# A convex relaxation for weakly supervised relation extraction

**Édouard Grave**
EECS Department
University of California, Berkeley
grave@berkeley.edu

## Abstract

A promising approach to relation extraction, called weak or distant supervision, exploits an existing database of facts as training data, by aligning it to an unlabeled collection of text documents. Using this approach, the task of relation extraction can easily be scaled to hundreds of different relationships. However, distant supervision leads to a challenging multiple instance, multiple label learning problem. Most of the proposed solutions to this problem are based on non-convex formulations, and are thus prone to local minima. In this article, we propose a new approach to the problem of weakly supervised relation extraction, based on discriminative clustering and leading to a convex formulation. We demonstrate that our approach outperforms state-of-the-art methods on the challenging dataset introduced by Riedel et al. (2010).

## 1 Introduction

Information extraction refers to the broad task of automatically extracting structured information from unstructured documents. An example is the extraction of named entities and the relations between those entities from natural language texts. In the age of the world wide web and big data, information extraction is quickly becoming pervasive. For example, in 2013, more than $130,000$ scientific articles were published about cancer. Keeping track with that quantity of information is almost impossible, and it is thus of utmost importance to transform the knowledge contained in this massive amount of documents into structured databases.

Traditional approaches to information extraction relies on supervised learning, yielding high precision and recall results (Zelenko et al., 2003). Unfortunately, these approaches need large amount of labeled data, and thus do not scale well to the great number of different types of fact found on the Web or in scientific articles. A promising approach, called distant or weak supervision, is to exploit an existing database of facts as training data, by aligning it to an unlabeled collection of text documents (Craven and Kumlien, 1999).

Knowledge base

| $r$ | $e_1$ | $e_2$ |
|---|---|---|
| BornIn | Lichtenstein | New York City |
| DiedIn | Lichtenstein | New York City |

| Sentences | Latent labels |
|---|---|
| *Roy Lichtenstein was born in New York City, into an upper-middle-class family.* | BornIn |
| *In 1961, Leo Castelli started displaying Lichtenstein's work at his gallery in New York.* | None |
| *Lichtenstein died of pneumonia in 1997 in New York City.* | DiedIn |

Figure 1: An example of a knowledge database comprising two facts and training sentences obtained by aligning this database to unlabeled text.

In this article, we are interested in weakly supervised extraction of binary relations. A challenge pertaining to weak supervision is that the obtained training data is noisy and ambiguous (Riedel et al., 2010). Let us start with an example: if the fact Attended(Turing, King′s College) exists in the knowledge database and we observe the sentence

> *Turing studied as an undergraduate from 1931 to 1934 at King's College, Cambridge.*

which contains mentions of both entities Turing

*Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1580–1590,
October 25-29, 2014, Doha, Qatar. ©2014 Association for Computational Linguistics

and King's College, then this sentence might express the fact that Alan Turing attended King's College, and thus, might be a useful example for learning to extract the relation Attended. However, the sentence

> *Celebrations for the centenary of Alan Turing are being planned at King's College.*

also contains mentions of Turing and King's College, but do not express the relation Attended. Thus, weak supervision lead to noisy examples. As noted by Riedel et al. (2010), such negative extracted sentences for existing facts can represent more than 30% of the data. Moreover, a given pair of entities, such as (Roy Lichtenstein, New York City), car verify multiple relations, such as BornIn and DiedIn. Weak supervision thus lead to ambiguous examples.

This challenge is illustrated in Fig. 1. A solution to address it is to formulate the task of weakly supervised relation extraction as a multiple instance, multiple label learning problem (Hoffmann et al., 2011; Surdeanu et al., 2012). However, these formulations are often non-convex and thus suffer from local minimum.

In this article, we make the following contributions:

- We propose a new convex relaxation for the problem of weakly supervised relation extraction, based on discriminative clustering,
- We propose an efficient algorithm to solve the associated convex program,
- We demonstrate that our approach obtains state-of-the-art results on the dataset introduced by Riedel et al. (2010).

To our knowledge, this paper is the first to propose a convex formulation for solving the problem of weakly supervised relation extraction.

## 2 Related work

**Supervised learning.** Many approaches based on supervised learning have been proposed to solve the problem of relation extraction, and the corresponding literature is to large to be summarized here. One of the first supervised method for relation extraction was inspired by syntactic parsing: the system described by Miller et al. (1998) combines syntactic and semantic knowledge, and thus, part-of-speech tagging, parsing, named entity recognition and relation extraction all happen at the same time. The problem of relation extraction was later formulated as a classification problem: Kambhatla (2004) proposed to solve this problem using maximum entropy models using lexical, syntactic and semantic features. Kernel methods for relation extraction, based on shallow parse trees or dependency trees were introduced by Zelenko et al. (2003), Culotta and Sorensen (2004) and Bunescu and Mooney (2005).

**Unsupervised learning.** The open information extraction paradigm, simultaneously proposed by Shinyama and Sekine (2006) and Banko et al. (2007), does not rely on any labeled data or even existing relations. Instead, open information extraction systems only use an unlabeled corpus, and output a set of extracted relations. Such systems are based on clustering (Shinyama and Sekine, 2006) or self-supervision (Banko et al., 2007). One of the limitations of these systems is the fact that they extract uncanonicalized relations.

**Weakly supervised learning.** Weakly supervised learning refers to a broad class of methods, in which the learning system only have access to partial, ambiguous and noisy labeling. Craven and Kumlien (1999) were the first to propose a weakly supervised relation extractor. They aligned a knowledge database (the Yeast Protein Database) with scientific articles mentioning a particular relation, and then used the extracted sentences to learn a classifier for extracting that relation.

Later, many different sources of weak labelings have been considered. Bellare and McCallum (2007) proposed a method to extract bibliographic relations based on conditional random fields and used a database of BibTex entries as weak supervision. Wu and Weld (2007) described a method to learn relations based on Wikipedia infoboxes. Knowledge databases, such as Freebase[1] (Mintz et al., 2009; Sun et al., 2011) and YAGO[2] (Nguyen and Moschitti, 2011) were also considered as a source of weak supervision.

**Multiple instance learning.** The methods we previously mentionned transform the weakly supervised problem into a fully supervised one, leading to noisy training datasets (see Fig. 1). Mul-

[1] www.freebase.com

[2] www.mpi-inf.mpg.de/yago-naga/yago

tiple instance learning (Dietterich et al., 1997) is a paradigm in which the learner receives bags of examples instead of individual examples. A positively labeled bag contains *at least one* positive example, but might also contains negative examples. In the context of relation extraction, Bunescu and Mooney (2007) introduced a kernel method for multiple instance learning, while Riedel et al. (2010) proposed a solution based on a graphical model.

Both these methods allow only one label per bag, which is an asumption that is not true for relation extraction (see Fig. 1). Thus, Hoffmann et al. (2011) proposed a multiple instance, multiple label method, based on an undirected graphical model, to solve the problem of weakly supervised relation extraction. Finally, Surdeanu et al. (2012) also proposed a graphical model to solve this problem. One of their main contributions is to capture dependencies between relation labels, such as the fact that two labels cannot be generated jointly (*e.g.* the relations `SpouseOf` and `BornIn`).

**Discriminative clustering.** Our approach is based on the discriminative clustering framework, introduced by Xu et al. (2004). The goal of discriminative clustering is to find a labeling of the data points leading to a classifier with low classification error. Different formulations of discriminative clustering have been proposed, based on support vector machines (Xu et al., 2004), the squared loss (Bach and Harchaoui, 2007) or the logistic loss (Joulin et al., 2010). A big advantage of discriminative clustering is that weak supervision or prior information can easily be incorporated. Our work is closely related to the method proposed by Bojanowski et al. (2013) for learning the names of characters in movies.

## 3 Weakly supervised relation extraction

In this article, our goal is to extract *binary relations* between entities from natural language text. Given a set of entities, a binary relation $r$ is a collection of ordered pairs of entities. The statement that a pair of entities $(e_1, e_2)$ belongs to the relation $r$ is denoted by $r(e_1, e_2)$ and this triple is called a *fact* or *relation instance*. For example, the fact that Ernest Hemingway was born in Oak Park is denoted by `BornIn(Ernest Hemingway, Oak Park)`. A given pair of entities, such as `(Edouard Manet, Paris)`, can belong to different relations, such as `BornIn` and `DiedIn`.

An *entity mention* is a contiguous sequence of tokens refering to an entity, while a *pair mention* or *relation mention candidate* is a sequence of text in which a pair of entities is mentioned. In the following, relation mention candidates will be restricted to pair of entities that are mentioned in the same sentence. For example, the sentence:

*Ernest Hemingway was born in Oak Park.*

contains two entity mentions, corresponding to two relation mention candidates. Indeed, the pairs `(Hemingway, Oak Park)` and `(Oak Park, Hemingway)` are two distinct pairs of entities, where only the first one verifies the relation `BornIn`.

Given a text corpus, *aggregate extraction* corresponds to the task of extracting a set of facts, such that each extracted fact is expressed at least once in the corpus. On the other hand, the task of *sentential extraction* corresponds to labeling each relation mention candidate by the relation it expresses, or by a `None` label if it does not express any relation. Given a solution to the sentential extraction problem, it is possible to construct a solution for the aggregate extraction problem by returning all the facts that were detected. We will follow this approach, by building an instance level classifier, and aggregating the results by extracting the facts that were detected at least once in the corpus.

In the following, we will describe a method to learn such a classifier using a database of facts instead of a set of labeled sentences. This setting is known as *distant supervision* or *weak supervision*, since we do not have access to labeled data on which we could directly train a sentence level relation extractor.

## 4 General approach

In this section, we propose a two step procedure to solve the problem of weakly supervised relation extraction:

1. First, we describe a method to infer the relation labels corresponding to each relation mention candidate of our training set,

2. Second, we train a supervised instance level relation extractor, using the labels infered during step 1.

In the second step of our approach, we will simply use a multinomial logistic regression model. We

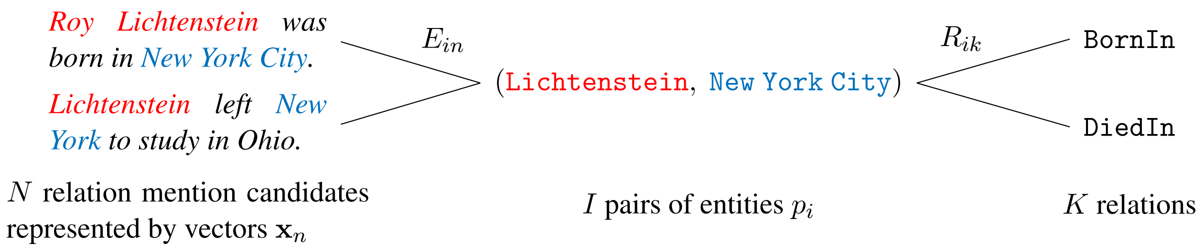
Roy Lichtenstein was born in New York City.
Lichtenstein left New York to study in Ohio.
$E_{in}$
(Lichtenstein, New York City)
$R_{ik}$
BornIn
DiedIn
$N$ relation mention candidates represented by vectors $\mathbf{x}_n$
$I$ pairs of entities $p_i$
$K$ relations

Figure 2: Instance of the weakly supervised relation extraction problem, with notations used in the text.

now describe the approach we propose for the first step.

### 4.1 Notations

Let $(p_i)_{1 \leq i \leq I}$ be a collection of $I$ pairs of entities. We suppose that we have $N$ relation mention candidates, represented by the vectors $(\mathbf{x}_n)_{1 \leq n \leq N}$. Let $\mathbf{E} \in \mathbb{R}^{I \times N}$ be a matrix such that $E_{in} = 1$ if the relation mention candidate $n$ corresponds to the pair of entities $i$, and $E_{in} = 0$ otherwise. The matrix $\mathbf{E}$ thus indicates which relation mention candidate corresponds to which pair of entities. We suppose that we have $K$ relations, indexed by the integers $\{1, ..., K\}$. Let $\mathbf{R} \in \mathbb{R}^{I \times K}$ be a matrix such that $R_{ik} = 1$ if the pair of entities $i$ verifies the relation $k$, and $R_{ik} = 0$ otherwise. The matrix $\mathbf{R}$ thus represents the knowledge database. See Fig. 2 for an illustration of these notations.

### 4.2 Problem formulation

Our goal is to infer a binary matrix $\mathbf{Y} \in \{0,1\}^{N \times (K+1)}$, such that $Y_{nk} = 1$ if the relation mention candidate $n$ express the relation $k$ and $Y_{nk} = 0$ otherwise (and thus, the integer $K+1$ represents the relation None).

We take an approach inspired by the discriminative clustering framework of Xu et al. (2004). We are thus looking for a $(K+1)$-class indicator matrix $\mathbf{Y}$, such that the classification error of an optimal multiclass classifier $f$ is minimum. Given a multiclass loss function $\ell$ and a regularizer $\Omega$, this problem can be formulated as:

$$\min_{\mathbf{Y}} \min_{f} \sum_{n=1}^{N} \ell(\mathbf{y}_n, f(\mathbf{x}_n)) + \Omega(f), \quad \text{s.t. } \mathbf{Y} \in \mathcal{Y}$$

where $\mathbf{y}_n$ is the $n$th line of $\mathbf{Y}$. The constraints $\mathbf{Y} \in \mathcal{Y}$ are added in order to take into account the information from the weak supervision. We will describe in the next section what kind of constraints are considered.

### 4.3 Weak supervision by constraining $\mathbf{Y}$

In this section, we show how the information from the knowledge base can be expressed as constraints on the matrix $\mathbf{Y}$.

First, we suppose that each relation mention candidate express exactly one relation (including the None relation). This means that the matrix $\mathbf{Y}$ contains exactly one 1 per line, which is equivalent to the constraint:

$$\forall n \in \{1, ..., N\}, \sum_{k=1}^{K} Y_{nk} = 1.$$

Second, if the pair $i$ of entities verifies the relation $k$ we suppose that *at least one* relation mention candidate indeed express that relation. Thus we want to impose that for at least one relation mention candidate $n$ such that $E_{in} = 1$, we have $Y_{nk} = 1$. This is equivalent to the constraint:

$$\forall (i,k) \text{ such that } R_{ik} = 1, \sum_{n=1}^{N} E_{in} Y_{nk} \geq 1.$$

Third, if the pair $i$ of entities does not verify the relation $k$, we suppose that no relation mention candidate express that relation. Thus, we impose that for all mention candidate $n$ such that $E_{in} = 1$, we have $Y_{nk} = 0$. This is equivalent to the constraint:

$$\forall (i,k) \text{ such that } R_{ik} = 0, \sum_{n=1}^{N} E_{in} Y_{nk} = 0.$$

Finally, we do not want too many relation mention candidates to be classified as None. We thus impose

$$\forall i \in \{1, ..., I\}, \sum_{n=1}^{N} E_{in} Y_{n(K+1)} \leq c \sum_{n=1}^{N} E_{in},$$

where $c$ is the proportion of relation mention candidates that do not express a relation, for entity pairs that appears in the knowledge database.

We can rewrite these constraints using only matrix operations in the following way:

$$\begin{aligned}&\mathbf{Y}\mathbf{1} = \mathbf{1}\\ &(\mathbf{E}\mathbf{Y}) \circ \mathbf{S} \geq \tilde{\mathbf{R}},\end{aligned} \qquad (1)$$

where $\circ$ is the Hadamard product (a.k.a. the elementwise product), the matrix $\mathbf{S} \in \mathbb{R}^{I \times (K+1)}$ is defined by

$$S_{ik} = \begin{cases} 1 & \text{if } R_{ik} = 1 \\ -1 & \text{if } R_{ik} = 0 \text{ or } k = K+1, \end{cases}$$

and the matrix $\tilde{\mathbf{R}} \in \mathbb{R}^{I \times (K+1)}$ is defined by

$$\tilde{\mathbf{R}} = [\mathbf{R}, -c\mathbf{E}\mathbf{1}].$$

The set $\mathcal{Y}$ is thus defined as the set of matrices $\mathbf{Y} \in \{0,1\}^{N \times (K+1)}$ that verifies those two linear constraints. It is important to note that besides the boolean constraints, the two other constraints are convex.

## 5 Squared loss and convex relaxation

In this section, we describe the problem we obtain when using the squared loss, and its associated convex relaxation. We then introduce an efficient algorithm to solve this problem, by computing its dual.

### 5.1 Primal problem

Following Bach and Harchaoui (2007), we use linear classifiers $\mathbf{W} \in \mathbb{R}^{D \times (K+1)}$, the squared loss and the squared $\ell_2$-norm as the regularizer. In that case, our formulation becomes:

$$\begin{aligned}\min_{\mathbf{Y},\mathbf{W}} \quad & \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\mathbf{W}\|_F^2 + \frac{\lambda}{2}\|\mathbf{W}\|_F^2,\\ \text{s.t.} \quad & \mathbf{Y} \in \{0,1\}^{N \times (K+1)}\\ & \mathbf{Y}\mathbf{1} = \mathbf{1},\\ & (\mathbf{E}\mathbf{Y}) \circ \mathbf{S} \geq \mathbf{R}.\end{aligned}$$

where $\|\cdot\|_F$ is the Frobenius norm and the matrix $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N]^\top \in \mathbb{R}^{N \times D}$ represents the relation mention candidates. Thanks to using the squared loss, we have a closed form solution for the matrix $\mathbf{W}$:

$$\mathbf{W} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_D)^{-1}\mathbf{X}^\top\mathbf{Y}.$$

Replacing the matrix $\mathbf{W}$ by its optimal solution, we obtain the following cost function:

$$\min_{\mathbf{Y}} \frac{1}{2}\mathbf{Y}^\top(\mathbf{I}_N - \mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_D)^{-1}\mathbf{X}^\top)\mathbf{Y}.$$

Then, by applying the Woodbury matrix identity and relaxing the constraint $\mathbf{Y} \in \{0,1\}^{N \times (K+1)}$ into $\mathbf{Y} \in [0,1]^{N \times (K+1)}$, we obtain the following convex quadratic problem in $\mathbf{Y}$:

$$\begin{aligned}\min_{\mathbf{Y}} \quad & \frac{1}{2}\text{tr}\left(\mathbf{Y}^\top(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_N)^{-1}\mathbf{Y}\right),\\ \text{s.t.} \quad & \mathbf{Y} \geq \mathbf{0},\\ & \mathbf{Y}\mathbf{1} = \mathbf{1},\\ & (\mathbf{E}\mathbf{Y}) \circ \mathbf{S} \geq \mathbf{R}.\end{aligned}$$

Since the inequality constraints might be infeasible, we add the penalized slack variables $\xi \in \mathbb{R}^{I \times (K+1)}$, finally obtaining:

$$\begin{aligned}\min_{\mathbf{Y},\xi} \quad & \frac{1}{2}\text{tr}\left(\mathbf{Y}^\top(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_N)^{-1}\mathbf{Y}\right) + \mu\|\xi\|_1\\ \text{s.t.} \quad & \mathbf{Y} \geq \mathbf{0}, \quad \xi \geq 0,\\ & \mathbf{Y}\mathbf{1} = \mathbf{1},\\ & (\mathbf{E}\mathbf{Y}) \circ \mathbf{S} \geq \mathbf{R} - \xi.\end{aligned}$$

This convex problem is a quadratic program. In the following section, we will describe how to solve this problem efficiently, by exploiting the structure of its dual problem.

### 5.2 Dual problem

The matrix $\mathbf{Q} = (\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_N)$ appearing in the quadratic program is an $N$ by $N$ matrix, where $N$ is the number of mention relation candidates. Computing its inverse is thus expensive, since $N$ can be large. Instead, we propose to solve the dual of this problem. Introducing dual variables $\Lambda \in \mathbb{R}^{I \times (K+1)}$, $\Sigma \in \mathbb{R}^{N \times (K+1)}$ and $\nu \in \mathbb{R}^N$, the dual problem is equal to

$$\begin{aligned}\min_{\Lambda,\Sigma,\nu} \quad & \frac{1}{2}\text{tr}\left(\mathbf{Z}^\top\mathbf{Q}\mathbf{Z}\right) - \text{tr}\left(\Lambda^\top\mathbf{R}\right) - \nu^\top 1\\ \text{s.t.} \quad & 0 \leq \Lambda_{ik} \leq \mu, \quad 0 \leq \Sigma_{nk},\end{aligned}$$

where

$$\mathbf{Z} = \mathbf{E}^\top(\mathbf{S} \circ \Lambda) + \Sigma + \nu\mathbf{1}^\top.$$

The derivation of this dual problem is given in Appendix A.

Solving the dual problem instead of the primal has two main advantages. First, the dual does not depend on the inverse of the matrix $\mathbf{Q}$, while the primal does. Since traditional features used for relation extraction are indicators of lexical, syntactic and named entities properties of the relation mention candidates, the matrix $\mathbf{X}$ is extremely sparse.

Using the dual problem, we can thus exploit the sparsity of the matrix $\mathbf{X}$ in the optimization procedure. Second, the constraints imposed on dual variables are simpler than constraints imposed on primal variables. Again, we will exploit this structure in the proposed optimization procedure.

Given a solution of the dual problem, the associated primal variable $\mathbf{Y}$ is equal to:

$$\mathbf{Y} = (\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I}_N)\mathbf{Z}.$$

Thus, we do not need to compute the inverse of the matrix $(\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I}_N)$ to obtain a solution to the primal problem once we have solved the dual.

### 5.3 Optimization of the dual problem

We propose to solve the dual problem using the accelerated projected gradient descent algorithm (Nesterov, 2007; Beck and Teboulle, 2009). Indeed, computing the gradient of the dual cost function is efficient, since the matrix $\mathbf{X}$ is sparse. Moreover, the constraints on the dual variables are simple and it is thus efficient to project onto this set of constraints. See Appendix B for more details.

**Complexity.** The overall complexity of one step of the accelerated projected gradient descent algorithm is $O(NFK)$, where $F$ is the average number of features per relation mention candidate. This means that the complexity of solving the quadratic problem corresponding to our approach is linear with respect to the number $N$ of relation mention candidates, and thus our algorithm can scale to large datasets.

### 5.4 Discussion

Before moving to the experimental sections of this article, we would like to discuss some properties of our approach.

**Kernels.** First of all, one should note that our proposed formulation only depends on the (linear) kernel matrix $\mathbf{X}\mathbf{X}^T$. It is thus possible to replace this matrix by any other kernel. However, in the case of a general kernel, the optimization algorithm presented in the previous section has a quadratic complexity $O(KN^2)$ with respect to the number $N$ of relation mention candidates, and it is thus not applicable *as is*. We plan to explore the use of kernels in future work.

**Rounding.** Given a continuous solution $\mathbf{Y} \in [0,1]^{N \times (K+1)}$ of the relaxed problem, a very simple way to obtain a relation label for each relation mention candidate of the training set is to compute the orthogonal projection of the matrix $\mathbf{Y}$ on the set of indicator matrices

$$\left\{ \mathbf{M} \in \{0,1\}^{N \times (K+1)} \mid \mathbf{M}\mathbf{1} = \mathbf{1} \right\}.$$

This projection consists in taking the maximum value along the rows of the matrix $\mathbf{Y}$. It should be noted that the obtained matrix does not necessarily verify the inequality constraints defined in Eq. 1. In the following, we will use this rounding, refered to as *argmax rounding*, to obtain relation labels for each relation mention candidate.

## 6 Dataset and features

In this section, we describe the dataset used in the experimental section and the features used to represent the data.

### 6.1 Dataset

We consider the dataset introduced by Riedel et al. (2010). This dataset consists of articles from the New York Times corpus (Sandhaus, 2008), from which named entities where extracted and tagged using the Stanford named entity recognizer (Finkel et al., 2005). Consecutive tokens with the same category were treated as a single mention. These named entity mentions were then aligned with the Freebase knowledge database, by using a string match between the mentions and the canonical names of entities in Freebase.

### 6.2 Features

We use the features extracted by Riedel et al. (2010), which were first introduced by Mintz et al. (2009). These features capture how two entity mentions are related in a given sentence, based on syntactic and lexical properties. Lexical features include: the sequence of words between the two entities, a window of $k$ words before the first entity and after the second entity, the corresponding part-of-speech tags, *etc.*. Syntactic features are based on the dependency tree of the sentence, and include: the path between the two entities, neighbors of the two entities that do not belong to the path. The OpenNLP[3] part-of-speech tagger and the Malt parser (Nivre et al., 2007) were used to extract those features.

[3] `opennlp.apache.org`

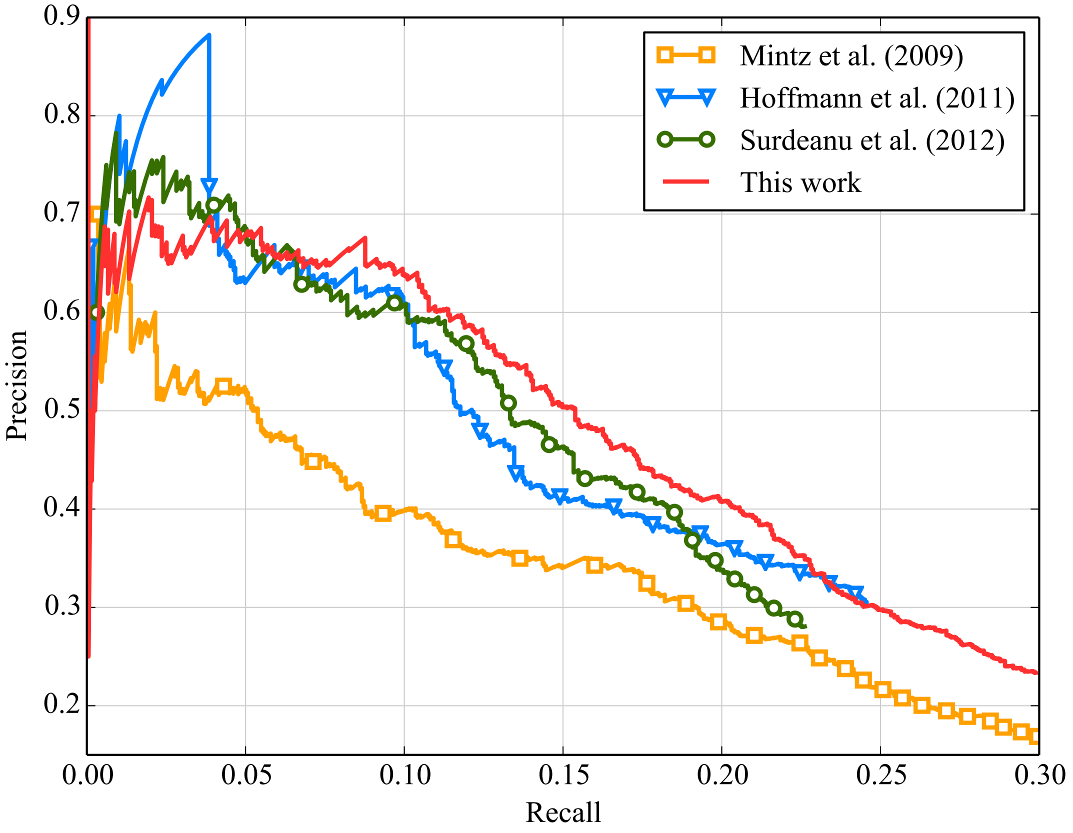

Figure 3: Precision/recall curves for different methods on the Riedel et al. (2010) dataset, for the task of aggregate extraction.

### 6.3 Implementation details

In this section, we discuss some important implementation details.

**Kernel normalization.** We normalized the kernel matrix $\mathbf{X}\mathbf{X}^\top$, so that its diagonal coefficients are equal to 1. This corresponds to normalizing the vectors $\mathbf{x}_n$ so that they have a unit $\ell_2$-norm.

**Choice of parameters.** We kept 20% of the examples from the training set as a validation set, in order to choose the parameters of our method. We then re-train a model on the whole training set, using the chosen parameters.

## 7 Experimental evaluation

In this section, we evaluate our approach to weakly supervised relation extraction by comparing it to state-of-the art methods.

### 7.1 Baselines

We now briefly present the different methods we compare to.

**Mintz et al.** This baseline corresponds to the method described by Mintz et al. (2009). We use the implementation of Surdeanu et al. (2012), which slightly differs from the original method: each relation mention candidate is treated independently (and not collapsed across mentions for a given entity pair). This strategy allows to predict multiple labels for a given entity pair, by OR-ing the predictions for the different mentions.

**Hoffmann et al.** This method, introduced by Hoffmann et al. (2011), is based on probabilistic graphical model of multi-instance multi-label learning. They proposed a learning method for this model, based on the perceptron algorithm (Collins, 2002) and a greedy search for the inference. We use the publicly available code of Hoffmann *et al.*[4].

**Surdeanu et al.** Finally, we compare our method to the one described by Surdeanu et al. (2012). This method is based on a two-layer graphical model, the first layer corresponding to

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

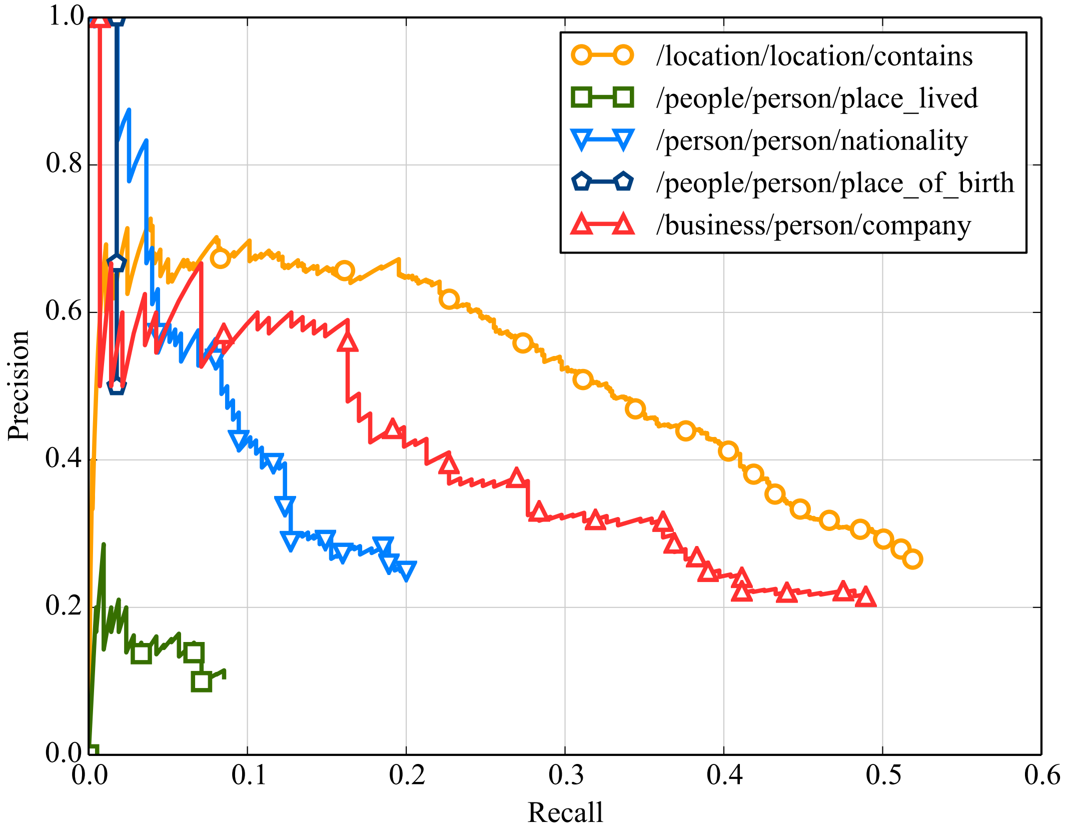

Figure 4: Precision/recall curves per relation for our method, on the Riedel et al. (2010) dataset, for the task of aggregate extraction.

a relation classifier at the mention level, while the second layer is aggregating the different prediction for a given entity pair. In particular, this second layer capture dependencies between relation labels, such as the fact that two labels cannot be generated jointly (*e.g.* the relations `SpouseOf` and `BornIn`). This model is trained by using hard discriminative Expectation-Maximization. We use the publicly available code of Surdeanu *et al.*[5].

### 7.2 Precision / recall curves

Following standard practices in relation extraction, we report precision/recall curves for the different models. In order to rank aggregate extractions for our model, the score of an extracted fact $r(e_1, e_2)$ is set to the maximal score of the different extractions of that fact. This is sometimes refered to as the soft-OR function.

### 7.3 Discussion

**Comparison with the state-of-the-art.** We report results for the different methods on the dataset introduced by Riedel et al. (2010) in Fig. 3. We observe that our approach generally outperforms the state of the art. Indeed, at equivalent recall, our method achieves better (or similar) precision than the other methods, except for very low recall (smaller than $0.05$). The improvement over the methods proposed by Hoffmann et al. (2011) and Surdeanu et al. (2012), which are currently the best published results on this dataset, can be as high as 5 points in precision for the same recall point. Moreover, our method achieves a higher recall $(0.30)$ than these two methods $(0.25)$.

**Performance per relation.** The dataset introduced by Riedel et al. (2010) is highly unbalanced: for example, the most common relation, `/location/location/contains`, represents almost half of the positive relations, while some relations are mentioned less than ten times. We thus decided to also report precision/recall curves for the five most common relations of that dataset in Fig. 4. First, we observe that the perfomances vary a lot from a relation to another. The frequence of the different relations is not the

[5] `nlp.stanford.edu/software/mimlre.shtml`

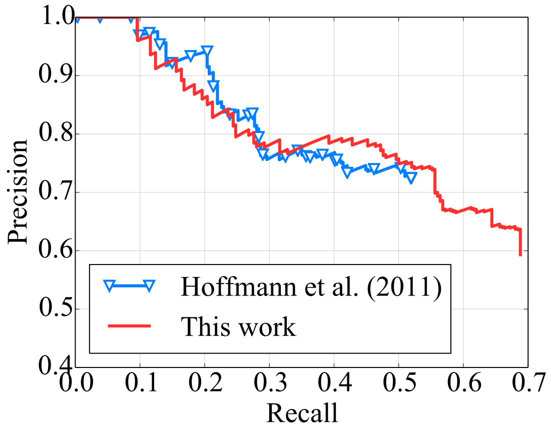

Figure 5: Precision/recall curves for the task of sentential extraction, on the manually labeled dataset of Hoffmann et al. (2011).

only factor in those discrepancies. Indeed, the relation /people/person/place_lived and the relation /people/person/place_of_birth are more frequent than the relation /business/person/company, but the extraction of the later works much better than the extraction of the two first.

Upon examination of the data, this can partly be explained by the fact that almost no sentences extracted for the relation /people/person/place_of_birth in fact express this relation. In other words, many facts present in Freebase are not expressed in the corpus, and are thus impossible to extract. On the other hand, most facts for the relation /people/person/place_lived are missing in Freebase. Therefore, many extractions produced by our system are considered false, but are in fact true positives. The problem of incomplete knowledge base was studied by Min et al. (2013).

**Sentential extraction.** We finally report precision/recall curves for the task of sentential extraction, in Fig. 5, using the manually labeled dataset of Hoffmann et al. (2011). We observe that for most values of recall, our method achieves similar precision that the one proposed by Hoffmann et al. (2011), while extending the highest recall from 0.52 to 0.68. Thanks to this higher recall, our method achieves a highest F1 score of 0.66, compared to 0.61 obtained by the method proposed by Hoffmann et al. (2011).

| Method | Runtime |
|---|---|
| Mintz et al. (2009) | 7 min |
| Hoffmann et al. (2011) | 2 min |
| Surdeanu et al. (2012) | 3 hours |
| This work | 3 hours |

Table 1: Comparison of running times for the different methods compared in the experimental section.

## 8 Conclusion

In this article, we introduced a new formulation for weakly supervised relation extraction. Our method is based on a constrained discriminative formulation of the multiple instance, multiple label learning problem. Using the squared loss, we obtained a convex relaxation of this formulation, allowing us to obtain an approximate solution to the initial integer quadratic program. Thus, our method is not sensitive to initialization. We demonstrated the competitiveness of our approach on the dataset introduced by Riedel et al. (2010), on which our method outperforms the state of the art methods for weakly supervised relation extraction, on both aggregate and sentential extraction.

As noted earlier, another advantage of our method is the fact that it is easily kernelizable. We would like to explore the use of kernels, such as the ones introduced by Zelenko et al. (2003), Culotta and Sorensen (2004) and Bunescu and Mooney (2005), in future work. We believe that such kernels could improve the relatively low recall obtained so far by weakly supervised method for relation extraction.

## Acknowledgments

The author is supported by a grant from Inria (Associated-team STATWEB) and would like to thank Armand Joulin for helpful discussions.

## References

Francis Bach and Zaïd Harchaoui. 2007. DIFFRAC: a discriminative and flexible framework for clustering. In *Adv. NIPS*.

Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open information extraction for the web. In *IJCAI*.

Amir Beck and Marc Teboulle. 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1).

Kedar Bellare and Andrew McCallum. 2007. Learning extractors from unlabeled text using relevant databases. In *Sixth international workshop on information integration on the web*.

Piotr Bojanowski, Francis Bach, Ivan Laptev, Jean Ponce, Cordelia Schmid, and Josef Sivic. 2013. Finding actors and actions in movies. In *Proceedings of ICCV*.

Razvan Bunescu and Raymond Mooney. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of HLT-EMNLP*.

Razvan Bunescu and Raymond Mooney. 2007. Learning to extract relations from the web using minimal supervision. In *Proceedings of the ACL*.

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of EMNLP*.

Mark Craven and Johan Kumlien. 1999. Constructing biological knowledge bases by extracting information from text sources. In *ISMB*, volume 1999.

Aron Culotta and Jeffrey Sorensen. 2004. Dependency tree kernels for relation extraction. In *Proceedings of the ACL*.

Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1).

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the ACL*.

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 ACL*.

Armand Joulin, Jean Ponce, and Francis Bach. 2010. Efficient optimization for discriminative latent class models. In *Adv. NIPS*.

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

Scott Miller, Michael Crystal, Heidi Fox, Lance Ramshaw, Richard Schwartz, Rebecca Stone, and Ralph Weischedel. 1998. Algorithms that learn to extract information. In *Proceedings of MUC-7*.

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 HLT-NAACL*.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the ACL-IJCNLP*.

Yurii Nesterov. 2007. Gradient methods for minimizing composite objective function.

Truc-Vien T Nguyen and Alessandro Moschitti. 2011. End-to-end relation extraction using distant supervision from external semantic repositories. In *Proceedings of the ACL*.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(02).

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases*.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12).

Yusuke Shinyama and Satoshi Sekine. 2006. Preemptive information extraction using unrestricted relation discovery. In *Proceedings of the HLT-NAACL*.

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, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of EMNLP-CoNLL*.

Fei Wu and Daniel S Weld. 2007. Autonomously semantifying wikipedia. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*.

Linli Xu, James Neufeld, Bryce Larson, and Dale Schuurmans. 2004. Maximum margin clustering. In *Adv. NIPS*.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *The Journal of Machine Learning Research*, 3.

## Appendix A Derivation of the dual

In this section, we derive the dual problem of the quadratic program of section 5. We introduce dual variables $\Lambda \in \mathbb{R}^{I\times(K+1)}$, $\Sigma \in \mathbb{R}^{N\times(K+1)}$, $\Omega \in \mathbb{R}^{I\times(K+1)}$ and $\nu \in \mathbb{R}^N$, such that $\Lambda \geq 0$, $\Sigma \geq 0$ and $\Omega \geq 0$.

The Lagrangian of the problem is

$$\begin{aligned} &\frac{1}{2}\text{tr}\left(\mathbf{Y}^\top(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_N)^{-1}\mathbf{Y}\right) + \mu\sum_{i,k}\xi_{ik} \\ &\quad - \text{tr}\left(\Lambda^\top((\mathbf{E}\mathbf{Y})\circ\mathbf{S} - \mathbf{R} + \xi)\right) \\ &\quad - \text{tr}(\Sigma^\top\mathbf{Y}) - \text{tr}(\Omega^\top\xi) - \nu^\top(\mathbf{Y}\mathbf{1} - \mathbf{1}). \end{aligned}$$

To find the dual function $g$ we minimize the Lagrangian over $\mathbf{Y}$ and $\xi$. Minimizing over $\xi$, we find that the dual function is equal to $-\infty$ unless $\mu - \Lambda_{ik} - \Omega_{ik} = 0$, in which case, we are left with

$$\begin{aligned} &\frac{1}{2}\text{tr}\left(\mathbf{Y}^\top(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_N)^{-1}\mathbf{Y}\right) \\ &- \text{tr}((\Lambda\circ\mathbf{S})^\top\mathbf{E}\mathbf{Y}) - \text{tr}(\Sigma^\top\mathbf{Y}) - \text{tr}(\mathbf{1}\nu^\top\mathbf{Y}) \\ &\qquad + \text{tr}(\Lambda^\top\mathbf{R}) + \nu^\top\mathbf{1}. \end{aligned}$$

Minimizing over $\mathbf{Y}$, we then obtain

$$\mathbf{Y} = (\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_N)(\mathbf{E}^\top(\mathbf{S}\circ\Lambda) + \Sigma + \nu\mathbf{1}^\top).$$

Replacing $\mathbf{Y}$ by its optimal value, we then obtain the dual function

$$-\frac{1}{2}\text{tr}\left(\mathbf{Z}^\top\mathbf{Q}\mathbf{Z}\right) + \text{tr}\left(\Lambda^\top\mathbf{R}\right) + \nu^\top\mathbf{1}.$$

where

$$\begin{aligned} \mathbf{Q} &= (\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_N), \\ \mathbf{Z} &= \mathbf{E}^\top(\mathbf{S}\circ\Lambda) + \Sigma + \nu\mathbf{1}^\top. \end{aligned}$$

Thus, the dual problem is

$$\begin{aligned} \max_{\Lambda,\Sigma,\nu} \quad & -\frac{1}{2}\text{tr}\left(\mathbf{Z}^\top\mathbf{Q}\mathbf{Z}\right) + \text{tr}\left(\Lambda^\top\mathbf{R}\right) + \nu^\top 1 \\ \text{s.t.} \quad & 0 \leq \Lambda_{ik}, \quad 0 \leq \Sigma_{nk}, \quad 0 \leq \Omega_{ik}, \\ & \mu - \Lambda_{ik} - \Omega_{ik} = 0. \end{aligned}$$

We can then eliminate the dual variable $\Omega$, since the constraints $\Omega_{ik} = \mu - \Lambda_{ik}$ and $\Omega_{ik} \geq 0$ are equivalent to $\mu \geq \Lambda_{ik}$. We finally obtain

$$\begin{aligned} \max_{\Lambda,\Sigma,\nu} \quad & -\frac{1}{2}\text{tr}\left(\mathbf{Z}^\top\mathbf{Q}\mathbf{Z}\right) + \text{tr}\left(\Lambda^\top\mathbf{R}\right) + \nu^\top 1 \\ \text{s.t.} \quad & 0 \leq \Lambda_{ik} \leq \mu, \quad 0 \leq \Sigma_{nk}. \end{aligned}$$

## Appendix B Optimization details

**Gradient of the dual cost function.** The gradient of the dual cost function $f$ with respect to the dual variables $\Sigma$, $\Lambda$ and $\nu$ is equal to

$$\begin{aligned} \nabla_\Sigma f &= (\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_N)\mathbf{Z}, \\ \nabla_\Lambda f &= \left((\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_N)\mathbf{Z}\mathbf{E}^\top\right)\circ\mathbf{S} - \mathbf{R}, \\ \nabla_\nu f &= (\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_N)\mathbf{Z}\mathbf{1} - \mathbf{1}. \end{aligned}$$

The most expensive step to compute those gradients is to compute the matrix product $\mathbf{X}\mathbf{X}^\top\mathbf{Z}$. Since the matrix $\mathbf{X}$ is sparse, we efficiently compute this product by first computing the product $\mathbf{X}^\top\mathbf{Z}$, and then by left multiplying the result by $\mathbf{X}$. The complexity of these two operations is $O(NFK)$, where $F$ is the average number of features per relation mention candidate.

**Projecting $\Sigma$ and $\Lambda$.** The componentwise projection operators associated to the constraints on $\Sigma$ and $\Lambda$ are defined by:

$$\begin{aligned} \text{proj}_\Sigma(\Sigma_{nk}) &= \max(0, \Sigma_{nk}), \\ \text{proj}_\Lambda(\Lambda_{ik}) &= \max(0, \min(\mu, \Lambda_{ik})). \end{aligned}$$

The complexity of projecting $\Sigma$ and $\Lambda$ is $O(NK)$. Thus, the cost of those operations is ne gligible compared to the cost of computing the gradients of the dual cost function.