Infusion of Labeled Data into Distant Supervision for Relation Extraction

Maria Pershina [+] Bonan Min [^] [*] Wei Xu [#] Ralph Grishman [+]
[+]New York University, New York, NY
{pershina, grishman}@cs.nyu.edu
[^]Raytheon BBN Technologies, Cambridge, MA
bmin@bbn.com
[#]University of Pennsylvania, Philadelphia, PA
xwe@cis.upenn.edu

Abstract

Distant supervision usually utilizes only unlabeled data and existing knowledge bases to learn relation extraction models. However, in some cases a small amount of human labeled data is available. In this paper, we demonstrate how a state-of-the-art multi-instance multi-label model can be modified to make use of these reliable sentence-level labels in addition to the relation-level distant supervision from a database. Experiments show that our approach achieves a statistically significant increase of 13.5% in F-score and 37% in area under the precision recall curve.

1 Introduction

Relation extraction is the task of tagging semantic relations between pairs of entities from free text. Recently, distant supervision has emerged as an important technique for relation extraction and has attracted increasing attention because of its effective use of readily available databases (Mintz et al., 2009; Bunescu and Mooney, 2007; Snyder and Barzilay, 2007; Wu and Weld, 2007). It automatically labels its own training data by heuristically aligning a knowledge base of facts with an unlabeled corpus. The intuition is that any sentence which mentions a pair of entities (e_1 and e_2) that participate in a relation, r, is likely to express the fact $r(e_1,e_2)$ and thus forms a positive training example of r.

One of most crucial problems in distant supervision is the inherent errors in the automatically generated training data (Roth et al., 2013). Table 1 illustrates this problem with a toy example. Sophisticated multi-instance learning algorithms (Riedel et al., 2010; Hoffmann et al., 2011; Surdeanu et al., 2012) have been proposed to address the issue by loosening the distant supervision assumption. These approaches consider all mentions of the same pair (e_1,e_2) and assume that *at-least-one* mention actually expresses the relation. On top of that, researchers further improved performance by explicitly adding preprocessing steps (Takamatsu et al., 2012; Xu et al., 2013) or additional layers inside the model (Ritter et al., 2013; Min et al., 2013) to reduce the effect of training noise.

True Positive	... to get information out of captured al-Qaida *leader* **Abu Zubaydah**.
False Positive	**...Abu Zubaydah** and former Taliban *leader* Jalaluddin Haqqani ...
False Negative	**...Abu Zubaydah** is one of Osama bin Laden's *senior operational planners*...

Table 1: Classic errors in the training data generated by a toy knowledge base of only one entry personTitle(Abu Zubaydah, *leader*).

However, the potential of these previously proposed approaches is limited by the inevitable gap between the relation-level knowledge and the instance-level extraction task. In this paper, we present the first effective approach, Guided DS (distant supervision), to incorporate labeled data into distant supervision for extracting relations from sentences. In contrast to simply taking the union of the hand-labeled data and the corpus labeled by distant supervision as in the previous work by Zhang et al. (2012), we generalize the labeled data through feature selection and model this additional information directly in the latent variable approaches. Aside from previous semi-supervised work that employs labeled and unlabeled data (Yarowsky, 2013; Blum and Mitchell, 1998; Collins and Singer, 2011; Nigam, 2001, and others), this is a learning scheme that combines unlabeled text and two training sources whose quantity and quality are radically different (Liang et al., 2009).

To demonstrate the effectiveness of our pro-

[*] Most of the work was done when this author was at New York University

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 732–738, Baltimore, Maryland, USA, June 23-25 2014.

Guideline $g = \{g_i \| i = 1, 2, 3\}$: types of entities, $dependency\ path$, span word (optional)	Relation $r(g)$
person_person, $nsubj \rightarrow\leftarrow dobj$, married	personSpouse
person_organization, $nsubj \rightarrow\leftarrow prep\_of$, became	personMemberOf
organization_organization, $nsubj \rightarrow\leftarrow prep\_of$, company	organizationSubsidiaries
person_person, $poss \rightarrow\leftarrow appos$, sister	personSiblings
person_person, $poss \rightarrow\leftarrow appos$, father	personParents
person_title, $\leftarrow nn$	personTitle
organization_person, $prep\_of \rightarrow appos \rightarrow$	organizationTopMembersEmployees
person_cause, $nsubj \rightarrow\leftarrow prep\_of$	personCauseOfDeath
person_number, $\leftarrow appos$	personAge
person_date, $nsubjpass \rightarrow\leftarrow prep\_on \leftarrow num$	personDateOfBirth

Table 2: Some examples from the final set $\mathbf{G}$ of extracted guidelines.

posed approach, we extend MIML (Surdeanu et al., 2012), a state-of-the-art distant supervision model and show a significant improvement of 13.5% in F-score on the relation extraction benchmark TAC-KBP (Ji and Grishman, 2011) dataset. While prior work employed tens of thousands of human labeled examples (Zhang et al., 2012) and only got a 6.5% increase in F-score over a logistic regression baseline, our approach uses much less labeled data (about 1/8) but achieves much higher improvement on performance over stronger baselines.

2 The Challenge

Simply taking the union of the hand-labeled data and the corpus labeled by distant supervision is not effective since hand-labeled data will be swamped by a larger amount of distantly labeled data. An effective approach must recognize that the hand-labeled data is more reliable than the automatically labeled data and so must take precedence in cases of conflict. Conflicts cannot be limited to those cases where all the features in two examples are the same; this would almost never occur, because of the dozens of features used by a typical relation extractor (Zhou et al., 2005). Instead we propose to perform feature selection to generalize human labeled data into *training guidelines*, and integrate them into latent variable model.

2.1 Guidelines

The sparse nature of feature space dilutes the discriminative capability of useful features. Given the small amount of hand-labeled data, it is important to identify a small set of features that are general enough while being capable of predicting quite accurately the type of relation that may hold between two entities.

We experimentally tested alternative feature sets by building supervised Maximum Entropy (MaxEnt) models using the hand-labeled data (Table 3), and selected an effective combination of three features from the full feature set used by Surdeanu et al., (2011):

- the semantic types of the two arguments (e.g. person, organization, location, date, title, ...)
- the sequence of dependency relations along the path connecting the heads of the two arguments in the dependency tree.
- a word in the sentence between the two arguments

These three features are strong indicators of the type of relation between two entities. In some cases the semantic types of the arguments alone narrows the possibilities to one or two relation types. For example, entity types such as person and title often implies the relation personTitle. Some lexical items are clear indicators of particular relations, such as "brother" and "sister" for a sibling relationship

We extract guidelines from hand-labeled data. Each guideline g=$\{g_i|i$=1,2,3$\}$ consists of a pair of semantic types, a dependency path, and optionally a span word and is associated with a particular relation $r(g)$. We keep only those guidelines

Model	Precision	Recall	F-score
MaxEnt$^{\text{all}}$	18.6	6.3	9.4
MaxEnt$^{\text{two}}$	24.13	10.75	14.87
MaxEnt$^{\text{three}}$	40.27	12.40	18.97

Table 3: Performance of a MaxEnt, trained on hand-labeled data using all features (Surdeanu et al., 2011) vs using a subset of two (types of entities, dependency path), or three (adding a span word) features, and evaluated on the test set.

which make the correct prediction for all and at least k=3 examples in the training corpus (threshold 3 was obtained by running experiments on the development dataset). Table 2 shows some examples in the final set $\mathbf{G}$ of extracted guidelines.

3 Guided DS

Our goal is to jointly model human-labeled ground truth and structured data from a knowledge base in distant supervision. To do this, we extend the MIML model (Surdeanu et al., 2012) by adding a new layer as shown in Figure 1.

The input to the model consists of (1) distantly supervised data, represented as a list of n bags[1] with a vector $\mathbf{y_i}$ of binary gold-standard labels, either $Positive(P)$ or $Negative(N)$ for each relation $r \in R$; (2) generalized human-labeled ground truth, represented as a set $\mathbf{G}$ of feature conjunctions g=$\{g_i|i$=1,2,3$\}$ associated with a unique relation $r(g)$. Given a bag of sentences, $\mathbf{x_i}$, which mention an ith entity pair (e_1, e_2), our goal is to correctly predict which relation is mentioned in each sentence, or NR if none of the relations under consideration are mentioned. The vector $\mathbf{z_i}$ contains the latent mention-level classifications for the ith entity pair. We introduce a set of latent variables $\mathbf{h_i}$ which model human ground truth for each mention in the ith bag and take precedence over the current model assignment $\mathbf{z_i}$.

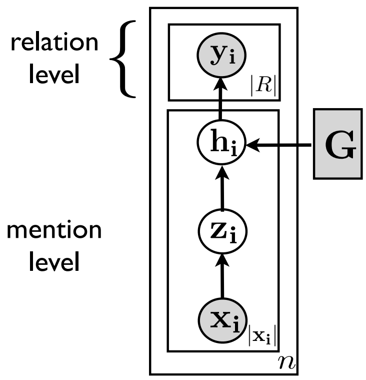


Figure 1: Plate diagram of Guided DS

Let i, j be the index in the bag and the mention level, respectively. We model mention-level extraction $p(z_{ij}|\mathbf{x}_{ij}; \mathbf{w}_z)$, human relabeling $h_{ij}(\mathbf{x}_{ij}, z_{ij})$ and multi-label aggregation $p(y_i^r|\mathbf{h}_i; \mathbf{w_y})$. We define:

- $y_i^r \in \{P, N\}$: r holds for the ith bag or not.
- $\mathbf{x}_{ij}$ is the feature representation of the jth relation mention in the ith bag. We use the same set of features as in Surdeanu et al. (2012).
- $z_{ij} \in R \cup NR$: a latent variable that denotes the relation of the jth mention in the ith bag
- $h_{ij} \in R \cup NR$: a latent variable that denotes the refined relation of the mention $\mathbf{x}_{ij}$

We define relabeled relations h_{ij} as following:

$$h_{ij}(x_{ij}, z_{ij}) = \begin{cases} r(g), \text{ if } \exists! g \in \mathbf{G} \text{ s.t. } g = \{g_k\} \subseteq \{\mathbf{x}_{ij}\} \\ z_{ij} \text{ , otherwise} \end{cases}$$

Thus, relation $r(g)$ is assigned to h_{ij} iff there exists a unique guideline $g \in \mathbf{G}$, such that the feature vector $\mathbf{x}_{ij}$ contains all constituents of g, i.e. entity types, a dependency path and maybe a span word, if g has one. We use mention relation z_{ij} inferred by the model only in case no such a guideline exists or there is more than one matching guideline. We also define:

- $\mathbf{w}_z$ is the weight vector for the multi-class relation mention-level classifier[2]
- $\mathbf{w}_y^r$ is the weight vector for the rth binary top-level aggregation classifier (from mention labels to bag-level prediction). We use $\mathbf{w}_y$ to represent $\mathbf{w}_y^1, \mathbf{w}_y^2, \ldots, \mathbf{w}_y^{|R|}$.

Our approach is aimed at improving the mention-level classifier, while keeping the multi-instance multi-label framework to allow for joint modeling.

4 Training

We use a hard expectation maximization algorithm to train the model. Our objective function is to maximize log-likelihood of the data:

$$LL(\mathbf{w_y}, \mathbf{w_z}) = \sum_{i=1}^{n} \log p(\mathbf{y_i}|\mathbf{x_i}, \mathbf{w_y}, \mathbf{w_z}, \mathbf{G})$$

$$= \sum_{i=1}^{n} \log \sum_{\mathbf{h_i}} p(\mathbf{y_i}, \mathbf{h_i}|\mathbf{x_i}, \mathbf{w_y}, \mathbf{w_z}, \mathbf{G})$$

$$= \sum_{i=1}^{n} \log \sum_{\mathbf{h_i}} \prod_{j=1}^{|\mathbf{h_i}|} p(h_{ij}|\mathbf{x}_{ij}, \mathbf{w_z}, \mathbf{G}) \prod_{r \in P_i \cup N_i} p(y_i^r|\mathbf{h}_i, \mathbf{w}_\mathbf{y}^r)$$

where the last equality is due to conditional independence. Because of the non-convexity of $LL(\mathbf{w_y}, \mathbf{w_z})$ we approximate and maximize the joint log-probability $p(\mathbf{y_i}, \mathbf{h_i}|\mathbf{x_i}, \mathbf{w_y}, \mathbf{w_z}, \mathbf{G})$ for each entity pair in the database:

$$\log p(\mathbf{y_i}, \mathbf{h_i}|\mathbf{x_i}, \mathbf{w_y}, \mathbf{w_z}, \mathbf{G})$$

$$= \sum_{j=1}^{|\mathbf{h_i}|} \log p(h_{ij}|\mathbf{x}_{ij}, \mathbf{w_z}, \mathbf{G}) + \sum_{r \in P_i \cup N_i} \log p(y_i^r|\mathbf{h}_i, \mathbf{w}_\mathbf{y}^r).$$

[1]A bag is a set of mentions sharing same entity pair.

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

Iteration	1	2	3	4	5	6	7	8
(a) Corrected relations:	2052	718	648	596	505	545	557	535
(b) Retrieved relations:	10219	860	676	670	621	599	594	592
Total relabelings	12271	1578	1324	1264	1226	1144	1153	1127

Table 4: Number of relabelings for each training iteration of Guided DS: (a) relabelings due to corrected relations, e.g. personChildren $\rightarrow$ personSiblings (b) relabelings due to retrieved relations, e.g. notRelated(NR) $\rightarrow$ personTitle

Algorithm 1 : **G**uided DS training

1: Phase 1: build set $\mathbf{G}$ of guidelines
2: Phase 2: EM training
3: **for** $\text{iteration} = 1, \dots, T$ **do**
4:     **for** $i = 1, \dots, n$ **do**
5:         **for** $j = 1, \dots, |\mathbf{x_i}|$ **do**
6:             $z^*_{ij} = \text{argmax}_{z_{ij}} p(z_{ij}|\mathbf{x_i},\mathbf{y_i},\mathbf{w_z},\mathbf{w_y})$
7:             $h^*_{ij} = \begin{cases} r(g), \text{ if } \exists!\, g \in \mathbf{G} : \{g_k\} \subseteq \{\mathbf{x}_{ij}\} \\ {z_{ij}}^* \text{ , otherwise} \end{cases}$
8:             **update** $\mathbf{h_i}$ **with** h^*_{ij}
9:         **end for**
10:     **end for**
11:     $\mathbf{w^*_z} = \text{argmax}_{\mathbf{w}} \sum_{i=1}^{n} \sum_{j=1}^{|\mathbf{x_i}|} \log p(h_{ij}|\mathbf{x}_{ij},\mathbf{w})$
12:     **for** $r \in R$ **do**
13:         $\mathbf{w^{r*}_y} = \text{argmax}_{\mathbf{w}} \sum_{1 \le i \le n \ s.t.\ r \in P_i \cup N_i} \log p(y^r_i|\mathbf{h_i},\mathbf{w})$
14:     **end for**
15: **end for**
16: **return** $\mathbf{w_z}, \mathbf{w_y}$

The pseudocode is presented as algorithm 1.

The following approximation is used for inference at step 6:

$$p(z_{ij}|\mathbf{x_i},\mathbf{y_i},\mathbf{w_z},\mathbf{w_y}) \propto p(\mathbf{y_i}, z_{ij}|\mathbf{x_i}, \mathbf{w_y}, \mathbf{w_z}) \approx p(z_{ij}|x_{ij}, \mathbf{w_z}) p(\mathbf{y_i}|\mathbf{h'_i}, \mathbf{w_y}) = p(z_{ij}|x_{ij}, \mathbf{w_z}) \prod_{r \in P_i \cup N_i} p(y^r_i|\mathbf{h'_i}, \mathbf{w^r_y}),$$

where $\mathbf{h'_i}$ contains previously inferred and maybe further relabeled mention labels for group i (steps 5-10), with the exception of component j whose label is replaced by z_{ij}. In the M-step (lines 12-15) we optimize model parameters $\mathbf{w_z}, \mathbf{w_y}$, given the current assignment of mention-level labels $\mathbf{h_i}$.

Experiments show that Guided DS efficiently learns new model, resulting in a drastically decreasing number of needed relabelings for further iterations (Table 4). At the inference step we first classify all mentions:

$$z^*_{ij} = \text{argmax}_{z \in R \cup NR} \ p(z|x_{ij}, \mathbf{w_z})$$

Then final relation labels for ith entity tuple are obtained via the top-level classifiers:

$$y^{r*}_i = \text{argmax}_{y \in \{P,N\}} \ p(y|\mathbf{z^*_i}, \mathbf{w^r_y})$$

5 Experiments

5.1 Data

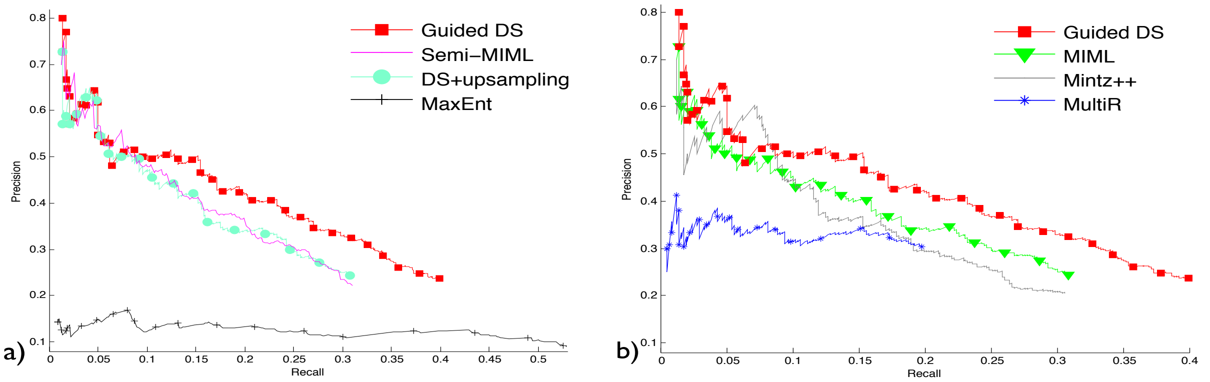
We use the KBP (Ji and Grishman, 2011) dataset[3] which is preprocessed by Surdeanu et al. (2011) using the Stanford parser[4] (Klein and Manning, 2003). This dataset is generated by mapping Wikipedia infoboxes into a large unlabeled corpus that consists of 1.5M documents from KBP source corpus and a complete snapshot of Wikipedia.

The KBP 2010 and 2011 data includes 200 query named entities with the relations they are involved in. We used 40 queries as development set and the rest 160 queries (3334 entity pairs that express a relation) as the test set. The official KBP evaluation is performed by pooling the system responses and manually reviewing each response, producing a hand-checked assessment data. We used KBP 2012 assessment data to generate guidelines since queries from different years do not overlap. It contains about 2500 labeled sentences of 41 relations, which is less than 0.09% of the size of the distantly labeled dataset of 2M sentences. The final set **G** consists of 99 guidelines (section 2.1).

5.2 Models

We implement Guided DS on top of the MIML (Surdeanu et al., 2012) code base[5]. Training MIML on a simple fusion of distantly-labeled and human-labeled datasets does not improve the maximum F-score since this hand-labeled data is swamped by a much larger amount of distant-supervised data of much lower quality. Upsampling the labeled data did not improve the performance either. We experimented with different upsampling ratios and report best results using ratio 1:1 in Figure 2.

[3] Available from Linguistic Data Consortium (LDC) at http://projects.ldc.upenn.edu/kbp/data.
[4] http://nlp.stanford.edu/software/lex-parser.shtml
[5] Available at http://nlp.stanford.edu/software/mimlre.shtml.



o el			1		o el			1	
ax nt	2	12	1	1	MultiR	3	1	2	
+ sam ling	32 2	2 31	2 2	12	Mintz++	2 1	2	2 1	1
emi	3 2	2 21	2	12 31	MIML	2	2	2 3	11
					Guided DS				

Figure 2: Performance of Guided DS on KBP task compared to a) baselines: MaxEnt, DS+upsampling, Semi-MIML (Min et al., 2013) b) state-of-art models: Mintz++ (Mintz et al., 2009), MultiR (Hoffmann et al., 2011), MIML (Surdeanu et al., 2012)

Our baselines: 1) MaxEnt is a supervised maximum entropy baseline trained on a human-labeled data; 2) DS+upsampling is an upsampling experiment, where MIML was trained on a mix of a distantly-labeled and human-labeled data; 3) Semi-MIML is a recent semi-supervised extension. We also compare Guided DS with three state-of-the-art models: 1) MultiR and 2) MIML are two distant supervision models that support multi-instance learning and overlapping relations; 3) Mintz++ is a single-instance learning algorithm for distant supervision. The difference between Guided DS and all other systems is significant with p-value less than 0.05 according to a paired t-test assuming a normal distribution.

5.3 Results

We scored our model against all 41 relations and thus replicated the actual KBP evaluation. Figure 2 shows that our model consistently outperforms all six algorithms at almost all recall levels and improves the maximum F-score by more than 13.5% relative to MIML (from 28.35% to 32.19%) as well as increases the area under precision-recall curve by more than 37% (from 11.74 to 16.1). Also, Guided DS improves the overall recall by more than 9% absolute (from 30.9% to 39.93%) at a comparable level of precision (24.35% for MIML vs 23.64% for Guided DS), while increases the running time of MIML by only 3%. Thus, our approach outperforms state-of-the-art model for relation extraction using much less labeled data that was used by Zhang et al., (2012) to outperform logistic regression baseline. Performance of Guided DS also compares favorably with best scored hand-coded systems for a similar task such as Sun et al., (2011) system for KBP 2011, which reports an F-score of 25.7%.

6 Conclusions and Future Work

We show that relation extractors trained with distant supervision can benefit significantly from a small number of human labeled examples. We propose a strategy to generate and select guidelines so that they are more generalized forms of labeled instances. We show how to incorporate these guidelines into an existing state-of-art model for relation extraction. Our approach significantly improves performance in practice and thus opens up many opportunities for further research in RE where only a very limited amount of labeled training data is available.

Acknowledgments

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory (AFRL) contract number FA8650-10-C-7058. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, or the U.S. Government.

References

Avrim Blum and Tom M. Mitchell. 1998. Combining labeled and unlabeled sata with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT)*, pages 92–100.

Razvan C. Bunescu and Raymond J. 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)*.

Michael Collins and Yorav Singer. 1999. Unsupervised models for named entity classification. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-VLC)*. ,

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 (ISMB)*, pages 77–86.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. 2008. Open information extraction from the web. *Communications of the ACM*, 51(12):68–74.

Raphael Hoffmann, Congle Zhang, and Daniel S. Weld. 2010. Learning 5000 relational extractors. In *Proceedings of the 49th Annual Meetings of the Association for Computational Linguistics (ACL)*, pages 286–295.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke S. Zettlemoyer, and Daniel S. Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 541–550.

Heng Ji and Ralph Grishman. 2011. Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1148–1158.

Heng Ji, Ralph Grishman, and Hoa Trang Dang. 2011. Overview of the TAC-2011 knowledge base population track. In *Text Analysis Conference Workshop*.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41th Annual Meetings of the Association for Computational Linguistics (ACL)*.

Percy Liang, Michael I.Jordan and Dan Klein. 2009. Learning From Measurements in Exponential Families. In Proceedings of the 26th Annual International Conference on Machine Learning (ICML), pages = 641–648

Bonan Min, Ralph Grishman, Li Wan, Chang Wang, and David Gondek. 2013. Distant supervision for relation extraction with an incomplete knowledge base. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedigns of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing (ACL)*, pages 1003–1011.

Ramesh Nallapati. 2004. Discriminative models for information retrieval. In *Proceedigns of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 64–71.

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 (ACL)*, pages 277–282.

Kamal Paul Nigam. 2001. Using Unlabeled Data to Improve Text Classification. Ph.D. thesis, School of Computer Science, Carnegie Mellon University.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedigns of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*, pages 148–163.

Alan Ritter, Luke Zettlemoyer, Mausam, and Oren Etzioni. 2013. Modeling missing data in distant supervision for information extraction. *Transactions of the Association for Computational Linguistics*.

Benjamin Roth, Tassilo Barth, Michael Wiegand, and Dietrich Klakow 2013. A Survey of Noise Reduction Methods for Distant Supervision. In *Proceedings of Conference on Information and Knowledge Management (CIKM-AKBC)*.

Benjamin Snyder and Regina Barzilay 2007. Database-text alignment via structured multilabel classification. In *Proceedings of IJCAI*.

Ang Sun, Ralph Grishman, Wei Xu, and Bonan Min. 2011. New york university 2011 system for kbp slot filling. In *Text Analysis Conference (TAC-KBP)*.

Mihai Surdeanu, J. Turmo, and A. Ageno. 2006. A hybrid approach for the acquisition of information extraction patterns. In *Proceedings of the 11th Conference of the European Chapter of the Associate for Computational Linguistics Workshop on Adaptive Text Extraction and Mining (EACL)*.

Mihai Surdeanu, Sonal Gupta, John Bauer, David McClosky, Angel X. Chang, Valentin I. Spitkovsky, and Christopher D.Manning. 2011. Stanford's

Distantly-Supervised Slot-Filling System. In *Proceedings of the Text Analysis Conference (TAC-KBP).*

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 455–465.

Shingo Takamatsu, Issei Sato, and Hiroshi Nakagawa. 2012. Reducing wrong labels in distant supervision for relation extraction. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 721–729.

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

Wei Xu, Raphael Hoffmann, Zhao Le, and Ralph Grishman. 2013. Filling knowledge base gaps for distant supervision of relation extraction. In *Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics (ACL).*

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33th Annual Meeting of the Association for Computational Linguistics (ACL).*

Ce Zhang, Feng Niu, Christopher Ré, and Jude Shavlik. 2012. Big data versus the crowd: Looking for relationships in all the right places. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 825–834. Association for Computational Linguistics.

Guodong Zhou, Jian Su, Jie Zhang, and Min Zhang. 2005. Exploring various knowledge in relation extraction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL).*