentions two related entities expresses the corresponding relation. Motivated by the observation that this assumptions frequently does not hold, in particular when considering external knowledge bases, we propose to relax it. Instead we assume that at least one sentence which mentions two related entities expresses the corresponding relation.

To model this assumption we make two contributions. First, we introduce a novel undirected graphical model that captures both the task of predicting relations between entities, and the task of predicting which sentences express these relations. Second, we propose to train this graphical model by framing distant supervision as an instance of constraint-driven semi-supervision. In particular, we use SampleRank, a discriminative learning algorithm for large factor graphs, and inject the expressed-at-least-once assumption through a truth function.

Empirically this approach improves precision substantially. For the task of extracting 1000 Freebase relation instances from the New York Times, we measure a precision of 91% for at-least-once supervision, and 87% for distant supervision. This amounts to an error reduction rate of 31%.

A crucial aspect of our approach is its extensibility: framed exclusively in terms of factor graphs and truth functions, it is conceptually easy to apply it to larger tasks such as the joint prediction of relations and entity types. In future work we will exploit this aspect and extend our model to jointly perform other relevant tasks for the automatic construction of KBs.

Acknowledgements This work was supported in part by the Center for Intelligent Information Retrieval, in part by SRI International subcontract #27-001338 and ARFL prime contract #FA8750-09-C-0181, in part by The Central Intelligence Agency, the National Security Agency and National Science Foundation under NSF grant #IIS-0326249, and in part by UPenn NSF medium IIS-0803847. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

References

1. Bellare, K., McCallum, A.: Generalized expectation criteria for bootstrapping extractors using record-text alignment. In: Empirical Methods in Natural Language Processing (EMNLP). pp. 131–140 (2009)
2. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: SIGMOD '08: Proceedings of the 2008 ACM SIGMOD international conference on Management of data. pp. 1247–1250. ACM, New York, NY, USA (2008)
3. Bunescu, R.C., Mooney, R.J.: Learning to extract relations from the web using minimal supervision. In: Proceedings of the 45rd Annual Meeting of the Association for Computational Linguistics (ACL' 07) (2007)
4. Chang, M.W., Goldwasser, D., Roth, D., Tu, Y.: Unsupervised constraint driven learning for transliteration discovery. In: NAACL '09: Proceedings of Human Language Technologies: Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 299–307 (2009)
5. Chang, M.W., Ratinov, L., Rizzolo, N., Roth, D.: Learning and inference with constraints. In: AAAI Conference on Artificial Intelligence. pp. 1513–1518. AAAI Press (2008)
6. Chang, M.W., Ratinov, L., Roth, D.: Guiding semi-supervision with constraint-driven learning. In: Annual Meeting of the Association for Computational Linguistics (ACL). pp. 280–287 (2007)
7. Collins, M.: Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In: Proceedings of the Conference on Empirical methods in natural language processing (EMNLP '02). vol. 10, pp. 1–8 (2002)
8. Craven, M., Kumlien, J.: Constructing biological knowledge-bases by extracting information from text sources. In: Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology. pp. 77–86. Germany (1999)
9. Culotta, A., McCallum, A.: Joint deduplication of multiple record types in relational data. In: Proceedings of the 14th ACM international conference on Information and knowledge management (CIKM '05). pp. 257–258. ACM, New York, NY, USA (2005)
10. Dietterich, T., Lathrop, R., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence 89(1-2), 31–71 (1997)
11. Dimitry Zelenko, C.A., Richardella, A.: Kernel methods for relation extraction. JMLR 3(6), 1083 – 1106 (2003)
12. Finkel, J.R., Grenager, T., Manning, C.: 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). pp. 363–370 (June 2005)

13. Geman, S., Geman, D.: Stochastic relaxation, gibbs distributions, and the bayesian restoration of images pp. 452–472 (1990)
14. Jensen, C.S., Kong, A., Kjaerulff, U.: Blocking gibbs sampling in very large probabilistic expert systems. International Journal of Human Computer Studies. Special Issue on Real-World Applications of Uncertain Reasoning 42, 647–666 (1993)
15. Lafferty, J.D., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: International Conference on Machine Learning (ICML) (2001)
16. Mann, G.S., McCallum, A.: Generalized expectation criteria for semi-supervised learning of conditional random fields. In: Annual Meeting of the Association for Computational Linguistics (ACL). pp. 870–878 (2008)
17. McCallum, A., Schultz, K., Singh, S.: Factorie: Probabilistic programming via imperatively defined factor graphs. In: Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C.K.I., Culotta, A. (eds.) Advances in Neural Information Processing Systems 22, pp. 1249–1257 (2009)
18. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proceedings of the 47rd Annual Meeting of the Association for Computational Linguistics (ACL' 09). pp. 1003–1011. Association for Computational Linguistics (2009)
19. Morgan, A.A., Hirschman, L., Colosimo, M., Yeh, A.S., Colombe, J.B.: Gene name identification and normalization using a model organism database. J. of Biomedical Informatics 37(6), 396–410 (2004)
20. Nivre, J., Hall, J., Nilsson, J.: Memory-based dependency parsing. In: Proceedings of CoNLL. pp. 49–56 (2004)
21. Rohanimanesh, K., Wick, M., McCallum, A.: Inference and learning in large factor graphs with a rank based objective. Tech. Rep. UM-CS-2009-08, University of Massachusetts, Amherst (2009)
22. Sandhaus, E.: The New York Times Annotated Corpus. Linguistic Data Consortium, Philadelphia (2008)
23. Singh, S., Schultz, K., McCallum, A.: Bi-directional joint inference for entity resolution and segmentation using imperatively-defined factor graphs. In: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD). pp. 414–429 (2009)
24. Singh, S., Yao, L., Riedel, S., McCallum, A.: Constraint-driven rank-based learning for information extraction. In: North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT) (2010)
25. Smith, N.A., Eisner, J.: Contrastive estimation: training log-linear models on unlabeled data. In: ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. pp. 354–362. Association for Computational Linguistics, Morristown, NJ, USA (2005)
26. Sun, X., Matsuzaki, T., Okanohara, D., Tsujii, J.: Latent variable perceptron algorithm for structured classification. In: IJCAI'09: Proceedings of the 21st international jont conference on Artifical intelligence. pp. 1236–1242. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2009)
27. Wick, M., Rohanimanesh, K., Culotta, A., McCallum, A.: Samplerank: Learning preferences from atomic gradients. In: Neural Information Processing Systems (NIPS), Workshop on Advances in Ranking (2009)
28. Wu, F., Weld, D.S.: Autonomously semantifying wikipedia. In: Proceedings of the 16th ACM international conference on information and knowledge management (CIKM '07). pp. 41–50. ACM, New York, NY, USA (2007)