A convex relaxation for weakly supervised relation extraction

Édouard Grave

University of California, Berkeley
grave@berkeley.edu
During World War II, Turing worked for the Government Code and Cypher School (GC&CS) at Bletchley Park.
Relation extraction

*During World War II, Turing worked for the Government Code and Cypher School (CG&CS) at Bletchley Park.*
Relation extraction

*During World War II, Turing worked for the Government Code and Cypher School (CG&CS) at Bletchley Park.*

Employee(*Alan Turing, CG&CS*)
Contains(*Bletchley Park, CG&CS*)
Relation extraction

During World War II, Turing worked for the Government Code and Cypher School (CG&CS) at Bletchley Park.

Employee(Alan Turing, CG&CS)
Contains(Bletchley Park, CG&CS)

Historically: supervised learning.

Limitations: need labeled data (expensive and time consuming).
  • lot of different relations;
  • different languages
  • etc.
Distant supervision for relation extraction
Craven and Kumlien (1999); Mintz et al. (2009)

Knowledge base

<table>
<thead>
<tr>
<th>$r$</th>
<th>$e_1$</th>
<th>$e_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>BornIn</td>
<td>Lichtenstein</td>
<td>New York City</td>
</tr>
<tr>
<td>DiedIn</td>
<td>Lichtenstein</td>
<td>New York City</td>
</tr>
</tbody>
</table>

Sentences

Roy Lichtenstein was born in New York City, into an upper-middle-class family.

In 1961, Leo Castelli started displaying Lichtenstein's work at his gallery in New York.

Roy Lichtenstein died of pneumonia in 1997 in New York City.
Distant supervision for relation extraction
Craven and Kumlien (1999); Mintz et al. (2009)

Knowledge base

<table>
<thead>
<tr>
<th>$r$</th>
<th>$e_1$</th>
<th>$e_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>BornIn</td>
<td>Lichtenstein</td>
<td>New York City</td>
</tr>
<tr>
<td>DiedIn</td>
<td>Lichtenstein</td>
<td>New York City</td>
</tr>
</tbody>
</table>

Sentences

*Roy Lichtenstein* was born in *New York City*, into an upper-middle-class family.

*In 1961, Leo Castelli started displaying Lichtenstein's work at his gallery in New York.*

*Roy Lichtenstein* died of pneumonia in 1997 in *New York City*.
Distant supervision for relation extraction
Craven and Kumlien (1999); Mintz et al. (2009)

Knowledge base

<table>
<thead>
<tr>
<th>r</th>
<th>e₁</th>
<th>e₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>BornIn</td>
<td>Lichtenstein</td>
<td>New York City</td>
</tr>
<tr>
<td>DiedIn</td>
<td>Lichtenstein</td>
<td>New York City</td>
</tr>
</tbody>
</table>

Sentences

Roy Lichtenstein was born in New York City, into an upper-middle-class family.

In 1961, Leo Castelli started displaying Lichtenstein’s work at his gallery in New York.

Roy Lichtenstein died of pneumonia in 1997 in New York City.

Latent label

BornIn
None
DiedIn
Multiple instance, multiple label learning
Bunescu and Mooney (2007); Riedel et al. (2010); Hoffmann et al. (2011); Surdeanu et al. (2012)

Roy Lichtenstein was born in New York City. Lichtenstein left New York to study in Ohio.

BornIn (Lichtenstein, New York City) DiedIn

Roy Lichtenstein
was born in New York City.
Lichtenstein left New York to study in Ohio.
Multiple instance, multiple label learning
Bunescu and Mooney (2007); Riedel et al. (2010); Hoffmann et al. (2011); Surdeanu et al. (2012)

Roy Lichtenstein was born in New York City.
Lichtenstein left New York to study in Ohio.

N pair mentions represented by vectors $x_n$

$E_{in} = 1$ if pair mention $n$ corresponds to entity pair $i$

$R_{ik} = 1$ if entity pair $i$ verifies relation $k$
Overview

Two steps procedure:
1. infer labels for each pair mention;
2. train supervised instance level relation extractor.

Goal: infer a binary matrix $Y$ such that:
- $Y_{nk} = 1$ if pair mention $n$ express relation $k$;
- $Y_{nk} = 0$ otherwise.

Approach based on discriminative clustering.
I. Discriminative clustering
Discriminative clustering
Xu et al. (2004); Bach and Harchaoui (2007)
Discriminative clustering
Xu et al. (2004); Bach and Harchaoui (2007)
Discriminative clustering
Xu et al. (2004); Bach and Harchaoui (2007)
Given a loss function $\ell$ and a regularizer $\Omega$:

$$\min_{\mathbf{Y}} \min_f \sum_{n=1}^{N} \ell(y_n, f(x_n)) + \Omega(f),$$

s.t. $\mathbf{Y} \in \mathcal{Y}$
II. Weak supervision by constraining Y
Weak supervision by constraining $\mathbf{Y}$

Each pair mention express exactly one relation:
Each pair mention express exactly one relation:

$$\forall n \in \{1, \ldots, N\}, \sum_{k=1}^{K+1} Y_{nk} = 1.$$
Weak supervision by constraining $\Upsilon$

If entity pair $i$ verifies relation $k$, then at least one pair mention $n$ corresponding to the pair $i$ express that relation:
Weak supervision by constraining $\mathbf{Y}$

If entity pair $i$ verifies relation $k$, then at least one pair mention $n$ corresponding to the pair $i$ express that relation:

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

$E_{in} = 1$ if pair mention $n$ corresponds to entity pair $i$
If entity pair \( i \) verifies relation \( k \), then at least one pair mention \( n \) corresponding to the pair \( i \) express that relation:

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

\( E_{in} = 1 \) if pair mention \( n \) corresponds to entity pair \( i \)
Weak supervision by constraining $\mathbf{Y}$

If entity pair $i$ does not verify relation $k$, then no pair mention $n$ corresponding to pair $i$ express that relation:
Weak supervision by constraining $\mathbf{Y}$

If entity pair $i$ does not verify relation $k$, then no pair mention $n$ corresponding to pair $i$ express that relation:

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

$E_{in} = 1$ if pair mention $n$ corresponds to entity pair $i$
Weak supervision by constraining $Y$

If entity pair $i$ does not verify relation $k$, then no pair mention $n$ corresponding to pair $i$ express that relation:

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

$E_{in} = 1$ if pair mention $n$ corresponds to entity pair $i$
Weak supervision by constraining $\mathbf{Y}$

For a given entity pair $i$, at most $c$ percent of pair mentions classified as none:
Weak supervision by constraining $\mathbf{Y}$

For a given entity pair $i$, at most $c$ percent of pair mentions classified as none:

\[ \forall i \in \{1, \ldots, I\}, \quad \sum_{n=1}^{N} E_{in} Y_{n(K+1)} \leq c \sum_{n=1}^{N} E_{in}, \]
Weak supervision by constraining $\mathbf{Y}$

These constraints are equivalent to:

\[ Y_1 = 1, \]
\[ (EY) \circ S \geq R. \]
III. Problem formulation
Problem formulation

We use linear classifiers \( W \in \mathbb{R}^{D \times (K+1)} \) and the squared loss:

\[
\min_{Y, W} \quad \frac{1}{2} \| Y - XW \|_F^2 + \frac{\lambda}{2} \| W \|_F^2,
\]

s.t. \( Y \in \{0, 1\}^{N \times (K+1)} \)

\( Y1 = 1, \)

\((EY) \circ S \geq R.\)
Problem formulation

We use linear classifiers $W \in \mathbb{R}^{D \times (K+1)}$ and the squared loss:

$$\min_{Y, W} \frac{1}{2} \|Y - XW\|_F^2 + \frac{\lambda}{2} \|W\|_F^2,$$

s.t. $Y \in \{0, 1\}^{N \times (K+1)}$

$Y1 = 1$,

$(EY) \circ S \geq R$.

Closed form solution for $W$:

$$W = (X^\top X + \lambda I_D)^{-1} X^\top Y.$$
Problem formulation

Replacing $\mathbf{W}$ by its optimal value:

$$\min_{\mathbf{Y}} \frac{1}{2} \text{tr} \left( \mathbf{Y}^\top (\mathbf{X} \mathbf{X}^\top + \lambda \mathbf{I}_N)^{-1} \mathbf{Y} \right),$$

s.t. $\mathbf{Y} \in \{0, 1\}^{N \times (K+1)}$

$\mathbf{Y}\mathbf{1} = 1,$

$(\mathbf{EY}) \circ \mathbf{S} \geq \mathbf{R}.$

This is a quadratic integer program.
Convex relaxation

Relaxing the constraints $Y \in \{0, 1\}^{N \times (K+1)}$ into $Y \in [0, 1]^{N \times (K+1)}$:

$$\min_Y \frac{1}{2} \text{tr} \left( Y^T (XX^T + \lambda I_N)^{-1} Y \right),$$

s.t. $Y \in [0, 1]^{N \times (K+1)}$

$Y1 = 1,$

$(EY) \circ S \geq R.$

This is a convex quadratic program.

It only depends on the kernel $XX^T$. 
Rounding

Given a solution of the relaxed problem $\mathbf{Y}$, orthogonal projection on:

$$\left\{ \mathbf{M} \in \{0, 1\}^{N \times (K+1)} \mid \mathbf{M1} = 1 \right\}.$$ 

Consists in taking the argmax along the rows of $\mathbf{Y}$. 
Optimization

We optimize the dual because:

- there is no matrix inverse (easy to compute the gradient),
- constraints are simpler (easy to project on the constraints).

We use accelerated projected gradient descent algorithm (FISTA).

Overall complexity: $O(NFK)$.

- $N$ is the number of pair mentions (sentences);
- $F$ is the average number of features;
- $K$ is the number of classes.
IV. Experiments
Experiments: dataset

Dataset introduced by Riedel et al. (2010):

- Articles from the New York Times corpus.
- Entities extracted using Stanford named entities recognizer.
- Entity mentions aligned to Freebase using a string match.

There are

- 52 relations,
- 4,200 entity pairs,
- 120,000 pair mentions.
Experiments: features

We use the features proposed by Mintz et al. (2009):

• Lexical features, such as:
  • sequence of words between entities;
  • window of $k$ words before/after the first/second entity;
  • corresponding part-of-speeches;

• Syntactic features, such as:
  • path in the dependency tree between the two entities;
  • neighbors of the two entities that are not in the path.
Figure: Precision/recall curves for different methods on the Riedel et al. (2010) dataset, for the task of aggregate extraction.
Experiments: results

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

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

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

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

Distant supervision for relation extraction:

- based on a convex formulation;
- discriminative clustering, where distant supervision is used as constraint;
- Competitive to state-of-the-art on aggregate and sentential extraction.
Conclusion

Distant supervision for relation extraction:

- based on a convex formulation;
- discriminative clustering, where distant supervision is used as constraint;
- Competitive to state-of-the-art on aggregate and sentential extraction.

Future work:

- Faster optimization methods for our approach;
- Kernelization of our methods.
Thank you for your attention.
References I


