Global Distant Supervision for Relation Extraction

Xianpei Han **Le Sun**

State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences
{xianpei, sunle}@nfs.iscas.ac.cn

Abstract

Machine learning approaches to relation extraction are typically supervised and require expensive labeled data. To break the bottleneck of labeled data, a promising approach is to exploit easily obtained indirect supervision knowledge – which we usually refer to as *distant supervision (DS)*. However, traditional DS methods mostly only exploit one specific kind of indirect supervision knowledge – the relations/facts in a given knowledge base, thus often suffer from the problem of lack of supervision. In this paper, we propose a global distant supervision model for relation extraction, which can: 1) compensate the lack of supervision with a wide variety of indirect supervision knowledge; and 2) reduce the uncertainty in DS by performing joint inference across relation instances. Experimental results show that, by exploiting the consistency between relation labels, the consistency between relations and arguments, and the consistency between neighbor instances using *Markov logic*, our method significantly outperforms traditional DS approaches.

Introduction

Relation extraction (RE) aims to identify and categorize relations between pairs of entities in text. For example, a RE system will extract *CEO-of(Jobs, Apple)* from the sentence "*Jobs is the CEO of Apple*". In recent years, with the expectation to build large scale, machine-readable knowledge bases which can support natural language understanding and human-like reasoning (e.g., *Yago*[1], *DBPedia*[2] and *Freebase*[3]), there is an increasing need for extracting relations/facts from large scale corpus (e.g., the Web). Unfortunately, machine learning approaches to relation extraction are typically supervised and require expensive labeled data, therefore are unlikely to be scaled to the web situation.

To break the bottleneck of labeled data, a promising approach is to exploit easily obtained indirect supervision knowledge – which we usually refer to as *distant supervision (DS)*. For example, as shown in Figure 1, we can collect training instances by heuristically aligning relations in a knowledge base (KB) with the sentences in a given corpus, then these instances can be used to build relation extractors using classifiers such as SVM and logistic classifier.

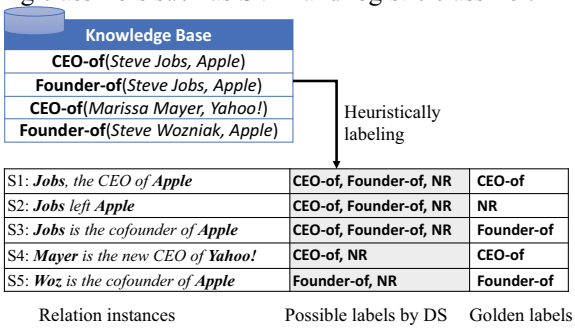

Figure 1.Training instance labeling using KB relations

However, traditional DS methods mostly only exploit one specific kind of indirect supervision knowledge – the relations/facts in a given knowledge base such as *Freebase* and *Yago*. Ignoring many other kinds of indirect supervision knowledge, traditional methods often suffer from the problem of lack of supervision. Specifically, the lack of supervision will introduce a lot of uncertainty and will result in wrongly labeled training instances. For example, in Figure 1 if only using the relations in KB as supervision knowledge, we will not be able to accurately label the three training instances *S1-S3* because they can be labeled as either *CEO-of*, *Founder-of* or *NR (Not a Relation)*, and there will be totally $3\times3\times3\times2\times2=108$ possible states for the labels of the five example instances. Since most machine learning techniques require accurately labelled training instances, the label uncertainty will result in big challenges.

To resolve the above problem, this paper proposes a global distant supervision model for relation extraction, which can: 1) compensate the lack of supervision with a wide variety of indirect supervision knowledge; and 2) reduce the uncertainty in DS by performing joint inference across relation instances. The idea of our method is that we

[1] http://www.mpi-inf.mpg.de/yago-naga/yago
[2] http://www.dbpedia.org/
[3] http://www.freebase.com/

can reduce uncertainty by accumulating evidence from many kinds of weak supervision knowledge and learning models which are globally consistent with all these weak supervision. For instance, apart from the relations in KB, we human can also derive supervision from world knowledge, such as selectional preference (*a CEO usually is a business person*), relation entailment (*the capital of a country is also a city of the country*), relation co-occurrence (*a company's founder usually is its CEO, too*) and label consistency between neighbors (*similar instances tend to express the same kind of relations*). It is clear that if we can model and exploit all these kinds of indirect supervision knowledge together, we will enhance the performance of distant supervision.

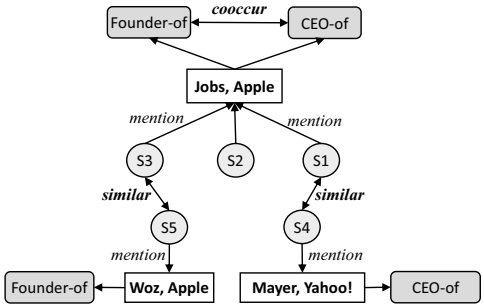

Figure 2. Dependencies between the objects in Figure 1

Our method will further reduce the uncertainty in DS by performing joint inference across relation instances. Through joint inference, evidence can be propagated across dependent decisions, and the "*easy/unambiguous*" decisions can be leveraged to help related "*hard/ambiguous*" ones. For example, in Figure 2, we can label the ambiguous instances *S1* and *S3* correspondingly as *CEO-of* and *Founder-of* using the evidence from the unambiguous instances *S5* and *S4*, where the evidence can be propagated through the *similar* dependencies between (*S3, S5*) and (*S1, S4*).

To materialize the above vision, there are several challenges to overcome. Firstly, our model should be expressive enough to accommodate a wide variety of supervision knowledge. Secondly, our model should compactly encode complex dependencies between different decisions. Thirdly, our model should be able to make globally consistent decisions under a lot of uncertainty and complex dependency structure. In this paper, we employ *Markov logic* (Richardson and Domingos, 2006) as representation language, and propose a globally consistent *Markov logic network* for DS which can address all above three challenges. We test our model on a publicly available data set. Experimental results show that our method significantly outperforms traditional DS methods.

This paper is organized as follows. Section 2 reviews related work and introduces Markov logic. Section 3 describes the global distant supervision model. Section 4 describes the learning and the inference of our model. Section 5 discusses experimental results. Finally Section 6 concludes this paper.

Background

Distant Supervision for Relation Extraction

As described above, a fundamental challenge of distant supervision is the label uncertainty of training instances. A straightforward solution is to turn the distant supervision problem into a direct supervision problem using the heuristic alignment assumption "*any sentence that contains a pair of entities is likely to express their relation in a KB*" (Craven and Kumlien, 1999; Wu et al., 2007; Mintz et al., 2009). Unfortunately, the DS assumption often fails and results in wrongly labeled training instances.

One common solution of label uncertainty is to use multi-instance learning techniques, which can exploit the dependencies between the labels of an entity pair and the labels of its mentions (Bunescu and Mooney, 2007; Riedel et al., 2010; Yao et al., 2010). Hoffmann et al. (2010) and Surdeanu et al. (2012) extended the multi-instance model into multi-instance multi-label model so that an entity pair can have multiple labels. Xu et al. (2013), Min et al. (2013), Ritter et al. (2013) and Zhang et al. (2013) further extended the model to resolve the incompleteness of KB. The main drawback of these methods is that they only exploit one specific kind of indirect supervision knowledge, but ignore many other kinds of useful supervision knowledge.

There were also some DS methods which try to eliminate wrongly labeled instances using additional knowledge. Takamatsu et al. (2012), Roth and Klakow (2013) and Han and Sun (2014) exploited the distributional statistics of instances/ patterns/features. Hoffmann et al. (2010) and Zhang et al. (2010) focused on learning dynamic lexicon. Nguyen and Moschitti (2011) and Pershina et al. (2014) constructed extractors by infusing labeled corpus with heuristically labeled corpus of DS. Riedel et al. (2013) and Fan et al. (2014) exploited the co-occurrence statistics between relation types/instances/features. The main drawback of these methods is that they are specialized to the used supervision knowledge, cannot exploit different kinds of indirect supervision knowledge together.

Markov Logic

In this study, the main representation challenges of our model include how to accommodate a wide variety of supervision knowledge and how to compactly encode complex dependencies. One of the most powerful representation languages for the above challenges is *Markov logic* (Richardson and Domingos, 2006), which is a probabilistic extension of first-order logic. First-order logic can compactly encode very complex dependencies, and has been extensively studied for knowledge representation. Markov logic extends first-order logic by softening formulas with weights, so that it can also handle noise and uncertainty.

Specifically, a *Markov Logic Network (MLN)* is a set of weighted first-order clauses. Together with a set of constants, it defines a *Markov network* with one node per ground predicate and one feature per ground clause. The weight of a feature is the weight of the first-order clause that originated it. The probability of a state x in such a network is given by the log-linear model:

$$P(x) = \frac{1}{Z}\exp(\sum_i w_i f_i(x))$$

where Z is a normalization constant, w_i is the weight of the ith formula, and $f_i = 1$ if the ith clause is true, and $f_i = 0$ otherwise. MLN has been successfully applied to many different tasks, such as unsupervised coreference resolution (Poon and Domingos, 2008), relational pattern clustering (Kok and Domingos, 2008), etc.

The Global Distant Supervision Model for Relation Extraction

This section describes our global distant supervision model which uses *Markov logic* as representation language. We first propose a base MLN which can encode the relations in a KB as supervision knowledge, then we extend the base MLN with more supervision knowledge and more decision dependencies.

We assume that there exist k distinct relation types of interest $R = \{r_1, ..., r_k\}$, therefore there are $k+1$ possible relation type labels $T = R \cup NR$ for each instance. In relation extraction, a dataset contains a set of entity pairs $P = \{p_1, ..., p_{|P|}\}$ with each entity pair $p = (arg1, arg2)$, and a set of relation instances $M = \{m_1, ..., m_{|M|}\}$ with each instance is a triple $m = (arg1, arg2, sentence)$. The sentence of an instance is represented as a feature vector, which is extracted using the method in Surdeanu et al. (2012). We use M_p to denote the set of all instances mentioning entity pair p.

Base MLN

Given a data set, the goal of our model is to predict the relation types of all entity pairs and the labels of all instances, which are represented using the following query predicates:

- `HasRel(p,r)`, which is true iff (i.e., *if and only if*) the entity pair `p` has a relation of type `r`;
- `Label(m,t!)`, which is true iff the label of instance `m` is `t` (`t ∈ T`). The "`t!`" notation signifies that for each instance `m`, this predicate is true only for a unique value of `t`, i.e., an instance cannot have multiple relation labels.

The main local evidence predicates of an instance contain:

- `HasFea(m,f)`, which is true iff instance `m` has feature `f`;
- `Arg1Type(m,c)`, which is true iff the argument 1 of `m` is of entity type `c`;
- `Arg2Type(m,c)`, which is true iff the argument 2 of `m` is of entity type `c`.

These evidence predicates are used to predict instance label through the *feature prediction rule* (where "⇒" is the logical entailment operator):

`HasFea(m,+f) ⇒ Label(m,+t)`

and the *selectional preference rule* (where "∧" is the logical conjunction operator):

`Arg1Type(m,+c) ∧ Label(m,+t)`
`Arg2Type(m,+c) ∧ Label(m,+t)`

The "+" notation signifies that MLN contains an instance of the rule with a separate weight for each value combination of the variables with a plus sign. For example, our model will contain a separate feature prediction rule for each (`feature, relation type`) value combination, such as '`HasFea(m,`*`ceo`*`)⇒Label(m,CEO-of)`' and '`HasFea(m,`*`co-found`*`)⇒Label(m,CEO-of)`'. We also model relation label priors using the unit clause:

`Label(+m,+t)`

Based on the above formulas, our model can predict the label of an instance by modeling the dependencies between an instance's relation label and its features and arguments. Once we know the label of an instance, the relation between its entity pair can be extracted using the hard *extraction rule*:

`Label(m,r) ∧ Mention(m,p) ⇒ HasRel(p,r)`

where the evidence predicate `Mention(m,p)` is true iff instance `m` mentions entity pair `p`. Notice that, the extraction rule doesn't indicate each true ground `HasRel(p,r)` predicate will have a supporting true ground `Label(m,r)` predicate in M_p. Therefore once we know the relation between a pair of entities, we use an additional *alignment rule* to predict the labels of its mention instances:

`HasRel(p,+r) ∧ AtLeastOne(`M_p`,+r)`

where the predicate `AtLeastOne(`M_p`,r)` will be true iff at least one instance in M_p has label `r`. The alignment rule is mainly used to encode the supervision from the relations in a KB: in the training of DS models, all values of the ground `HasRel` predicates are given by a KB, and which will be exploited to label the ground `Label` predicates through the above alignment rule.

Consistency between Relation Labels

The base MLN model only exploits the supervision from the relations in a KB. In many cases, there are rich inter-dependencies between the different relation labels of an entity pair. Several main label dependencies include relation entailment (e.g., *Capital-of* ⇒ *City-of*), relation mutual exclusion (e.g., *Parent-of* cannot co-occur with *Spouse-of*), and relation co-occurrence (e.g., *CEO-of* and *Founder-of*). The above label dependencies can provide additional supervision and reduce the uncertainty in DS by forcing consistency between relation labels. For example, if the entity pair (*Paris, France*) has a relation *Capital-of*, then it must also has a relation *City-of* based on the entailment rule "*Capital-of* ⇒ *City-of*".

Specifically, we model the consistency between relation labels using the *relation co-occurrence rule*:

HasRel(p,+r) ∧ HasRel(p,+r')

and the hard *relation entailment rule*:

HasRel(p,r) ∧ Entail(r,r') ⇒ HasRel(p,r')

where the evidence predicate Entail(r,r') is true iff relation type r entails r'. The relation co-occurrence rule can capture both co-occurrence consistency (if the rule weight for a (*relation type, relation type*) pair is positive) and mutual exclusion consistency (if the rule weight for a (*relation type, relation type*) pair is negative).

Consistency between Relation and Argument

Traditionally, the argument information of an instance is represented using its words and its entity type. However, the above representation can only capture limited argument information. For example, for the argument '*MIT*', this representation only captures the information that '*MIT is an organization*', but ignores many other useful information such as '*MIT is a school*', '*MIT is a university*', etc. In our model, we reduce uncertainty in relation extraction by exploiting more argument information.

Specifically, argument information can be exploited to reduce uncertainty in the following two ways:

1) The rich argument information can better model selectional preference. This is because traditional entity types (such as *Person, Location*, *Organization* and *Date*) are usually too coarse to precisely capture the selectional preference of a relation type. For example, argument 2 of *schools_attended* must be an *education institute*, but most current RE systems don't contain this entity type.

2) The argument information can be used to filter out invalid values. For example, argument 2 of the relation type *person:age* is of entity type *NUMBER*, but not all numbers are valid person age values because nearly all persons' ages are between 0 and 120. For demonstration, we investigated the *NUMBER* entities in the KBP data set (Surdeanu et al., 2012), and found that nearly 99% of *NUMBER* entities are invalid person age values.

Based on the above observation, our model encodes the consistency between a relation and its arguments using the *relation-argument consistency rules*:

Arg1HasFea(m,+f) ∧ Label(m,+t)

Arg2HasFea(m,+f) ∧ Label(m,+t)

And we extract argument features as follows:

1) Firstly, we identify the fine-grained Freebase entity types of an argument by matching its name with Freebase, then these entity types are used as its features;

2) Secondly, we extract an age validation feature for *NUMBER* entity – *IsValidPerAge*. We extract this feature if a *NUMBER* entity is an integer and its numeric value is within [0, 120].

Consistency between Neighbor Instances

In this section, we model the dependency between neighbor instances as indirect supervision knowledge. The start point of our method is *the nearest neighbor consistency assumption*, which means that nearby instances are likely to have the same label. Based on this assumption, the similarity between instances can reduce the uncertainty in DS: the classification results should be sufficiently smooth with respect to the underlying similarity structure, and the label of an instance should be consistent with its neighbors.

Specifically, we model the consistency between neighbor instances using *the nearest neighbor consistency rule*:

KNNOf(m,m') ∧ Label(m,t) ⇒ Label(m',t)

where the evidence predicate KNNOf(m,m') is true iff m is one of the *k*-nearest neighbors of m'.

To find *k*-nearest neighbors of an instance, we use a similarity measure as follows:

$$\mathrm{sim}(m, m') = \mathrm{sim}_{\mathrm{arg1}}(m, m') \times \mathrm{sim}_{\mathrm{arg2}}(m, m') \times \mathrm{sim}_{\mathrm{sent}}(m, m')$$

where $\mathrm{sim}_{\mathrm{sent}}$ is the cosine similarity between two instances' feature vectors, $\mathrm{sim}_{\mathrm{arg1}}$ and $\mathrm{sim}_{\mathrm{arg2}}$ are cosine similarities between arg1s' feature vectors and between arg2s' feature vectors. We use the features discussed in above section to represent an argument, where each feature is weighted using its maximum mutual information value over different relation types (Aggarwal & Zhai, 2012).

Learning and Inference

In this section, we describe the learning and the inference of our global distant supervision MLN model, and the extraction of new relations using the learned model.

Learning. In the learning of DS systems, the relations between entity pairs are given, but the labels of instances are unknown. Therefore, distantly supervised learning in Markov logic maximizes the conditional log-likelihood

$$\begin{aligned} L(x, y) &= \log P(Y = y \mid X = x) \\ &= \log \sum_z P(Y = y, Z = z \mid X = x) \end{aligned}$$

where X, Y, Z correspondingly are evidence predicates, known/observed query predicates, unknown/hidden query predicates in training data. In our global distant supervision MLN, Y includes HasRel, Z includes Label, AtLeastOne and X includes HasFea, Arg1Type, Arg2Type, Mention, Arg1HasFea, Arg2HasFea, KNNOf and Entail. The gradient of the above optimization problem is:

$$\frac{\partial}{\partial w_i} L(x, y) = E_{Z|y,x}[n_i] - E_{Y,Z|x}[n_i]$$

where n_i is the number of true groundings of the ith clause. The gradient is the difference of two expectations, each of which can be approximated using samples generated by algorithms such as SampleSAT(Wei et al., 2004) and MC-SAT(Poon & Domingos, 2006). Using the above gradient, many gradient based learning algorithms can be used to learn our MLN model, such as LBFGS and Conjugate Gradients. This study employs PSCG algorithm (Lowd & Domingos, 2007) with two adjustments: 1) we use a fixed step

length; 2) we use the inverse of the number of true groundings as our preconditioner, rather than the inverse diagonal Hessian. We found that these two adjustments can improve the quality of learned models.

Due to the hidden variables Z, the optimization problem of distantly supervised learning is not convex, so parameter initialization is important. In this study, we first learn the parameters of the base MLN model, then the learned weights of the base MLN model are used to initialize the learning of the full MLN model.

Inference. In MLN, all decisions are jointly inferred. Many inference algorithms have been developed for MLN, including MaxWalkSAT (Kautz et al., 1997), SampleSAT (Wei et al., 2004), MC-SAT (Poon & Domingos, 2006), etc. In this study, we use SampleSAT to generate samples for weight learning and use MaxWalkSAT for inference on new data. We extended SampleSAT and MaxWalkSAT with the ability to flip multi-atoms at each step, so that the hard rules in our model and the mutual exclusion of instance labels will always be satisfied.

For relation extraction on a new data set, we first initialize the MLN state by running a MaxWalkSAT pass with only the formulas in base MLN, then run another pass with all formulas. We found this can improve the quality of the optimal MLN state. Given the optimal MLN state, each true ground `HasRel` predicate will be extracted as a relation.

Extraction Ranking. As shown in previous methods (Hoffmann et al., 2011; Surdeanu et al., 2012), many factors can be used to estimate the confidence of extractions for a better precision/recall trade-off, such as the marginal probability of `HasRel(p,r)` and the redundancy of an extraction in a large corpus. Following these observations, we sort extractions using the confidence score:

$$\text{Conf}(p, r) = \max_{m \in M_p} P(z_m = r) \times \text{Reliability}(p, r)$$

where p is an entity pair, r is the extracted relation type between p, z_m is the label of instance m, $P(z_m = r) = P(\texttt{Label(m, r)} = \text{true} \mid Y, Z_{-m}; \boldsymbol{w})$ is the conditional probability of instance `m` having label `r` in MLN, given all the values of other predicates and the MLN parameters $\boldsymbol{w}$. Reliability(p, r) measures whether this extraction is reliable, following the idea of the internal consistency reliability in statistics and research (Trochim, 2000), that is, whether this relation can be extracted in different instances. Specifically, we compute Reliability(p, r) as the *Average Inter-item Correlation* between instance labels:

$$\text{Reliablity}(p, r) = \operatorname{ave}_{m,m' \in M_p^+} \text{Correlation}(z_m, z_{m'})$$

where M_p^+ include all mentions of `p` whose labels are not *NR*; $\text{Correlation}(z_m, z_{m'})$ is the correlation score between the two labels, its value will be 1.0 if $z_m = z_{m'}$, and if $z_m \neq z_{m'}$, its value will be the relation type co-occurrence probability:

$$P_{cooccur}(z_m, z_{m'}) = 1/1+\exp(-w)$$

where w is the weight of the relation co-occurrence rule '`HasRel(p,`z_m`)`$\wedge$`HasRel(p,`$z_{m'}$`)`'.

Experiments

In this section, we assess the performance of our method and compare it with traditional methods.

Data Set

We evaluate our method on a publicly available data set — KBP, which was developed by Surdeanu et al. (2012). KBP was constructed by aligning the relations from English Wikipedia infoboxes against a document collection which contains the corpus provided by the KBP shared task (Ji et al., 2010; Ji et al., 2011) and a complete snapshot of the June 2010 version of Wikipedia. KBP contains 183,062 training relations and 3,334 testing relations. This paper tunes and tests different methods use the same partitions and the same evaluation method as Surdeanu et al. (2012).

System and Baselines

We tune our global distant supervision model using the validation partition of KBP. After tuning for different MLN models, we used PSCG algorithm (5 samples, 10~20 iterations, step length 0.03) and SampleSAT inference algorithm (5,000,000 ~ 10,000,000 flips with 20% noise flips, 30% random ascent flips, and 50% SA flips) for learning. Because positive/negative instances are highly imbalanced in the training corpus, we put a higher misclassification cost (the tuned value is 2.0) to positive instances. For the `KNNOf` evidence predicates, we use 10 nearest neighbors for each instance (with similarity > 0.2).

We compare our method with three baselines:

Mintz++ – This is a traditional DS method proposed by Mintz et al.(2009), which labels training instances using heuristic DS assumption, and employs a multi-class logistic classifier for extraction.

Hoffmann – This is a multi-instance multi-label DS model proposed by Hoffmann et al. (2011), where the label dependency between an entity pair and its mentions is modeled using a deterministic *at-least-one* assumption.

Surdeanu – This is a multi-instance multi-label DS model proposed by Surdeanu et al. (2012), where the label dependency between an entity pair and its mentions is modeled using a relational classifier.

In our experiments, we use the implementations and the optimal settings of Stanford's MIMLRE package (Surdeanu et al., 2012) for all three baselines, which is open source and was shown to achieve state-of-the-art performance.

Overall Results

Following previous methods, we evaluate the different methods using the standard *Precision*, *Recall* and *F1-meas-*

ure on the ranked relation extractions, and provide the precision/recall curves of different methods. For our model, we use two different settings: the first is the base MLN model – *MLN-Base*; the second is the full MLN model – *MLN-Full*. The overall results are shown in Figure 3 and Table 1.

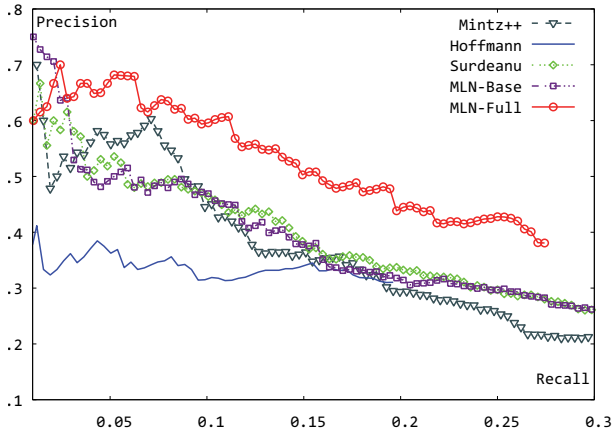

Figure 3. Precision/Recall curves on KBP dataset

System	P	R	F1
Mintz++	0.260	0.250	0.255
Hoffmann	0.306	0.198	0.241
Surdeanu	0.249	0.314	0.278
MLN-Base	**0.262**	**0.302**	**0.281**
MLN-Full	**0.426**	**0.259**	**0.322**

Table 1. The best F1-measures in P/R curves

From the above results, we can see that:

1) By accommodating a wide variety of supervision knowledge and making joint inference, our method significantly outperforms traditional DS methods: compared with the three baselines *Mintz++*, *Hoffmann* and *Surdeanu*, our *MLN-Full* model correspondingly achieved 26%, 34% and 16% F1 improvements.

2) Markov logic can effectively represent knowledge and encode complex dependencies between different decisions for relation extraction. Using the same supervision knowledge (i.e., the relations in a KB), *MLN-base* correspondingly achieved 17% and 1% F1 improvements over the multi-instance baselines: *Hoffmann* and *Surdeanu*.

3) The uncertainty in distant supervision can be greatly reduced by accommodating a wide variety of indirect supervision knowledge and performing joint inference across relation instances. In Table 1 we can see that, compared with the *MLN-Base* model, the *MLN-Full* model achieved 15% F1 improvement. In Figure 3 we can see that the *MLN-Full* model achieved a consistent precision improvement over the *MLN-Base* model on nearly all recall region.

Detailed Analysis

To analyze the effect of different kinds of indirect supervision knowledge, we incrementally accommodated consistency between relation labels (*LabelDep*), consistency between relation and argument (*ArgFea*) and consistency between neighbor instances (*NeighborSim*) into the base MLN model. The results are presented in Table 2.

Model	P	R	F1	ΔBase
MLN-Base	0.262	0.302	0.281	---
+ LabelDep	0.323	0.292	0.307	**9.3%**
+ ArgFea	0.390	0.260	0.312	**11.0%**
+ NeighborSim	0.426	0.259	0.322	**14.6%**

Table 2. The best F1-measures in P/R curves by incremental accommodating indirect supervision knowledge (where ΔBase is the F1 improvement over the MLN-Base model)

From Table 2, we can see that:

1) The accommodation of additional supervision knowledge is an effective way to improve the performance of DS systems. In our model, all three types of indirect supervision knowledge improved the performance of our MLN model.

2) The exploitation of relation label dependencies significantly improved the extraction performance. For instance, our model correspondingly achieved 9.3%, 27.4% and 10.4% F1 improvements over *MLN-Base*, *Hoffmann* and *Surdeanu*. We believe this is because Markov logic provides a flexible way to encode complex label dependencies, such as entailment, mutual exclusion and co-occurrence. Therefore our method can reduce uncertainty more than traditional multi-instance model based DS approaches, which model the relation labels of an entity pair independently.

3) The incorporation of rich argument features improved the relation extraction performance. We believe this is because rich argument features can provide additional useful information. For example, the information '*MIT is a University*' is useful for extracting relations of type *schools_attended*.

4) The consistency between neighbor instances improved the extraction performance. This verified both the effectiveness of using similarity between instances as additional supervision and the effectiveness of using joint inference to reduce uncertainty.

Conclusion

In this paper, we propose a global distant supervision model which can: 1) compensate the lack of direct supervision with a wide variety of indirect supervision knowledge; and 2) overcome the uncertainty in DS by performing joint inference across relation instances. Experimental results showed that our method can significantly improve the relation extraction performance on a publicly available data set.

Distant supervision is a challenging task, future directions include incorporating additional knowledge (e.g., intrinsic structure of data sets such as clusters), better modeling of dependency, and other techniques which can accommodate heterogeneous supervision for distant supervision.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grants no. 61433015, 61572477 and 61272324, and the National High Technology Development 863 Program of China under Grants no. 2015AA015405. Moreover, we sincerely thank the reviewers for their valuable comments.

References

Aggarwal, C. C., & Zhai, C. 2012. *A survey of text classification algorithms*. In: Mining text data, pp. 163-222. Springer US.

Bunescu, R. C. and Mooney, R. 2007. *Learning to extract relations from the web using minimal supervision*. In: Proceedings of ACL 2007, pp. 576-583.

Craven, M. and J. Kumlien. 1999. *Constructing biological knowledge bases by extracting information from text sources*. In: Proceedings of AAAI 1999, pp. 77-86.

Fan, M., et al. 2014. *Distant Supervision for Relation Extraction with Matrix Completion*. In: Proceedings of ACL 2014, pp. 839-849.

Gupta, R. and Sarawagi, S. 2011. *Joint training for open-domain extraction on the web: exploiting overlap when supervision is limited*. In: Proceedings of WSDM 2011, pp. 217-226.

Han, X. and Sun, L. 2014. *Semantic Consistency: A Local Subspace Based Method for Distant Supervised Relation Extraction*. In: Proceedings of ACL 2014, pp. 718-724.

Hoffmann, R., Zhang, C., et al. 2010. *Learning 5000 relational extractors*. In: Proceedings of ACL 2010, pp. 286-295.

Hoffmann, R., Zhang, C., et al. 2011. *Knowledge-based weak supervision for information extraction of overlapping relations*. In: Proceedings of ACL 2011, pp. 541-550.

Ji, H., Grishman, R., et al. 2010. *Overview of the TAC 2010 knowledge base population track.* In: Proceedings of the Text Analytics Conference 2010.

Ji, H., Grishman, R., et al. 2011. *Overview of the TAC 2011 knowledge base population track.* In: Proceedings of the Text Analytics Conference 2011.

Kautz, H.; Selman, B.; and Jiang, Y. 1997. *A general stochastic approach to solving problems with hard and soft constraints*. In: The Satisfiability Problem: Theory and Applications. AMS, pp.573-586.

Krause, S., Li, H., et al. 2012. *Large-Scale learning of relation-extraction rules with distant supervision from the web*. In: Proceedings of ISWC 2012, pp. 263-278.

Kok, S. and Domingos, P. 2008. *Extracting semantic networks from text via relational clustering*. In: Machine Learning and Knowledge Discovery in Databases, pp. 624-639.

Lowd, D. & Domingos, D. 2007. *Efficient weight learning for Markov logic networks*. In: Proceedings of PKDD-07, pp.200-211.

Mintz, M., Bills, S., et al. 2009. *Distant supervision for relation extraction without labeled data*. In: Proceedings of ACL-AFNLP 2009, pp. 1003-1011.

Min, B., Grishman, R., et al. 2013. *Distant Supervision for Relation Extraction with an Incomplete Knowledge Base*. In: Proceedings of NAACL-HLT 2013, pp. 777-782.

Nguyen, T. T. and Moschitti, A. 2011. *Joint distant and direct supervision for relation extraction*. In: Proceedings of IJCNLP 2011, pp. 732-740.

Pershina, M., et al. 2014. *Infusion of Labeled Data into Distant Supervision for Relation Extraction*. In: Proceedings of ACL 2014, pp. 732-738.

Poon, H., & Domingos, P. 2006. *Sound and efficient inference with probabilistic and deterministic dependencies*. In: Proceedings of AAAI 2006, pp. 458-463.

Poon, H., & Domingos, P. 2008. *Joint unsupervised coreference resolution with Markov logic*. In: Proceedings of EMNLP 2008, pp. 650-659.

Richardson, M. & Domingos, P. 2006. *Markov logic networks*. In: Machine Learning 62:107–136.

Riedel, S., Yao, L., et al. 2010. *Modeling relations and their mentions without labeled text*. In: Proceedings of Machine Learning and Knowledge Discovery in Databases, 2010, pp. 148-163.

Riedel, S., Yao, L., et al. 2013. *Relation Extraction with Matrix Factorization and Universal Schemas*. In: Proceedings of NAACL-HLT 2013, pp. 74-84.

Ritter, A., Zettlemoyer, L., Mausam, Etzioni, O. 2013. *Modeling Missing Data in Distant Supervision for Information Extraction*. In: Transactions of the Association for Computational Linguistics, Vol 1, pp. 367-378.

Roth, B. and Klakow, D. 2013. *Combining Generative and Discriminative Model Scores for Distant Supervision*. In: Proceedings of ACL 2013, pp. 24-29.

Surdeanu, M., Tibshirani, J., et al. 2012. *Multi-instance multi-label learning for relation extraction*. In: Proceedings of EMNLP-CoNLL 2012, pp. 455-465.

Takamatsu, S., Sato, I., et al. 2012. *Reducing wrong labels in distant supervision for relation extraction*. In: Proceedings of ACL 2012, pp. 721-729.

Trochim, W. 2000. *The Research Methods Knowledge Base, 2nd Edition*. Atomic Dog Publishing, Cincinnati, OH.

Wang, C., Fan, J., et al. 2011. *Relation extraction with relation topics*. In: Proceedings of EMNLP 2011, pp. 1426-1436.

Wei, W.; Erenrich, J. and Selman, B. 2004. *Towards efficient sampling: Exploiting random walk strategies*. In: Proceedings of AAAI 2004, pp.670-676.

Wu, F. and Weld, D. S. 2007. Autonomously semantifying Wikipedia. In: Proceedings of CIKM 2007, pp. 41-50.

Xu, W., Hoffmann, R., et al. 2013. *Filling Knowledge Base Gaps for Distant Supervision of Relation Extraction*. In: Proceedings of ACL 2013, pp. 665-670.

Yao, L., Riedel, S., et al. 2010. *Collective cross-document relation extraction without labelled data*. In: Proceedings of EMNLP 2010, pp. 1013-1023.

Zhang, C., Hoffmann, R., et al. 2012. *Ontological smoothing for relation extraction with minimal supervision*. In: Proceedings of AAAI 2012, pp. 157-163.

Zhang, X., Zhang, J., et al. 2013. *Towards Accurate Distant Supervision for Relational Facts Extraction*. In: Proceedings of ACL 2013, pp. 810-815.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., & Schölkopf, B. 2004. *Learning with local and global consistency*. In: Proceedings of NIPS 2004, pp. 321-328.