# Distant Supervision for Relation Extraction with an Incomplete Knowledge Base

**Bonan Min, Ralph Grishman, Li Wan**
New York University
New York, NY 10003
{min,grishman,wanli}
@cs.nyu.edu

**Chang Wang, David Gondek**
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598
{wangchan,dgondek}
@us.ibm.com

## Abstract

Distant supervision, heuristically labeling a corpus using a knowledge base, has emerged as a popular choice for training relation extractors. In this paper, we show that a significant number of "*negative*" examples generated by the labeling process are *false negatives* because the knowledge base is incomplete. Therefore the heuristic for generating negative examples has a serious flaw. Building on a state-of-the-art distantly-supervised extraction algorithm, we proposed an algorithm that learns from only *positive* and *unlabeled* labels at the pair-of-entity level. Experimental results demonstrate its advantage over existing algorithms.

## 1 Introduction

Relation Extraction is a well-studied problem (Miller et al., 2000; Zhou et al., 2005; Kambhatla, 2004; Min et al., 2012a). Recently, Distant Supervision (DS) (Craven and Kumlien, 1999; Mintz et al., 2009) has emerged to be a popular choice for training relation extractors without using manually labeled data. It automatically generates training examples by labeling relation mentions[1] in the source corpus according to whether the argument pair is listed in the target relational tables in a knowledge base (KB). This method significantly reduces human efforts for relation extraction.

The labeling heuristic has a serious flaw. Knowledge bases are usually highly incomplete. For exam-

ple, 93.8% of *persons* from Freebase[2] have no *place of birth*, and 78.5% have no *nationality* (section 3). Previous work typically assumes that if the argument entity pair is not listed in the KB as having a relation, all the corresponding relation mentions are considered *negative* examples.[3] This crude assumption labeled many entity pairs as *negative* when in fact some of their mentions express a relation. The number of such *false negative matches* even exceeds the number of *positive* pairs, by 3 to 10 times, leading to a significant problem for training. Previous approaches (Riedel et al., 2010; Hoffmann et al., 2011; Surdeanu et al., 2012) bypassed this problem by heavily under-sampling the "*negative*" class.

We instead deal with a learning scenario where we only have entity-pair level labels that are either *positive* or *unlabeled*. We proposed an extension to Surdeanu et al. (2012) that can train on this dataset. Our contribution also includes an analysis on the incompleteness of Freebase and the *false negative match* rate in two datasets of labeled examples generated by DS. Experimental results on a realistic and challenging dataset demonstrate the advantage of the algorithm over existing solutions.

## 2 Related Work

Distant supervision was first proposed by Craven and Kumlien (1999) in the biomedical domain.

---

[1] An occurrence of a pair of entities with the source sentence.

[2] Freebase is a large collaboratively-edited KB. It is available at http://www.freebase.com.

[3] There are variants of labeling heuristics. For example, Surdeanu et al. (2011) and Sun et al. (2011) use a pair $< e, v >$ as a negative example, when it is not listed in Freebase, but $e$ is listed with a different $v'$. These assumptions are also problematic in cases where the relation is not functional.

Since then, it has gain popularity (Mintz et al., 2009; Bunescu and Mooney, 2007; Wu and Weld, 2007; Riedel et al., 2010; Hoffmann et al., 2011; Surdeanu et al., 2012; Nguyen and Moschitti, 2011). To tolerate noisy labels in positive examples, Riedel et al. (2010) use Multiple Instance Learning (MIL), which assumes only at-least-one of the relation mentions in each "bag" of mentions sharing a pair of argument entities which bears a relation, indeed expresses the target relation. MultiR (Hoffmann et al., 2011) and Multi-Instance Multi-Label (MIML) learning (Surdeanu et al., 2012) further improve it to support multiple relations expressed by different sentences in a bag. Takamatsu et al. (2012) models the probabilities of a pattern showing relations, estimated from the heuristically labeled dataset. Their algorithm removes mentions that match low-probability patterns. Sun et al. (2011) and Min et al. (2012b) also estimate the probablities of patterns showing relations, but instead use them to relabel examples to their most likely classes. Their approach can correct highly-confident false negative matches.

## 3 Problem Definition

**Distant Supervision**: Given a KB $\mathcal{D}$ (a collection of relational tables $r(e_1, e_2)$, in which $r \epsilon R$ ($R$ is the set of relation labels), and $< e_1, e_2 >$ is a pair of entities that is known to have relation $r$) and a corpus $C$, the key idea of distant supervision is that we align $\mathcal{D}$ to $C$, label each *bag*[4] of relation mentions that share argument pair $< e_1, e_2 >$ with $r$, otherwise *OTHER*. This generates a dataset that has labels on entity-pair (bag) level. Then a relation extractor is trained with single-instance learning (by assuming all mentions have the same label as the bag), or Multiple-Instance Learning (by assuming at-least-one of the mentions expresses the bag-level label), or Multi-Instance Multi-Label learning (further assuming a bag can have multiple labels) algorithms. All of these works treat the *OTHER* class as examples that are labeled as *negative*.

**The incomplete KB problem**: KBs are usually incomplete because they are manually constructed, and it is not possible to cover all human knowledge

---

[4]A bag is defined as a set of relation mentions sharing the same entity pair as relation arguments. We will use the terms *bag* and *entity pair* interchangeably in this paper.

nor stay current. We took frequent relations, which involve an entity of type PERSON, from Freebase for analysis. We define the incompleteness $\partial(r)$ of a relation $r$ as follows:

$$\partial(r) = \frac{|\{e\}| - |\{e | \exists e', s.t. r(e, e') \epsilon \mathcal{D}\}|}{|\{e\}|}$$

$\partial(r)$ is the percentage of all *persons* $\{e\}$ that do not have an attribute $e'$ (with which $r(e, e')$ holds). Table 1 shows that 93.8% of *persons* have no *place of birth*, and 78.5% of them have no *nationality*. These are must-have attributes for a person. This shows that Freebase is highly incomplete.

| Freebase relation types | Incompleteness |
|---|---|
| /people/person/education | 0.792 |
| /people/person/employment_history | 0.923 |
| /people/person/nationality* | 0.785 |
| /people/person/parents* | 0.988 |
| /people/person/place_of_birth* | 0.938 |
| /people/person/places_lived* | 0.966 |

Table 1: The incompleteness of Freebase (* are must-have attributes for a person).

We further investigate the rate of *false negative matches*, as the percentage of entity-pairs that are not listed in Freebase but one of its mentions generated by DS does express a relation in the target set of types. We randomly picked 200 *unlabeled* bags[5] from each of the two datasets (Riedel et al., 2010; Surdeanu et al., 2012) generated by DS, and we manually annotate all relation mentions in these bags. The result is shown in Table 2, along with a few examples that indicate a relation holds in the set of false negative matches (bag-level). Both datasets have around 10% false negative matches in the *unlabeled* set of bags. Taking into consideration that the number of *positive* bags and *unlabeled* bags are highly imbalanced (1:134 and 1:37 in the *Riedel* and *KBP* dataset respectively, before undersampling the *unlabeled* class), the number of false negative matches are 11 and 4 times the number of *positive* bags in *Reidel* and *KBP* dataset, respectively. Such a large ratio shows false negatives do have a significant impact on the learning process.

## 4 A semi-supervised MIML algorithm

Our goal is to model the bag-level label noise, caused by the incomplete KB problem, in addition

---

[5]85% and 95.7% of the bags in the Riedel and KBP datasets have only one relation mention.

| Dataset (training) | # positive bags | # positive : # unlabeled | % are false negatives | # positive : # false negative | has human assessment | Examples of false negative mentions |
|---|---|---|---|---|---|---|
| *Riedel* | 4,700 | 1:134(BD*) | 8.5% | 1:11.4 | no | (/location/location/contains)... in **Brooklyn** 's **Williamsburg**. |
| | | | | | | (/people/person/place_lived) **Cheryl Rogowski** , a farmer from **Orange County** ... |
| *KBP* | 183,062 | 1:37(BD*) | 11.5% | 1:4 | yes | (per:city_of_birth) **Juan Martn Maldacena** (born September 10, 1968) is a theoretical physicist born in **Buenos Aires** |
| | | | | | | (per:employee_of)**Dave Matthews**, from the **ABC News**, ... |

Table 2: False negative matches on the *Riedel* (Riedel et al., 2010) and *KBP* dataset (Surdeanu et al., 2012). All numbers are on bag (pairs of entities) level. *BD\** are the numbers before downsampling the *negative* set to 10% and 5% in *Riedel* and *KBP* dataset, respectively.

to modeling the instance-level noise using a 3-layer MIL or MIML model (e.g., Surdeanu et al. (2012)). We propose a 4-layer model as shown in Figure 1.

The input to the model is a list of $n$ bags with a vector of binary labels, either *Positive (P)*, or *Unlabled (U)* for each relation $r$. Our model can be viewed as a ==semi-supervised[6]== framework that extends a state-of-the-art Multi-Instance Multi-Label (MIML) model (Surdeanu et al., 2012). Since the input to previous MIML models are bags with per-relation binary labels of either *Positive (P)* or *Negative (N)*, we add a set of latent variables $\ell$ which models the true bag-level labels, to bridge the observed bag labels $\mathbf{y}$ and the MIML layers. ==We consider this as our main contribution to the model.== Our hierarchical model is shown in Figure 1.
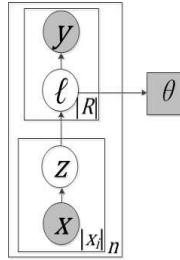


Figure 1: Plate diagram of our model.

Let $i, j$ be the index in the bag and mention level, respectively. Following Surdeanu et al. (2012), we model mention-level extraction $p(z_{ij}^r | \mathbf{x}_{ij}; \mathbf{w}_z)$ and multi-instance multi-label aggregation $p(\ell_i^r | \mathbf{z}_i; \mathbf{w}_\ell^r)$ in the bottom 3 layers. We define:

- $r$ is a relation label. $r \epsilon R \cup \{OTHER\}$, in which *OTHER* denotes no relation expressed.
- $y_i^r \epsilon \{P, U\}$: $r$ holds for $i$th bag or the bag is unlabeled.

---
[6]We use the term *semi-supervised* because the algorithm uses *unlabeled* bags but existing solutions requires bags to be labeled either *positive* or *negative*.

- $\ell_i^r \epsilon \{P, N\}$: a hidden variable that denotes whether $r$ holds for the $i$th bag.
- $\theta$ is an observed constant controlling the total number of bags whose latent label is *positive*.

We define the following conditional probabilities:

- $p(y_i^r | \ell_i^r) = \begin{cases} 1/2 & \text{if } y_i^r = P \wedge \ell_i^r = P; \\ 1/2 & \text{if } y_i^r = U \wedge \ell_i^r = P; \\ 1 & \text{if } y_i^r = U \wedge \ell_i^r = N; \\ 0 & \text{otherwise ;} \end{cases}$

 It encodes the constraints between true bag-level labels and the entity pair labels in the KB.

- $p(\theta | \ell) \sim \mathcal{N}(\frac{\sum_{i=1}^n \sum_{r \epsilon R} \delta(\ell_i^r, P)}{n}, \frac{1}{k})$ where $\delta(x, y) = 1$ if $x = y$, 0 otherwise. $k$ is a large number. $\theta$ is the fraction of the bags that are positive. It is an observed parameter that depends on both the source corpus and the KB used.

Similar to Surdeanu et al. (2012), we also define the following parameters and conditional probabilities (details are in Surdeanu et al. (2012)):

- $z_{ij} \epsilon R \cup \{OTHER\}$: a latent variable that denotes the relation type of the $j$th mention in the $i$th bag.
- $\mathbf{x}_{ij}$ is the feature representation of the $j$th relation mention in the $i$th bag. We use the set of features in Surdeanu et al. (2012).
- $\mathbf{w}_z$ is the weight vector for the multi-class relation mention-level classifier.
- $\mathbf{w}_\ell^r$ is the weight vector for the $r$th binary top-level aggregation classifier (from mention labels to bag-level prediction). We use $\mathbf{w}_\ell$ to represent $\mathbf{w}_\ell^1, \mathbf{w}_\ell^2, ... \mathbf{w}_\ell^{|R|}$.
- $p(\ell_i^r | \mathbf{z}_i; \mathbf{w}_\ell^r) \sim Bern(f_\ell(\mathbf{w}_\ell^r, \mathbf{z}_i))$ where $f_\ell$ is probability produced by the $r$th top-level classifier, from the mention-label level to the bag-label level.
- $p(z_{ij}^r | \mathbf{x}_{ij}; \mathbf{w}_z) \sim Multi(f_z(\mathbf{w}_z, \mathbf{x}_{ij}))$ where $f_z$

is probability produced by the mention-level classifier, from the mentions to the mention-label level.[7]

## 4.1 Training

We use hard <mark>Expectation-Maximization (EM)</mark> algorithm for training the model. Our objective function is to maximize log-likelihood:

$$L(\mathbf{w}_z, \mathbf{w}_\ell) = logp(\mathbf{y}, \theta | \mathbf{x}; \mathbf{w}_z, \mathbf{w}_\ell)$$

$$= log \sum_\ell p(\mathbf{y}, \theta, \ell | \mathbf{x}; \mathbf{w}_z, \mathbf{w}_\ell)$$

Since solving it exactly involves exploring an exponential assignment space for $\ell$, we approximate and iteratively set $\ell^* = \arg_\ell \max p(\ell | \mathbf{y}, \theta, \mathbf{x}; \mathbf{w}_z, \mathbf{w}_\ell)$

$$p(\ell | \mathbf{y}, \theta, \mathbf{x}, \mathbf{w}_z, \mathbf{w}_\ell) \propto p(\mathbf{y}, \theta, \ell | \mathbf{x}; \mathbf{w}_z, \mathbf{w}_\ell)$$

$$= p(\mathbf{y}, \theta | \ell, \mathbf{x}) p(\ell | \mathbf{x}; \mathbf{w}_z, \mathbf{w}_\ell)$$

$$= p(\mathbf{y} | \ell) p(\theta | \ell) p(\ell | \mathbf{x}; \mathbf{w}_z, \mathbf{w}_\ell)$$

Rewriting in log form:

$$logp(\ell | \mathbf{y}, \theta, \mathbf{x}, \mathbf{w}_z, \mathbf{w}_\ell)$$

$$= logp(\mathbf{y} | \ell) + logp(\theta | \ell) + logp(\ell | \mathbf{x}; \mathbf{w}_z, \mathbf{w}_\ell)$$

$$= \sum_{i=1}^{n} \sum_{r \epsilon R} logp(y_i^r | \ell_i^r) - k\left(\frac{\sum_{i=1}^{n} \sum_{r \epsilon R} \delta(\ell_i^r, P)}{n} - \theta\right)^2$$

$$+ \sum_{i=1}^{n} \sum_{r \epsilon R} logp(\ell_i^r | \mathbf{x}_i; \mathbf{w}_z, \mathbf{w}_\ell) + const$$

---

**Algorithm 1** Training (**E-step**:2-11; **M-step**:12-15)

1: **for** $i = 1, 2$ to $T$ **do**
2: $\quad \ell_i^r \leftarrow N$ for all $y_i^r = U$ and $r \epsilon R$
3: $\quad \ell_i^r \leftarrow P$ for all $y_i^r = P$ and $r \epsilon R$
4: $\quad I = \{< i, r > | \ell_i^r = N\}; I' = \{< i, r > | \ell_i^r = P\}$
5: $\quad$ **for** $k = 0, 1$ to $\theta n - |I'|$ **do**
6: $\qquad < i', r' > = \arg \max_{<i,r> \epsilon I} p(\ell_i^r | \mathbf{x}_i; \mathbf{w}_z, \mathbf{w}_\ell)$
7: $\qquad \ell_{i'}^{r'} \leftarrow P; I = I \backslash \{< i', r' >\}$
8: $\quad$ **end for**
9: $\quad$ **for** $i = 1, 2$ to $n$ **do**
10: $\qquad \mathbf{z}_i^* = \arg \max_{\mathbf{z}_i} p(\mathbf{z}_i | \ell_i, \mathbf{x}_i; \mathbf{w}_z, \mathbf{w}_\ell)$
11: $\quad$ **end for**
12: $\quad \mathbf{w}_z^* = \arg \max_{\mathbf{w}_z} \sum_{i=1}^{n} \sum_{j=1}^{|\mathbf{x}_i|} logp(z_{ij} | \mathbf{x}_{ij}, \mathbf{w}_z)$
13: $\quad$ **for all** $r \epsilon R$ **do**
14: $\qquad \mathbf{w}_\ell^{r(*)} = \arg \max_{w_\ell^r} \sum_{i=1}^{n} p(\ell_i^r | \mathbf{z}_i, \mathbf{w}_\ell^r)$
15: $\quad$ **end for**
16: **end for**
17: **return** $\mathbf{w_z}, \mathbf{w}_\ell$

---

In the E-step, we do a greedy search (steps 5-8 in algorithm 1) in all $p(\ell_i^r | \mathbf{x}_i; \mathbf{w}_z, \mathbf{w}_\ell)$ and update $\ell_i^r$ until the second term is maximized. $\mathbf{w}_z, \mathbf{w}_\ell$ are the model weights learned from the previous iteration.

After fixed $\ell$, we seek to maximize:

$$logp(\ell | \mathbf{x}_i; \mathbf{w}_z, \mathbf{w}_\ell) = \sum_{i=1}^{n} logp(\ell_i | \mathbf{x}_i; \mathbf{w}_z, \mathbf{w}_\ell)$$

$$= \sum_{i=1}^{n} log \sum_{z_i} p(\ell_i, \mathbf{z}_i | \mathbf{x}_i; \mathbf{w}_z, \mathbf{w}_\ell)$$

which can be solved with an approximate solution in Surdeanu et al. (2012) (step 9-11): update $\mathbf{z}_i$ independently with: $\mathbf{z}_i^* = \arg \max_{\mathbf{z}_i} p(\mathbf{z}_i | \ell_i, \mathbf{x}_i; \mathbf{w}_z, \mathbf{w}_\ell)$. More details can be found in Surdeanu et al. (2012).

In the M-step, we retrain both of the mention-level and the aggregation level classifiers.

The full EM algorithm is shown in algorithm 1.

## 4.2 Inference

Inference on a bag $\mathbf{x}_i$ is trivial. For each mention:

$$z_{ij}^* = \arg_{z_{ij} \epsilon R \cup \{OTHER\}} \max p(z_{ij} | x_{ij}, \mathbf{w}_z)$$

Followed by the aggregation (directly with $\mathbf{w}_\ell$):

$$y_i^{r(*)} = \arg_{y_i^r \epsilon \{P,N\}} \max p(y_i^r | \mathbf{z}_i; \mathbf{w}_\ell^r)$$

## 4.3 Implementation details

We <mark>implement</mark> our model on top of the <mark>MIML(Surdeanu et al., 2012) code base.</mark>[8] We use the same mention-level and aggregate-level feature sets as Surdeanu et al. (2012). We adopt the same idea of using cross validation for the E and M steps to avoid overfitting. We initialize our algorithm by sampling 5% *unlabeled* examples as *negative*, in essence using 1 epoch of MIML to initialize. Empirically it performs well.

## 5 Experiments

**Data set**: We use the KBP (Ji et al., 2011) dataset[9] prepared and publicly released by Surdeanu et al. (2012) for our experiment since it is 1) large and realistic, 2) publicly available, 3) most importantly, it is the only dataset that has associated human-labeled ground truth. Any KB held-out evaluation without manual assessment will be significantly affected by KB incompleteness. In KBP

---

[7]All classifiers are implemented with L2-regularized logistic regression with Stanford CoreNLP package.

[8]Available at http://nlp.stanford.edu/software/mimlre.shtml
[9]Available from Linguistic Data Consortium (LDC). http://projects.ldc.upenn.edu/kbp/data/
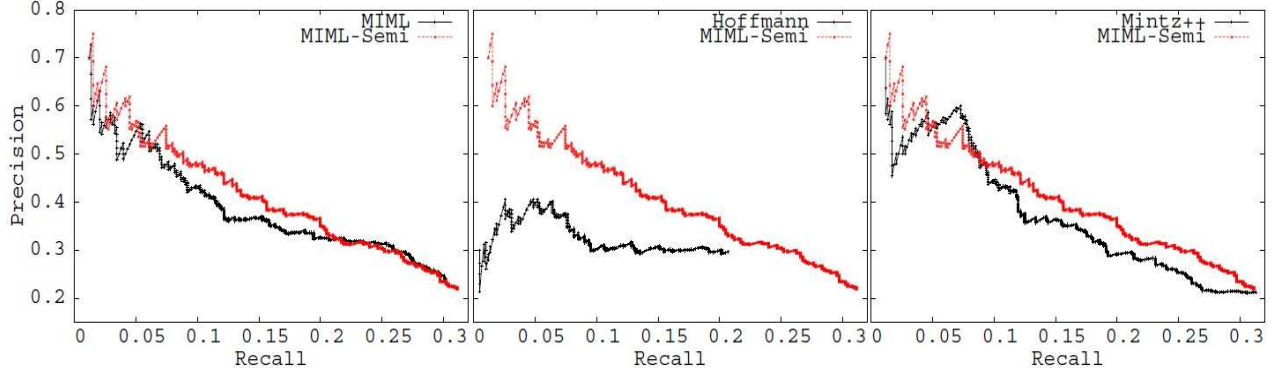
Figure 2: Performance on the KBP dataset. The figures on the left, middle and right show *MIML*, *Hoffmann*, and *Mintz++* compared to the same *MIML-Semi* curve, respectively. *MIML-Semi* is shown in red curves (lighter curves in black and white) while other algorithms are shown in black curves (darker curves in black and white).

dataset, the training bags are generated by mapping Wikipedia (http://en.wikipedia.org) infoboxes (after merging similar types following the KBP 2011 task definition) into a large unlabeled corpus (consisting of 1.5M documents from the KBP source corpus and a complete snapshot of Wikipedia). The KBP shared task provided 200 query named entities with their associated slot values (in total several thousand pairs). We use 40 queries as development dataset (*dev*), and the rest (160 queries) as evaluation dataset. We set $\theta = 0.25$ by tuning on the *dev* set and use it in the experiments. For a fair comparison, we follow Surdeanu et al. (2012) and begin by downsampling the "negative" class to 5%. We also set T=8 and use the following noisy-or (for $i$th bag) of mention-level probability to rank predicted types ($r$) of pairs and plot the precision-recall curves for all experiments.

$$Prob_i(r) = 1 - \prod_j \left(1 - p(z_{ij} = r | \mathbf{x}_{ij}; \mathbf{w}_z)\right)$$

**Evaluation**: We compare our algorithm (*MIML-semi*) to three algorithms: 1) *MIML* (Surdeanu et al., 2012), the Multiple-Instance Multiple Label algorithm which labels the bags directly with the KB ($y = \ell$). 2) *MultiR* (denoted as *Hoffmann*) (Hoffmann et al., 2011), a Multiple-Instance algorithm that supports overlapping relations. It also imposes $y = \ell$. 3) *Mintz++* (Surdeanu et al., 2012), a variant of the single-instance learning algorithm (section 3). The first two are stat-of-the-art Multi-Instance Multi-Label algorithms. *Mintz++* is a strong baseline (Surdeanu et al., 2012) and an improved version of Mintz et al. (2009). Figure 2 shows that our algorithm consistently outperforms all three al-

gorithms at almost all recall levels (with the exception of a very small region in the PR-curve). This demonstrates that by treating unla-beled data set differently and leveraging the missing positive bags, *MIML-semi* is able to learn a more accurate model for extraction. Although the proposed solution is a specific algorithm, we believe the idea of treating unlabeled data differently can be incorporated into any of these algorithms that only use unlabeled data as negative examples.

## 6   Conclusion

We show that the distant-supervision labeling process generates a significant number of false negatives because the knowledge base is incomplete. We proposed an algorithm that learns from only positive and unlabeled bags. Experimental results demonstrate its advantage over existing algorithms.

# References

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*.

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*.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Heng Ji, Ralph Grishman, and Hoa T. Dang. 2011. Overview of the TAC 2011 knowledge base population track. In *Proceedings of the Text Analytics Conference*.

Jing Jiang and ChengXiang Zhai. 2007. A systematic exploration of the feature space for relation extraction. In *Proceedings of HLT-NAACL-2007*.

Nanda Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction. In *Proceedings of ACL-2004*.

Scott Miller, Heidi Fox, Lance Ramshaw, and Ralph Weischedel. 2000. A novel use of statistical parsing to extract information from text. In *Proceedings of NAACL-2000*.

Bonan Min, Shuming Shi, Ralph Grishman and Chin-Yew Lin. 2012a. Ensemble Semantics for Large-scale Unsupervised Relation Extraction. In *Proceedings of EMNLP-CoNLL 2012*.

Bonan Min, Xiang Li, Ralph Grishman and Ang Sun. 2012b. New York University 2012 System for KBP Slot Filling. In *Proceedings of the Text Analysis Conference (TAC)* 2012.

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*.

Truc Vien T. Nguyen and Alessandro Moschitti. 2011. End-to-end relation extraction using distant supervision from external semantic repositories. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 10)*.

Ang Sun, Ralph Grishman, Wei Xu, and Bonan Min. 2011. New York University 2011 system for KBP slot filling. In *Proceedings of the Text Analytics Conference*.

Mihai Surdeanu, Sonal Gupta, John Bauer, David McClosky, Angel X. Chang, Valentin I. Spitkovsky, and Christopher D. Manning. 2011. Stanfords distantly-supervised slot-filling system. In *Proceedings of the Text Analytics Conference*.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, Christopher D. Manning. 2012. Multi-instance Multi-label Learning for Relation Extraction. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing and Natural Language Learning*.

TAC KBP 2011 task definition. 2011. http://nlp.cs.qc.cuny.edu/kbp/2011/KBP2011_TaskDefinition.pdf

Shingo Takamatsu, Issei Sato, Hiroshi Nakagawa. 2012. Reducing Wrong Labels in Distant Supervision for Relation Extraction. In *Proceedings of 50th Annual Meeting of the Association for Computational Linguistics*.

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

Guodong Zhou, Jian Su, Jie Zhang and Min Zhang. 2005. Exploring various knowledge in relation extraction. In *Proceedings of ACL-2005*.