Noisy Or-based model for Relation Extraction using Distant Supervision

Ajay Nagesh[1,2,3]
[1]IITB-Monash Research Academy
ajaynagesh@cse.iitb.ac.in

Gholamreza Haffari
[2]Faculty of IT, Monash University
gholamreza.haffari@monash.edu

Ganesh Ramakrishnan
[3]Dept. of CSE, IIT Bombay
ganesh@cse.iitb.ac.in

Abstract

Distant supervision, a paradigm of relation extraction where training data is created by aligning facts in a database with a large unannotated corpus, is an attractive approach for training relation extractors. Various models are proposed in recent literature to align the facts in the database to their mentions in the corpus. In this paper, we discuss and critically analyse a popular alignment strategy called the *"at least one"* heuristic. We provide a simple, yet effective relaxation to this strategy. We formulate the inference procedures in training as integer linear programming (*ILP*) problems and implement the relaxation to the *"at least one "* heuristic *via* a soft constraint in this formulation. Empirically, we demonstrate that this simple strategy leads to a better performance under certain settings over the existing approaches.

1 Introduction

Although supervised approaches to relation extraction (GuoDong et al., 2005; Surdeanu and Ciaramita, 2007) achieve very high accuracies, they do not scale as they are data intensive and the cost of creating annotated data is quite high. To alleviate this problem, Mintz et al. (2009) proposed relation extraction in the paradigm of *distant supervision*. In this approach, given a database of facts (e.g. Freebase[1]) and an unannotated document collection, the goal is to heuristically align the facts in the database to the sentences in the corpus which contain the entities mentioned in the fact. This is done to create weakly labeled training data to train a classifier for relation extraction. The underlying assumption is that all mentions of an entity pair[2] (i.e. sentences containing the entity pair) in the corpus express the same relation as stated in the database.

The above assumption is a weak one and is often violated in natural language text. For instance, the entity pair, (`Barack Obama`, `United States`) participate in more than one relation: `citizenOf`, `presidentOf`, `bornIn` and every mention expresses either one of these fixed set of relations or none of them.

Consequently, a number of models have been proposed in literature to provide better heuristics for the mapping between the entity pair in the database and its mentions in the sentences of the corpus. Riedel et al. (2010) tightens the assumption of distant supervision in the following manner: "Given a pair of entities and their mentions in sentences from a corpus, *at least one* of the mentions express the relation given in the database". In other words, it models the problem as that of multi-instance (mentions) single-label (relation) learning. Following this, Hoffmann et al. (2011) and Surdeanu et al. (2012) propose models that consider the mapping as that of multi-instance multi-label learning. The instances are the mentions of the entity pair in the sentences of the corpus and the entity pair can participate in more than one relation.

Although, these models work very well in practice, they have a number of shortcomings. One of them is the possibility that during the alignment, a fact in the database might not have an instantiation in the corpus. For instance, if our corpus only contains documents from the years 2000 to 2005, the fact `presidentOf`(`Barack Obama`, `United States`) will not be present in the corpus. In such cases, the distant supervision assumption fails to provide a mapping for the fact in the corpus.

In this paper, we address this situation with a

[1] `www.freebase.com`

[2] In this paper we restrict ourselves to binary relations

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1937–1941,
October 25-29, 2014, Doha, Qatar.

noisy-or model (Srinivas, 2013) in training the relation extractor by relaxing the *"at least one"* assumption discussed above. Our contributions in this paper are as follows: (i) We formulate the inference procedures in the training algorithm as integer linear programming (*ILP*) problems, (ii) We introduce a soft-constraint in the ILP objective to model noisy-or in training, and (iii) Empirically, our algorithm performs better than Hoffmann et al. (2011) procedure under certain settings on two benchmark datasets.

Our paper is organized as follows. In Section 2, we discuss our methodology. We review the approach of Hoffmann et al. (2011) and explain our modifications to it. In Section 3, we discuss related work. In Section 4, we discuss the experimental setup and our preliminary results. We conclude in Section 4.

2 Methodology

Our work extends the work of Hoffmann et al. (2011). So, we recapitulate Hoffmann's model in the following subsection. Following which our additions to this model is explained in detail.

Hoffmann's model

Hoffmann et al. (2011) present a multi-instance multi-label model for relation extraction through distant supervision. In this model, a pair of entities have multiple mentions (sentence containing the entity pair) in the corpus. An entity pair can have one or more relation labels (obtained from the database).

Objective function

Consider an entity pair (e_1, e_2) denoted by the index i. The set of sentences containing the entity pair is denoted $\mathbf{x_i}$ and the set of relation labels for the entity pair from the database is denoted by $\mathbf{y_i}$. The mention-level labels are denoted by the latent variable $\mathbf{z}$ (there is one variable z_j for each sentence j).

To learn the parameters θ, the training objective to maximize is the likelihood of the facts observed in the database conditioned on the sentences in the text corpus.

$$\begin{aligned}\theta^* &= \arg\max_{\theta} \prod_i Pr(\mathbf{y_i}|\mathbf{x_i};\theta) \\ &= \arg\max_{\theta} \prod_i \sum_z Pr(\mathbf{y_i}, \mathbf{z}|\mathbf{x_i};\theta)\end{aligned}$$

The expression $Pr(\mathbf{y_i}, \mathbf{z}|\mathbf{x_i})$ for a given entity pair is defined by two types of factors in the factor graph. They are *extract factors* for each mention and *mention factors* between a relation label and all the mentions.

The *extract factors* capture the local signal for each mention and consists of a bunch of lexical and syntactic features like POS tags, dependency path between the entities and so on (Mintz et al., 2009).

The *mention factors* capture the dependency between relation label and its mentions. Here, the *at least one* assumption that was discussed in Section 1 is modeled. It is implemented as a simple deterministic OR operator as given below:

$$f_{mention}(y_r, \mathbf{z}) = \begin{cases} 1 & \text{if } y_r \text{ is true } \wedge \exists i : z_i = r \\ 0 & \text{otherwise} \end{cases}$$

Training algorithm

The learning algorithm is a perceptron-style parameter update scheme with 2 modifications: i) online learning ii) Viterbi approximation. The inference is shown to reduce to the well-known weighted edge-cover problem which can be solved exactly, although Hoffmann et al. (2011) provide an approximate solution.

Algorithm 1: Hoffmann et al. (2011) : Training

```
Input  : i) $\Sigma$: set of sentences, ii) $E$: set of entities
         mentioned in the sentences, iii) $R$: set of
         relation labels, iv) $\Delta$: database of facts
Output: Extraction model : $\Theta$
begin
    for $t \leftarrow 1$ to $T$ ;          /* training iterations */
    do
        for $i \leftarrow 1$ to $N$ ;   /* No. of entity pairs */
        do
            $\widehat{\mathbf{y}}, \widehat{\mathbf{z}}, = \arg\max_{\mathbf{y},\mathbf{z}} Pr(\mathbf{y}, \mathbf{z}|\mathbf{x_i}; \Theta)$
            if $\widehat{\mathbf{y}}! = \mathbf{y_i}$ then
                $\mathbf{z}^* = \arg\max_{\mathbf{z}} Pr(\mathbf{z}|\mathbf{y_i}, \mathbf{x_i}; \Theta)$
                $\Theta^{new} = \Theta^{old} + \Phi(\mathbf{x_i}, \mathbf{z}^*) - \Phi(\mathbf{x_i}, \widehat{\mathbf{z}})$
end
```

Our additions to Hoffmann's model

In the training algorithm described above, there are two MAP inference procedures. Our contributions in this space is two-fold. Firstly, we have formulated these as ILP problems. As a result of this, the approximate inference therein is replaced by an exact inference procedure. Secondly, we replace the *deterministic-or* by a *noisy-or* which provides a soft-constraint instead of the hard-constraint of Hoffmann. (*"at least one"* assumption)

ILP formulations

Some notations:

- □ z_{ji} : The mention variable z_j (or jth sentence) taking the relation value i
- □ s_{ji} : Score for z_j taking the value of i. Scores are computed from the *extract* factors
- □ y_i : relation label being i
- □ m : number of mentions (sentences) for the given entity pair
- □ R: total number of relation labels (excluding the *nil* label)

Deterministic OR

The following is the ILP formulation for the exact inference $\arg\max Pr(\mathbf{y}, \mathbf{z}|\mathbf{x_i})$ in the model based on the *deterministic-or*:

$$\begin{aligned} & \max_{Z,Y} \left\{ \sum_{j=1}^{m} \sum_{i \in \{R,nil\}} \left[z_{ji} s_{ji} \right] \right\} \\ \textbf{s.t} \quad & 1. \quad \sum_{i \in \{R,nil\}} z_{ji} = 1 \quad \forall j \\ & 2. \quad z_{ji} \leq y_i \quad \forall j, \forall i \\ & 3. \quad y_i \leq \sum_{j=1}^{m} z_{ji} \quad \forall i \\ \text{where} \quad & z_{ji} \in \{0,1\}, \quad y_i \in \{0,1\} \end{aligned}$$

The first constraint restricts a mention to have only one label. The second and third constraints impose the *at least one* assumption. This is the same formulation as Hoffmann but expressed as an ILP problem. However, posing the inference as an ILP allows us to easily add more constraints to it.

Noisy OR

As a case-study, we add the *noisy-or* soft-constraint in the above objective function. The idea is to model the situation where a fact is present in the database but it is not instantiated in the text. This is a common scenario, as the facts populated in the database and the text of the corpus can come from different domains and there might not be a very good match.

$$\begin{aligned} & \max_{Z,Y,\epsilon} \left\{ \left(\sum_{j=1}^{m} \sum_{i \in \{R,nil\}} \left[z_{ji} s_{ji} \right] \right) - \left(\sum_{i \in R} \epsilon_i \right) \right\} \\ \textbf{s.t} \quad & 1. \quad \sum_{i \in \{R,nil\}} z_{ji} = 1 \quad \forall j \\ & 2. \quad z_{ji} \leq y_i \quad \forall j, \forall i \\ & 3. \quad y_i \leq \sum_{j=1}^{m} z_{ji} + \epsilon_i \quad \forall i \\ \text{where} \quad & z_{ji} \in \{0,1\}, \quad y_i \in \{0,1\}, \quad \epsilon_i \in \{0,1\} \end{aligned}$$

In the above formulation, the objective function is augmented with a soft penalty. Also the third constraint is modified with this penalty term. We call this new term ϵ_i and it is a binary variable to model noise. Through this term we encourage *at least one* type of configuration but will not disallow a configuration that does not conform to this. Essentially, the consequence of this is to allow the case where a fact is present in the database but is not instantiated in the text.

3 Related Work

Relation Extraction in the paradigm of distant supervision was introduced by Craven and Kumlien (1999). They used a biological database as the source of distant supervision to discover relations between biological entities. The progression of models for information extraction using distant supervision was presented in Section 1.

Surdeanu et al. (2012) discuss a noisy-or method for combining the scores of various sentence level models to rank a relation during evaluation. In our approach, we introduce the noisy-or mechanism in the training phase of the algorithm.

Our work is inspired from previous works like Roth and tau Yih (2004). The use of ILP for this problem facilitates easy incorporation of different constraints and to the best of our knowledge, has not been investigated by the community.

4 Experiments

The experimental runs were carried out using the publicly available Stanford's distantly supervised slot-filling system[3] (Surdeanu et al., 2011) and Hoffmann et al. (2011) code-base[4].

[3] http://nlp.stanford.edu/software/mimlre.shtml

[4] http://www.cs.washington.edu/ai/raphaelh/mr/

Datasets and Evaluation

We report results on two standard datasets used as benchmarks by the community namely KBP and Riedel datasets. A complete description of these datasets is provided in Surdeanu et al. (2012).

The evaluation setup and module is the same as that described in Surdeanu et al. (2012). We also use the same set of features used by the various systems in the package to ensure that the approaches are comparable. As in previous work, we report precision/recall (P/R) graphs to evaluate the various techniques.

We used the publicly available *lp_solve* package[5] to solve our inference problems.

Performance of ILP

Use of ILP raises concerns about performance as it is NP-hard. In our problem we solve a separate ILP for every entity pair. The number of variables is limited by the number of mentions for the given entity pair. Empirically, on the KBP dataset (larger of the two datasets), Hoffmann takes around 1hr to run. Our ILP formulation takes around 8.5hrs. However, MIMLRE algorithm (EM-based) takes around 23hrs to converge.

Results

We would primarily like to highlight two settings on which we report the P/R curves and contrast it with Hoffmann et al. (2011). Firstly, we replace the approximate inference in that work with our ILP-based exact inference; we call this setting the *hoffmann-ilp*. Secondly, we replace the deterministic-or in the model with a noisy-or, and call this setting the *noisy-or*. We further compare our approach with Surdeanu et al. (2012). The P/R curves for the various techniques on the two datasets are shown in Figures 1 and 2.

We further report the highest F1 point in the P/R curve for both the datasets in Tables 1 and 2.

Table 1 : Highest F1 point in P/R curve : KBP Dataset

	Precision	Recall	F1
Hoffmann	0.306451619	0.197916672	0.2405063349
MIMLRE	0.28061223	0.286458343	0.2835051518
Noisy-OR	0.297002733	0.189236104	0.2311770916
Hoffmann-ilp	0.293010741	0.189236104	0.2299577976

Discussion

We would like to discuss the results in the above two scenarios.

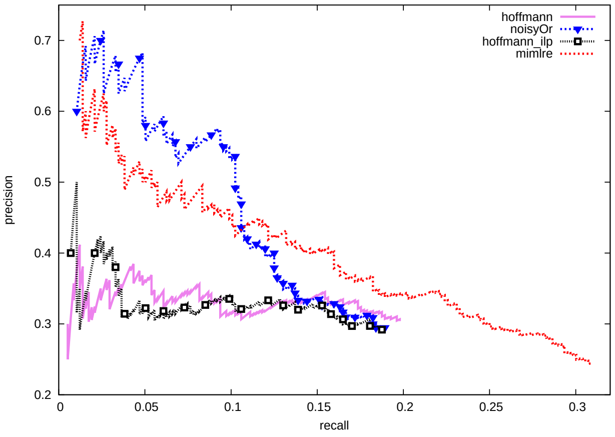

Figure 1: Results : KBP dataset

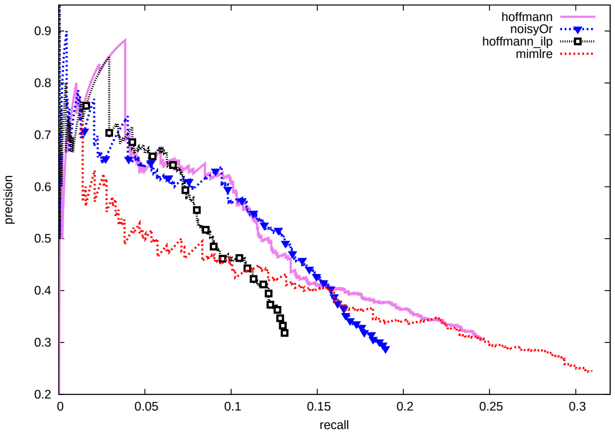

Figure 2: Results : Riedel dataset

1. Performance of *hoffmann-ilp*

   On the KBP dataset, we observe that *hoffmann-ilp* has higher precision in the range of 0.05 to 0.1 at lower recall (0 to 0.04). In other parts of the curve it is very close to the baseline (although hoffmann's algorithm is slightly better). In Table 1, we notice that recall of *hoffmann-ilp* is lower in comparison with hoffmann's algorithm.

   On the Riedel dataset, we observe that *hoffmann-ilp* has better precision (0.15 to 0.2) than MIMLRE within recall of 0.1. At recall $>$ 0.1, precision drops drastically. This is because, *hoffmann-ilp* predicts significantly more nil labels. However, nil labels are not part of the label-set in the P/R curves reported in the community. In Table 2, we see that *hoffmann-ilp* has higher precision (0.04) compared to Hoffmann's algorithm.

2. Performance of *noisy-or*

[5] `http://lpsolve.sourceforge.net/5.5/`

Table 2 : Highest F1 point in P/R curve : Riedel Dataset

	Precision	Recall	F1
Hoffmann	0.32054795	0.24049332	0.27480916
MIMLRE	0.28061223	0.28645834	0.28350515
Noisy-OR	0.317	0.18139774	0.23075178
Hoffmann-ilp	0.36701337	0.12692702	0.18862161

In Figure 1 we see that there is a big jump in precision (around 0.4) of *noisy-or* compared to Hoffmann's model in most parts of the curve on the KBP dataset. However, in Figure 2 (Riedel dataset), we do not see such a trend. Although, we do perform better than MIMLRE (Surdeanu et al., 2012) (precision > 0.15 for recall < 0.15).

On both datasets, *noisy-or* has higher precision than MIMLRE, as seen from Tables 1 and 2. However, the recall reduces. More investigation in this direction is part of future work.

5 Conclusion

In this paper we described an important addition to Hoffmann's model by the use of the *noisy-or* soft constraint to further relax the *at least one* assumption. Since we posed the inference procedures in Hoffmann using ILP, we could easily add this constraint during the training and inference.

Empirically, we showed that the resulting P/R curves have a significant performance boost over Hoffmann's algorithm as a result of this newly added constraint. Although our system has a lower recall when compared to MIMLRE (Surdeanu et al., 2012), it performs competitively w.r.t the precision at low recall.

As part of immediate future work, we would like to improve the system recall. Our ILP formulation provides a good framework to add new type of constraints to the problem. In the future, we would like to experiment with other constraints like modeling the selectional preferences of entity types.

References

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

Zhou GuoDong, Su Jian, Zhang Jie, and Zhang Min. 2005. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 427–434, Stroudsburg, PA, USA. Association for Computational Linguistics.

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 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 541–550, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 1003–1011, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases: Part III*, ECML PKDD'10, pages 148–163, Berlin, Heidelberg. Springer-Verlag.

Dan Roth and Wen tau Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *In Proceedings of CoNLL-2004*, pages 1–8.

Sampath Srinivas. 2013. A generalization of the noisy-or model. *CoRR*, abs/1303.1479.

Mihai Surdeanu and Massimiliano Ciaramita. 2007. Robust information extraction with perceptrons. In *Proceedings of the NIST 2007 Automatic Content Extraction Workshop (ACE07)*, March.

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 Fourth Text Analysis Conference (TAC 2011)*, Gaithersburg, Maryland, USA, November.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 455–465, Stroudsburg, PA, USA. Association for Computational Linguistics.