Sequence Learning from Data with Multiple Labels

Mark Dredze (Johns Hopkins Univ., USA)

Partha Pratim Talukdar (Univ. of Penn., USA)

Koby Crammer (Technion, Israel)

Penn
University of Pennsylvania

Motivation

- Labeled data is expensive
- Multiple cheap but noisy annotations may be available (e.g. Amazon Mechanical Turk)
  - The problem: **Adjudication**!
- Can we learn from multiple labels without adjudication?

Penn
University of Pennsylvania

Learning Setting

- Input:
  - Feature sequence (sentence)
  - *Set* of initial priors over labels at each position

John Blitzer studies at the University of Pennsylvania .

PER/0.7 PER/0.7 O/1.0 O/1.0 O/1.0 ORG/1.0 ORG/1.0 ORG/0.3 O/1.0
O/0.1 O/0.1 LOC/0.7
ORG/0.1
LOC/0.1 LOC/0.1

- Output: Trained sequence labeler (e.g. CRF)
  - Take label priors into account during training

Why Multiple Labels?

- Easy to encode guesses as to correct label
  - Users provide labels
  - Allows multiple conflicting labels
    - Don't need to resolve conflicts (saves time)

Comparison with Canonical Multi-Label Learning

Canonical Multi-Label

1. Multiple labels per instance during training
2. Each instance can have multiple valid labels

This Paper

1. Same, but only one of the labels is correct
2. Only one valid label per instance

Previous Work

- Jin and Ghahramani, NIPS 2003
  - Classification setting (simple output)
- This paper
  - Structured Prediction (complex output)

Generality of the Learning Setting

- Multi-Label setting encodes standard learning settings
  - Unsupervised
    - uniform prior over labels
  - Supervised
    - per-position prior of 1.0
  - Semi-supervised
    - combination of above

Learning with Multiple Labels

- Two learning goals
  - Find a model that best describes the data
  - Respect per-position input prior over labels, as much as possible
- Balance these two goals in a single objective function

Multi-CRF

CRF

$$P(\mathbf{y}|\mathbf{x},\theta) \propto \exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K}\theta_k f_k(y_{t-1}, y_t, \mathbf{x})\right)$$

Multi-CRF Objective

$$\min\ l(\theta) = \sum_{i=1}^{N}\sum_{\mathbf{y}\in S^{(i)}}\left[\mathrm{KL}(\hat{P}(\mathbf{y}|\mathbf{x}^{(i)}), \pi_{\mathbf{y}}^{(i)}) - \hat{P}(\mathbf{y}|\mathbf{x}^{(i)})\log P(\mathbf{y}|\mathbf{x}^{(i)},\theta)\right]$$

Estimated Prior **Initial Prior** **CRF**

Penn
University of Pennsylvania

Multi-EM Algorithm

- M-step
  - Learn a Multi-CRF that models *all* given labels at each position
  - Weigh possible labels by estimated label priors
- E-step
  - Re-estimate label priors based on model and initial prior
  - Balances between CRF's label estimates and the input priors

Experimental Setup

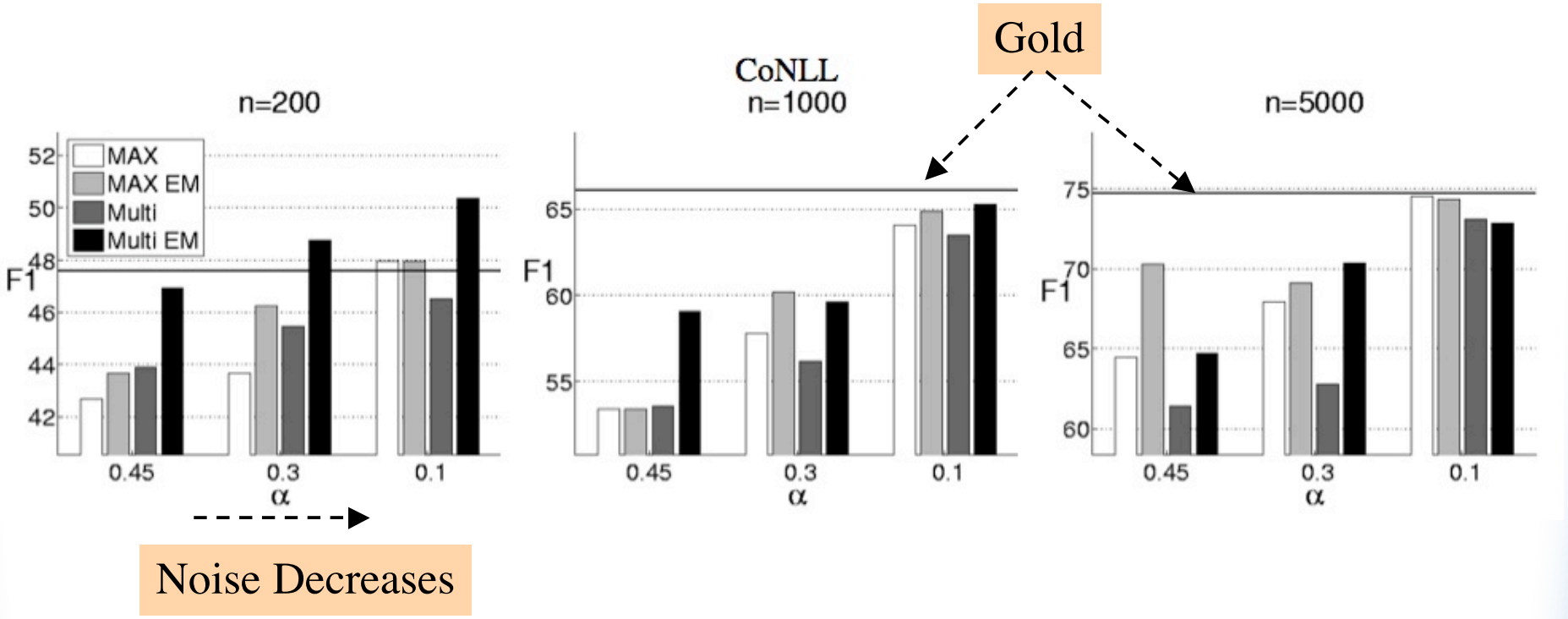
- Dataset
  - CoNLL-2003: Named Entity Dataset with PER, LOC and ORG tags, 3454 test instances
- Each instance has two different sequences
  - Gold labels
  - Labels generated by an HMM
- Noise level: α
  - probability of incorrect sequence getting higher prior (higher is noisier)

Penn
University of Pennsylvania

Variants

- MAX
  - Standard CRF with max prior at each position.
- MAX-EM
  - EM with MAX in M step
- Multi
  - Multi-CRF
- Multi-EM
  - EM with Multi-CRF in M step

Results on CoNLL Data



Multi-EM most effective on noisier data, especially when less supervision is available.

Penn
University of Pennsylvania

When is Learning Successful?

- Effective over single-label learning with
  - Small amount training data (low quantity)
  - Lots of noise (low quality)
- Additional label may add information in this setting.

Penn
University of Pennsylvania

Conclusion

- Presented novel models for learning structured predictors from multi-labeled data, in presence of noise.
- Experimental results on real world data
- Analyzed when learning in such setting is effective.

Penn
University of Pennsylvania

Thanks!

Penn
University of Pennsylvania