Optimizing Multivariate Performance Measures for Learning Relation Extraction Models

Ganesh Ramakrishnan (joint work with)
Gholamreza Haffari
Ajay Nagesh

June 16, 2015

Outline

Outline

Relation Extraction

Xerox Corporation is an American multinational company headquartered in Norwalk, Connecticut. On May 21, 2009, it was announced that Ursula Burns would be the CEO of Xerox.

Relation Extraction

Xerox Corporation is an American multinational company headquartered in Norwalk, Connecticut. On May 21, 2009, it was announced that Ursula Burns would be the CEO of Xerox.

Relation Extraction

Xerox Corporation is an American multinational company headquartered in Norwalk, Connecticut. On May 21, 2009, it was announced that Ursula Burns would be the CEO of Xerox.

```
Headquarters(Xerox Corp., Norwalk)
Contains(Norwalk, Connecticut)
CEO(Xerox Corp., Ursula Burns)
```

Relation Extraction

Xerox Corporation is an American multinational company headquartered in Norwalk, Connecticut. On May 21, 2009, it was announced that Ursula Burns would be the CEO of Xerox.

```
Headquarters(Xerox Corp., Norwalk)
Contains(Norwalk, Connecticut)
CEO(Xerox Corp., Ursula Burns)
```

Traditionally: supervised learning

Limitations: not scalable (expensive and time consuming to create labeled data)

Relation Extraction

Xerox Corporation is an American multinational company headquartered in Norwalk, Connecticut. On May 21, 2009, it was announced that Ursula Burns would be the CEO of Xerox.

```
Headquarters(Xerox Corp., Norwalk)
Contains(Norwalk, Connecticut)
CEO(Xerox Corp., Ursula Burns)
```

Traditionally: supervised learning

Limitations: not scalable (expensive and time consuming to create labeled data)

Can we leverage already existing high quality databases (e.g. Freebase) to learn good relation extractors ?

Distant Supervision for Relation Extraction

Mintz et al. (2009)

Knowledge base

r	e_1	e_2
BornIn	Barack Obama	U. S.
PresidentOf	Barack Obama	U. S.

Sentences

- *Barack Obama was born in Honolulu, Hawaii, United States.*
- *Obama left United States this Saturday for a UN summit in Geneva.*
- *President Obama defended his administrations' collection of phone records in the U.S.*

Distant Supervision based Relation Extraction

Mintz et al. (2009)

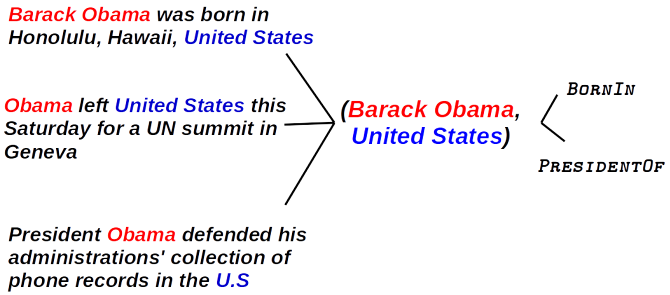
Knowledge base

r	e_1	e_2
BornIn	Barack Obama	U. S.
PresidentOf	Barack Obama	U. S.

Sentences	**Latent Label**
- *Barack Obama was born in Honolulu, Hawaii, United States.*	*BornIn*
- *Obama left United States this Saturday for a UN summit in Geneva.*	*none*
- *President Obama defended his administrations' collection of phone records in the U.S.*	*PresidentOf*

Multiple instance, multiple label

Riedel et al. (2010); Hoffmann et al. (2011); Surdeanu et al. (2012)



Motivation

- In existing approaches, model parameters are often learnt by optimizing performance measures (e.g.: conditional log-likelihood, error rate)

Motivation

- In existing approaches, model parameters are often learnt by optimizing performance measures (e.g.: conditional log-likelihood, error rate)
- However, these are not directly related to evaluation measures (e.g.: F1-score, area under ROC curve)

Motivation

- In existing approaches, model parameters are often learnt by optimizing performance measures (e.g.: conditional log-likelihood, error rate)
- However, these are not directly related to evaluation measures (e.g.: F1-score, area under ROC curve)
- **Training Objective $\nLeftrightarrow$ Evaluation Measure**

Motivation (cont.)

- Can we train better models by directly optimizing task specific performance measures while allowing latent variables to adapt their values

Motivation (cont.)

- Can we train better models by directly optimizing task specific performance measures while allowing latent variables to adapt their values
- Further, can we provide a knob in the training algorithm to favor *precision* more than *recall*

Introduction

- Our work: large margin method to learn parameters of models
  - That contain latent variables
  - Optimize performance measures that are non-linear (e.g. : F_β)

Introduction

- Our work: large margin method to learn parameters of models
  - That contain latent variables
  - Optimize performance measures that are non-linear (e.g. : F_β)
- Outline of our approach: Interleaves concave-convex procedure (CCCP) with dual decomposition

Introduction

- Our work: large margin method to learn parameters of models
  - That contain latent variables
  - Optimize performance measures that are non-linear (e.g. : F_β)
- Outline of our approach: Interleaves concave-convex procedure (CCCP) with dual decomposition
  - CCCP : used to populate latent variables

Introduction

- Our work: large margin method to learn parameters of models
  - That contain latent variables
  - Optimize performance measures that are non-linear (e.g. : F_β)
- Outline of our approach: Interleaves concave-convex procedure (CCCP) with dual decomposition
  - CCCP : used to populate latent variables
  - Dual decomposition: factorizes the hard optimization problem (during training) into independent sub-problems

Introduction

- Our work: large margin method to learn parameters of models
  - That contain latent variables
  - Optimize performance measures that are non-linear (e.g. : F_β)
- Outline of our approach: Interleaves concave-convex procedure (CCCP) with dual decomposition
  - CCCP : used to populate latent variables
  - Dual decomposition: factorizes the hard optimization problem (during training) into independent sub-problems
  - We present linear programming (LP) and local search methods to solve the sub-problems

Outline

Preliminaries

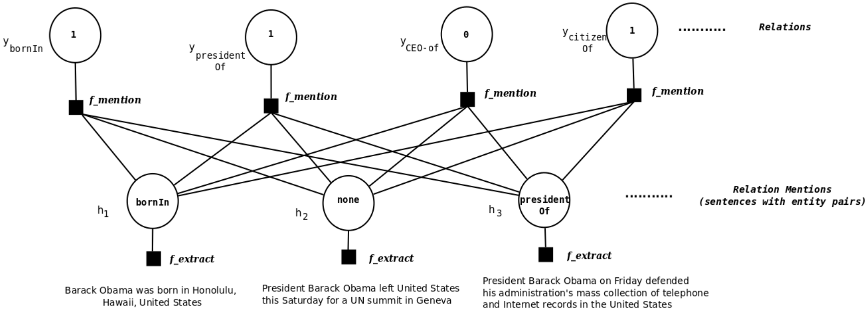
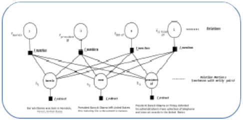


Figure: Graphical model instantiated for entity pair $\mathbf{x} := (\texttt{Barack Obama, United States})$

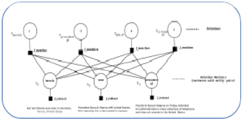
Preliminaries



one data-point

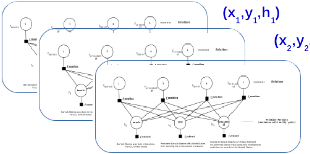
- **x**: entity pair
- **y**: relation labels
- **h**: latent mention labels

Preliminaries



one data-point

- **x**: entity pair
- **y**: relation labels
- **h**: latent mention labels

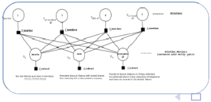


(x_1,y_1,h_1)

(x_2,y_2,h_2)

(x_3,y_3,h_3)

....



(x_n,y_n,h_n)

(X, Y, H) :
Set of all points
in the training dataset

Structured Prediction Learning

- Goal: To find $\mathbf{w} \in R^d$ that minimizes risk

$$R_{f_{\mathbf{w}}}^{\Delta} := \Delta\Big(\big(f_{\mathbf{w}}(\mathbf{x}_1), .., f_{\mathbf{w}}(\mathbf{x}_N)\big), \big(\mathbf{y}_1, .., \mathbf{y}_N\big)\Big)$$

Structured Prediction Learning

- Goal: To find $\mathbf{w} \in R^d$ that minimizes risk

$$R_{f_{\mathbf{w}}}^{\Delta} := \Delta\Big(\big(f_{\mathbf{w}}(\mathbf{x}_1), .., f_{\mathbf{w}}(\mathbf{x}_N)\big), \big(\mathbf{y}_1, .., \mathbf{y}_N\big)\Big)$$

- Most large margin learning algorithms assume that loss is decomposable. So $R_{f_{\mathbf{w}}}^{\Delta}$ is simplified to,

$$R_{f_{\mathbf{w}}}^{\Delta} := \sum_{i=1}^{N} \delta(f_{\mathbf{w}}(\mathbf{x}), \mathbf{y})$$

Structured Prediction Learning

- Goal: To find $\mathbf{w} \in R^d$ that minimizes risk

$$R^{\Delta}_{f_{\mathbf{w}}} := \Delta\Big(\big(f_{\mathbf{w}}(\mathbf{x}_1), .., f_{\mathbf{w}}(\mathbf{x}_N)\big), \big(\mathbf{y}_1, .., \mathbf{y}_N\big)\Big)$$

- Most large margin learning algorithms assume that loss is decomposable. So $R^{\Delta}_{f_{\mathbf{w}}}$ is simplified to,

$$R^{\Delta}_{f_{\mathbf{w}}} := \sum_{i=1}^{N} \delta(f_{\mathbf{w}}(\mathbf{x}), \mathbf{y})$$

- However, for non-decomposable loss functions like $-F_1$, Δ cannot be expressed in terms of δ

Outline

Multi-variate Structured Prediction

- In decomposable structure prediction task, our aim is to learn $\mathbf{w} \in \mathbb{R}^d$ such that for a new entity pair $\mathbf{x}$, we can find:

Multi-variate Structured Prediction

- In decomposable structure prediction task, our aim is to learn $\mathbf{w} \in \mathbb{R}^d$ such that for a new entity pair $\mathbf{x}$, we can find:

$$f_{\mathbf{w}}(\mathbf{x}) := \arg\max_{\mathbf{y}} \max_{\mathbf{h}} \mathbf{w} \cdot \Phi(\mathbf{x}, \mathbf{h}, \mathbf{y})$$

Multi-variate Structured Prediction

- In decomposable structure prediction task, our aim is to learn $\mathbf{w} \in \mathbb{R}^d$ such that for a new entity pair $\mathbf{x}$, we can find:

$$f_{\mathbf{w}}(\mathbf{x}) := \arg\max_{\mathbf{y}} \max_{\mathbf{h}} \mathbf{w} \cdot \Phi(\mathbf{x}, \mathbf{h}, \mathbf{y})$$

- Instead of learning a mapping function from an individual instance to its label, we learn a mapping from all instances to their labels

Multi-variate Structured Prediction

- In decomposable structure prediction task, our aim is to learn $\mathbf{w} \in \mathbb{R}^d$ such that for a new entity pair $\mathbf{x}$, we can find:

$$f_{\mathbf{w}}(\mathbf{x}) := \arg\max_{\mathbf{y}} \max_{\mathbf{h}} \mathbf{w} \cdot \Phi(\mathbf{x}, \mathbf{h}, \mathbf{y})$$

- Instead of learning a mapping function from an individual instance to its label, we learn a mapping from all instances to their labels
- We define the best labeling using the following linear discriminant function

$$\mathbf{f}(\mathbf{X}) := \arg\max_{\mathbf{Y}' \in \mathcal{Y}} \max_{\mathbf{H} \in \mathcal{H}} \left\{ \mathbf{w} \cdot \Psi(\mathbf{X}, \mathbf{H}, \mathbf{Y}') \right\}$$

where $\Psi(\mathbf{X}, \mathbf{H}, \mathbf{Y}') := \sum_{i=1}^{N} \Phi(\mathbf{x}_i, \mathbf{h}_i, \mathbf{y}'_i)$

Training Objective

- Based on margin re-scaling formulation of structured prediction problems, our training objective is:

$$\min_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||_2^2 + C \max_{\mathbf{y}'_1,..,\mathbf{y}'_N} \Big\{ \Delta\Big((\mathbf{y}_1,..,\mathbf{y}_N), (\mathbf{y}'_1,..,\mathbf{y}'_N) \Big)$$

$$+ \sum_{i=1}^{N} \max_{\mathbf{h}} \mathbf{w}\cdot\Phi(\mathbf{x}_i,\mathbf{h},\mathbf{y}'_i) - \sum_{i=1}^{N} \max_{\mathbf{h}} \mathbf{w}\cdot\Phi(\mathbf{x}_i,\mathbf{h},\mathbf{y}_i) \Big\}$$

Training Objective

- Based on margin re-scaling formulation of structured prediction problems, our training objective is:

$$\min_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||_2^2 + C \max_{\mathbf{y}'_1,..,\mathbf{y}'_N} \Bigg\{ \Delta\Big((\mathbf{y}_1,..,\mathbf{y}_N), (\mathbf{y}'_1,..,\mathbf{y}'_N) \Big) + \sum_{i=1}^{N} \max_{\mathbf{h}} \mathbf{w}\cdot\Phi(\mathbf{x}_i,\mathbf{h},\mathbf{y}'_i) - \sum_{i=1}^{N} \max_{\mathbf{h}} \mathbf{w}\cdot\Phi(\mathbf{x}_i,\mathbf{h},\mathbf{y}_i) \Bigg\}$$

- The above objective is non-convex since it is the difference of two convex functions

Convex-concave Procedure (CCCP)

- Refer to Yuille and Rangarajan (2001) and Sriperumbudur and Lanckriet (2012)
- CCCP is a special example of Majorization-Minimization (class of) algorithm(s)
- Elaborate Convergence proof for constrained version by Sriperumbudur and Lanckriet (2012) using Zangwill's global convergence framework

Training Objective

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C \max_{\mathbf{y}'_1,\ldots,\mathbf{y}'_N} \Big\{ \Delta\Big((\mathbf{y}_1,..,\mathbf{y}_N), (\mathbf{y}'_1,..,\mathbf{y}'_N) \Big) + \sum_{i=1}^{N} \max_{\mathbf{h}} \mathbf{w} \cdot \Phi(\mathbf{x}_i, \mathbf{h}, \mathbf{y}'_i) - \sum_{i=1}^{N} \max_{\mathbf{h}} \mathbf{w} \cdot \Phi(\mathbf{x}_i, \mathbf{h}, \mathbf{y}_i) \Big\}$$

Algorithm 1 The Training Algorithm

1: **procedure** OPT-LATENTSVM(X, Y)
2: Initialize $\mathbf{w}^{(0)}$ and set $t = 0$
3: **repeat**
4: **for** $i := 1$ to N **do**

Concave Step:

Convex-concave Procedure (CCCP)

Training Objective

$$\min_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||_2^2 + C \max_{\mathbf{y}'_1,\ldots,\mathbf{y}'_N} \left\{ \Delta\Big((\mathbf{y}_1,..,\mathbf{y}_N),(\mathbf{y}'_1,..,\mathbf{y}'_N)\Big) + \sum_{i=1}^{N} \max_{\mathbf{h}} \mathbf{w}\cdot\Phi(\mathbf{x}_i,\mathbf{h},\mathbf{y}'_i) - \sum_{i=1}^{N} \max_{\mathbf{h}} \mathbf{w}\cdot\Phi(\mathbf{x}_i,\mathbf{h},\mathbf{y}_i) \right\}$$

Algorithm 1 The Training Algorithm

```
procedure OPT-LATENTSVM(X, Y)
    Initialize w^(0) and set t = 0
    repeat
        for i := 1 to N do
            h*_i := arg max_h w^(t) · Φ(x_i, h, y_i)
        Convex Step:
        Given the best assignment of latent variables,
        Solve for w (via cutting plane algorithm)
    until some stopping condition is met
    return w^(t)
```

Figure: CCCP Algorithm

Convex-concave Procedure (CCCP)

Training Objective

$$\min_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||_2^2 + C \max_{\mathbf{y}'_1,\dots,\mathbf{y}'_N} \Big\{ \Delta\Big((\mathbf{y}_1,..,\mathbf{y}_N),(\mathbf{y}'_1,..,\mathbf{y}'_N)\Big) + \sum_{i=1}^{N} \max_{\mathbf{h}} \mathbf{w}\cdot\Phi(\mathbf{x}_i,\mathbf{h},\mathbf{y}'_i) - \sum_{i=1}^{N} \max_{\mathbf{h}} \mathbf{w}\cdot\Phi(\mathbf{x}_i,\mathbf{h},\mathbf{y}_i) \Big\}$$

Algorithm 1 The Training Algorithm

```
procedure OPT-LATENTSVM(X, Y)
    Initialize w^(0) and set t = 0
    repeat
        for i := 1 to N do
            h*_i := arg max_h w^(t) · Φ(x_i, h, y_i)
        // Optimizing Eq 12
        w^(t+1) := optSVM(X, H*, Y)
        t := t + 1
    until some stopping condition is met
    return w^(t)
```

Figure: CCCP Algorithm

Convex Step: Loss Augmented Inference

- Convex step (via cutting plane) to find the best **w**

Convex Step: Loss Augmented Inference

- Convex step (via cutting plane) to find the best $\mathbf{w}$
- Involves solving the following "loss-augmented inference"

$$\max_{\mathbf{y}'_1,..,\mathbf{y}'_N} \quad \Delta\Big((\mathbf{y}_1, .., \mathbf{y}_N), (\mathbf{y}'_1, .., \mathbf{y}'_N)\Big) + \sum_{i=1}^{N} \max_{\mathbf{h}} \mathbf{w} \cdot \Phi(\mathbf{x}_i, \mathbf{h}, \mathbf{y}'_i)$$

Convex Step: Loss Augmented Inference

- Convex step (via cutting plane) to find the best **w**
- Involves solving the following "loss-augmented inference"

$$\max_{\mathbf{y}'_1,..,\mathbf{y}'_N} \quad \Delta\Big((\mathbf{y}_1,..,\mathbf{y}_N),(\mathbf{y}'_1,..,\mathbf{y}'_N)\Big) + \sum_{i=1}^{N} \max_{\mathbf{h}} \mathbf{w}\cdot\Phi(\mathbf{x}_i,\mathbf{h},\mathbf{y}'_i)$$

- We employ dual decomposition to decouple the two terms and efficiently find an approximate solution

Dual Decomposition

$$\max_{\mathbf{y}'_1,..,\mathbf{y}'_N} \quad \Delta\Big((\mathbf{y}_1, .., \mathbf{y}_N), (\mathbf{y}'_1, .., \mathbf{y}'_N)\Big) + \sum_{i=1}^{N} \max_{\mathbf{h}} \mathbf{w} \cdot \Phi(\mathbf{x}_i, \mathbf{h}, \mathbf{y}'_i)$$

Dual Decomposition

$$\max_{\mathbf{y}'_1,..,\mathbf{y}'_N} \quad \Delta\Big((\mathbf{y}_1,..,\mathbf{y}_N),(\mathbf{y}'_1,..,\mathbf{y}'_N)\Big) + \sum_{i=1}^{N} \max_{\mathbf{h}} \mathbf{w}\cdot\Phi(\mathbf{x}_i,\mathbf{h},\mathbf{y}'_i)$$

$$\max_{\mathbf{y}'_1,\ldots,\mathbf{y}'_N,\mathbf{y}''_1,\ldots,\mathbf{y}''_N} \Delta\Big((\mathbf{y}_1,\ldots,\mathbf{y}_N),(\mathbf{y}'_1,\ldots,\mathbf{y}'_N)\Big) + \sum_{i=1}^{N} \max_{\mathbf{h}} \mathbf{w}\cdot\Phi(\mathbf{x}_i,\mathbf{h},\mathbf{y}_i{}'')$$

subject to

$$\forall i \in \{1,\ldots,N\}, \forall \ell \in \{1,\ldots,L\}, \quad y'_{i,\ell} = y''_{i,\ell}$$

Two Independent Sub-problems

Dual Decomposition (cont.)

- After forming lagrangian, the dual objective function is derived as:

$$L(\mathbf{\Lambda}) := \max_{\mathbf{Y}'} \Delta(\mathbf{Y}, \mathbf{Y}') + \sum_i \sum_\ell \lambda_i(\ell) y'_{i,\ell} +$$

→ *"loss-lagrangian"*

$$\max_{\mathbf{Y}''} \sum_{i=1}^{N} \max_{\mathbf{h}} \mathbf{w} \cdot \Phi(\mathbf{x}_i, \mathbf{h}, \mathbf{y}_i'') - \sum_i \sum_\ell \lambda_i(\ell) y''_{i,\ell}$$

→ *"model-lagrangian"*

Dual Decomposition (cont.)

- After forming lagrangian, the dual objective function is derived as:

$$L(\boldsymbol{\Lambda}) := \max_{\mathbf{Y}'} \Delta(\mathbf{Y}, \mathbf{Y}') + \sum_i \sum_\ell \lambda_i(\ell) y'_{i,\ell} + \quad \longrightarrow \text{“loss-lagrangian”}$$

$$\max_{\mathbf{Y}''} \sum_{i=1}^{N} \max_{\mathbf{h}} \mathbf{w} \cdot \Phi(\mathbf{x}_i, \mathbf{h}, \mathbf{y}_i'') - \sum_i \sum_\ell \lambda_i(\ell) y''_{i,\ell} \quad \longrightarrow \text{“model-lagrangian”}$$

- Since $L(\boldsymbol{\Lambda})$ is an upper-bound on the original loss-augmented inference, we find the tightest upper-bound as an approximate solution: $\min_{\boldsymbol{\Lambda}} L(\boldsymbol{\Lambda})$
- This is solved via sub-gradient descent method

Optimization of the Dual - Multivariate Loss

$$\mathbf{Y}'_* := \arg\max_{\mathbf{Y}'} \Delta(\mathbf{Y}, \mathbf{Y}') + \sum_i \sum_\ell \lambda_i^{(t-1)}(\ell) y'_{i,\ell}$$

Optimization of the Dual - Multivariate Loss

$$\mathbf{Y}'_* := \arg\max_{\mathbf{Y}'} \Delta(\mathbf{Y}, \mathbf{Y}') + \sum_i \sum_\ell \lambda_i^{(t-1)}(\ell) y'_{i,\ell}$$

- Optimizing the multivariate loss is also a hard problem since we have search over entire space of $\mathbf{Y}' \in \mathcal{Y}$ (exponential)

Optimization of the Dual - Multivariate Loss

$$\mathbf{Y}'_* := \arg\max_{\mathbf{Y}'} \Delta(\mathbf{Y}, \mathbf{Y}') + \sum_i \sum_\ell \lambda_i^{(t-1)}(\ell) y'_{i,\ell}$$

- Optimizing the multivariate loss is also a hard problem since we have search over entire space of $\mathbf{Y}' \in \mathcal{Y}$ (exponential)
- However, loss term can be expressed in terms of aggregate statistics over $\mathbf{Y}'$ (false positives (FPs) and false negatives (FNs))

Optimization of the Dual - Multivariate Loss

$$\mathbf{Y}'_* := \arg\max_{\mathbf{Y}'} \Delta(\mathbf{Y}, \mathbf{Y}') + \sum_i \sum_\ell \lambda_i^{(t-1)}(\ell) y'_{i,\ell}$$

- Optimizing the multivariate loss is also a hard problem since we have search over entire space of $\mathbf{Y}' \in \mathcal{Y}$ (exponential)
- However, loss term can be expressed in terms of aggregate statistics over $\mathbf{Y}'$ (false positives (FPs) and false negatives (FNs))
- Since FPs and FNs are integral it can take finite values which can be represented on a two-dimensional grid and efficiently searched via a local search algorithm

Local Search Algorithm

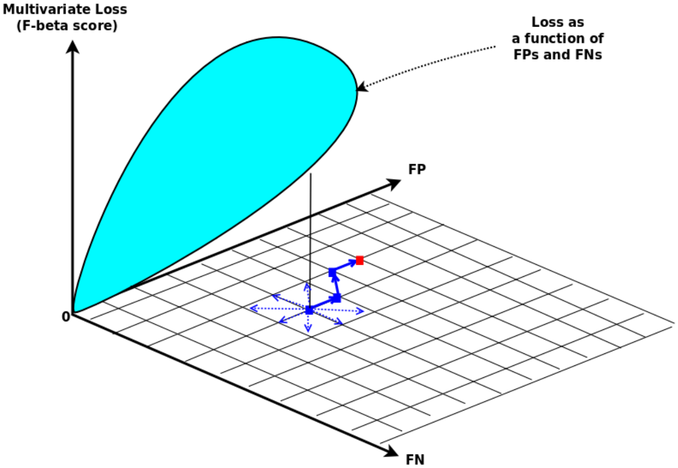


Figure: Local Search Algorithm : An Illustration

Optimization of the Dual - Model Lagrangian

$$\mathbf{Y}''_* := \arg\max_{\mathbf{Y}''} \sum_{i=1}^{N} \max_{\mathbf{h}} \mathbf{w} \cdot \Phi(\mathbf{x}_i, \mathbf{h}, \mathbf{y}_i'') - \sum_i \sum_\ell \lambda_i^{(t-1)}(\ell) y''_{i,\ell}$$

Optimization of the Dual - Model Lagrangian

$$\mathbf{Y}''_* := \arg\max_{\mathbf{Y}''} \sum_{i=1}^{N} \max_{\mathbf{h}} \mathbf{w} \cdot \Phi(\mathbf{x}_i, \mathbf{h}, \mathbf{y}_i'') - \sum_{i} \sum_{\ell} \lambda_i^{(t-1)}(\ell) y''_{i,\ell}$$

- This problem is as difficult as the MAP inference in the underlying graphical model (NP-hard for loopy graphs)

Optimization of the Dual - Model Lagrangian

$$\mathbf{Y}''_* := \arg\max_{\mathbf{Y}''} \sum_{i=1}^{N} \max_{\mathbf{h}} \mathbf{w} \cdot \Phi(\mathbf{x}_i, \mathbf{h}, \mathbf{y}_i{}'') - \sum_i \sum_\ell \lambda_i^{(t-1)}(\ell) y''_{i,\ell}$$

- This problem is as difficult as the MAP inference in the underlying graphical model (NP-hard for loopy graphs)
- We use ILP formulations (relaxed to LP) to solve this

Outline

Experimental Setup

- Dataset: We used the benchmark dataset created by Riedel et. al. (2010)

Experimental Setup

- Dataset: We used the benchmark dataset created by Riedel et. al. (2010)
- Baseline: Hoffmann et. al.'s (2011) state-of-the-art distantly supervised relation extractor

Experimental Setup

- Dataset: We used the benchmark dataset created by Riedel et. al. (2010)
- Baseline: Hoffmann et. al.'s (2011) state-of-the-art distantly supervised relation extractor
- Our approaches:
  - Max-margin which optimizes simple decomposable (Hamming) loss
  - Max-margin which optimizes non-decomposable $-F_\beta$ loss

$$F_\beta := \frac{1}{\frac{\beta}{\text{Precision}} + \frac{1-\beta}{\text{Recall}}}$$

Training on sub-samples of data

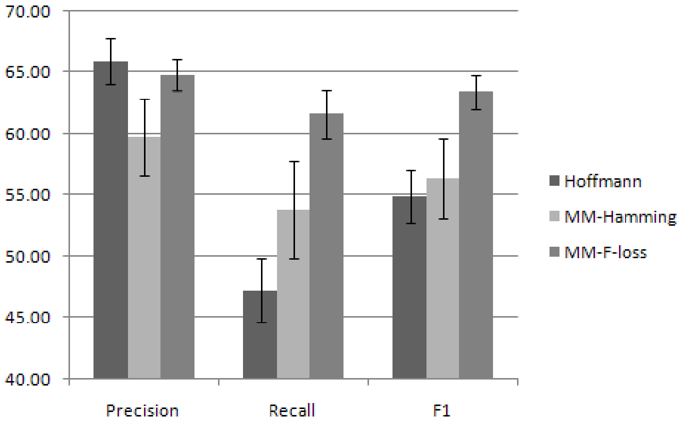


Figure: Experiments on 10% Riedel datasets.

Tuning towards Precision/Recall

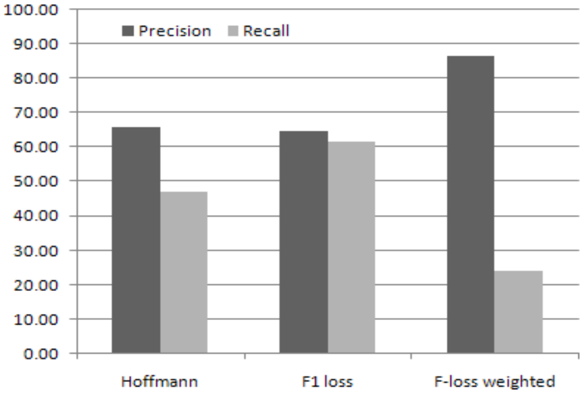


Figure: Weighting of Precision and Recall ($\beta = 0.833$)

Accuracies on the entire dataset

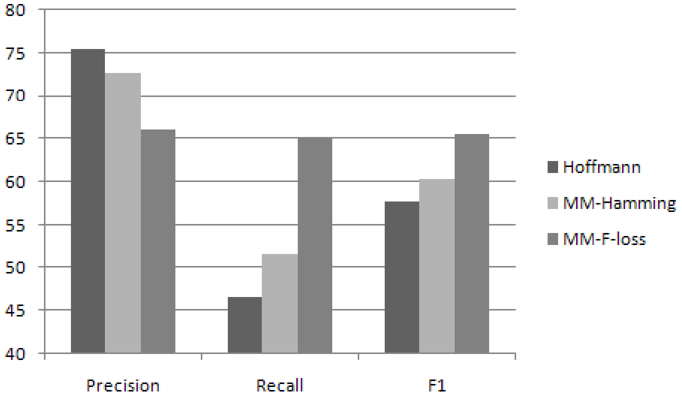


Figure: Overall accuracies Riedel dataset

Outline

Conclusion

- Described a novel max-margin approach to optimize non-linear performance measures, such as F_β, in distant supervision of information extraction models

Conclusion

- Described a novel max-margin approach to optimize non-linear performance measures, such as F_β, in distant supervision of information extraction models
- Our approach is general and can be applied to other latent variable models in NLP

Conclusion

- Described a novel max-margin approach to optimize non-linear performance measures, such as F_β, in distant supervision of information extraction models
- Our approach is general and can be applied to other latent variable models in NLP
- Our approach involves solving the hard-optimization problem in learning by interleaving Concave-Convex Procedure with dual decomposition

Conclusion

- Under several conditions, we have shown our technique outperforms very strong baselines, and results in up to 8.5% improvement in F_1-score

Conclusion

- Under several conditions, we have shown our technique outperforms very strong baselines, and results in up to 8.5% improvement in F_1-score
- For future work:
  - Maximize other performance measures, such as area under the curve, for information extraction models
  - Explore our approach for other latent variable models in NLP, such as those in machine translation

Acknowledgements

- National ICT Australia (**NICTA**) for their generous funding, as part of the Machine Learning Collaborative Research Projects.
- Xerox Research Centre India (**XRCI**) for their generous student travel grant.

Conclusion

Thank You!

Code present at :

https://github.com/ajaynagesh/lsvm_relationextraction

References

- 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
- T. Joachims. 2005. A support vector method for multi-variate performance measures. In International Conference on Machine Learning (ICML)
- Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In Proceedings 47th Annual Meeting of the ACL
- 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

- Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. 2012. Multi-instance multi-label learning for relation extraction. In the proceedings of EMNLP-CoNLL
- Edouard Grave. 2014. A convex relaxation for weakly supervised relation extraction. Proceedings of EMNLP.
- Chun-Nam John Yu and Thorsten Joachims. 2009. Learning structural svms with latent variables. In Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009
- Mani Ranjbar, Tian Lan, Yang Wang, Stephen N. Robinovitch, Ze-Nian Li, and Greg Mori. 2013. Optimizing nondecomposable loss functions in structured prediction. IEEE Trans. Pattern Anal. Mach. Intell.
- Alexander M. Rush and Michael Collins. 2012. A tutorial on dual decomposition and lagrangian relaxation for inference in natural language processing. J. Artif. Intell. Res. (JAIR)
- I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. 2004. Support vector machine learning for interdependent and structured output spaces. In International Conference on Machine Learning (ICML)