Efficiently Maintaining Conditional Random Fields

Xixuan (Aaron) Feng

Guangde Chen

One-Slide Summary

- Statistical Information Extraction (IE) is important in both academic and industrial applications
  - To build knowledge bases
  - To improve search, mining, etc.
  - E.g.: isWiki, MSR Academic Search, Ali Baba, YAGO

- **Problem: Corpus is evolving!**
  - A single new training example forces us to train the model from scratch and so is too slow to apply to real applications
- **Goal: To incorporate new examples efficiently**
  - We focus on the most poplar statistical model for IE – conditional random fields (CRFs)
- **Result: Evaluation over a text chunking dataset**

Baselines

- **Two extremes**
  - Retrain-All: to train the model from scratch as every new example comes in
  - Retrain-New: only to use new examples

Retrain-All

- $\mathbf{1}.\ \boldsymbol{w} \leftarrow \mathbf{0}$
- $\mathbf{2}.\ \boldsymbol{w} \leftarrow \boldsymbol{w} - \sum \nabla \boldsymbol{l_i}(\boldsymbol{w})$
- $\mathbf{3}.$ If $\boldsymbol{w}$ is not converged GOTO 2

Retrain-New

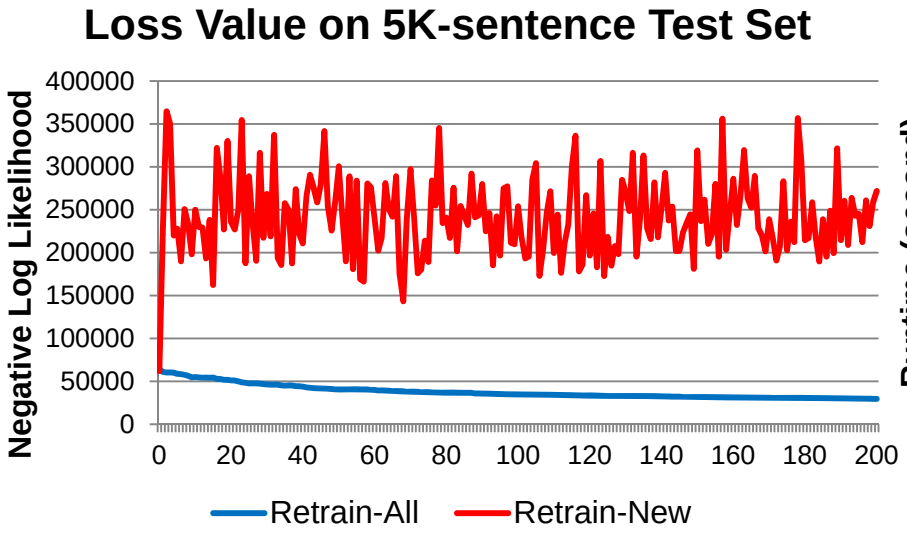
- $\mathbf{1}.\ \boldsymbol{w} \leftarrow \mathbf{0}$
- $\mathbf{2}.\ \boldsymbol{w} \leftarrow \boldsymbol{w} - \nabla \boldsymbol{l_{new}}(\boldsymbol{w})$
- $\mathbf{3}.$ If $\boldsymbol{w}$ is not converged GOTO 2

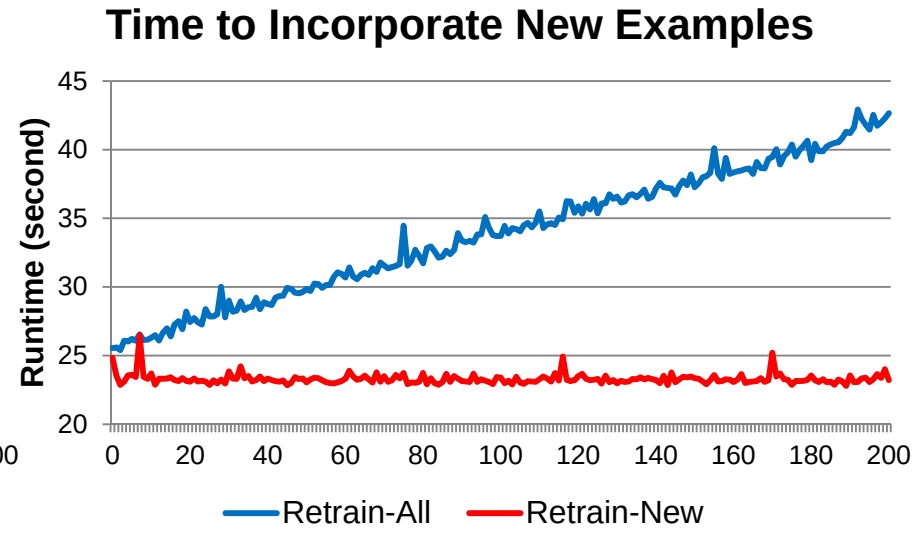
Experiment Setup

- **Dataset**
  - CoNLL 2000 Shared Tasks
    http://www.cnts.ua.ac.be/conll2000/chunking/
  - Wall Street Journal
  - ~9K sentences = 5K (test set) + ~4K (training pool)
- **Task: Text Chunking**
- **Features: 7.4 million**
  - Word-based regular expressions
  - Part-of-speech
- **20 initial examples + 200 examples streaming in**
  - From training pool
  - Evaluate quality using 5K-sentence test set

Baselines

- **Two extremes**
  - Retrain-All: to train the model from scratch as every new example comes in
  - Retrain-New: only to use new examples





Increment Instead of Retrain

- **Keep the model as initial guess when new examples come**
  - Incr-All and Incr-New
  - We expected Incr-New gives bad quality results

Incr-All

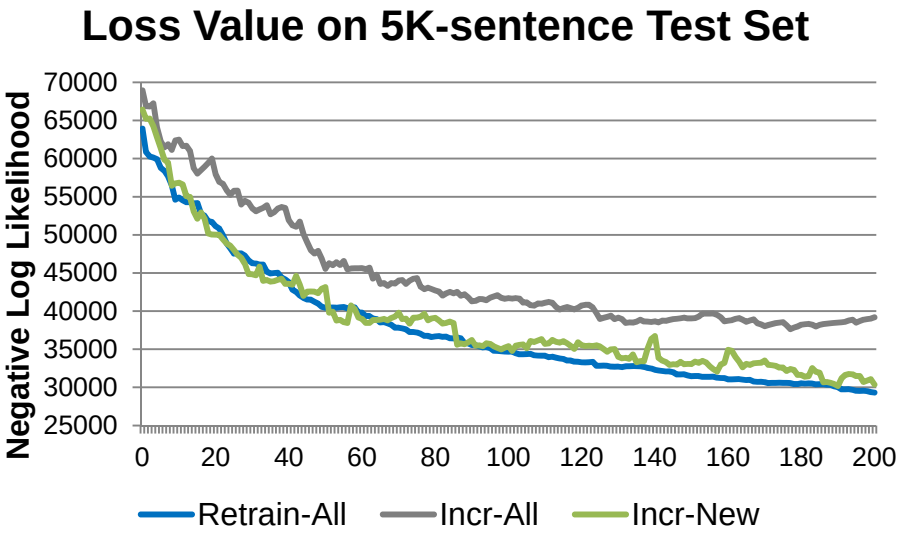
- $\mathbf{1.}\ \boldsymbol{w} \leftarrow \boldsymbol{w}_{previous}$
- $\mathbf{2.}\ \boldsymbol{w} \leftarrow \boldsymbol{w} - \sum \nabla \boldsymbol{l}_{\boldsymbol{i}}(\boldsymbol{w})$
- $\mathbf{3.}$ If $\boldsymbol{w}$ is not converged GOTO 2

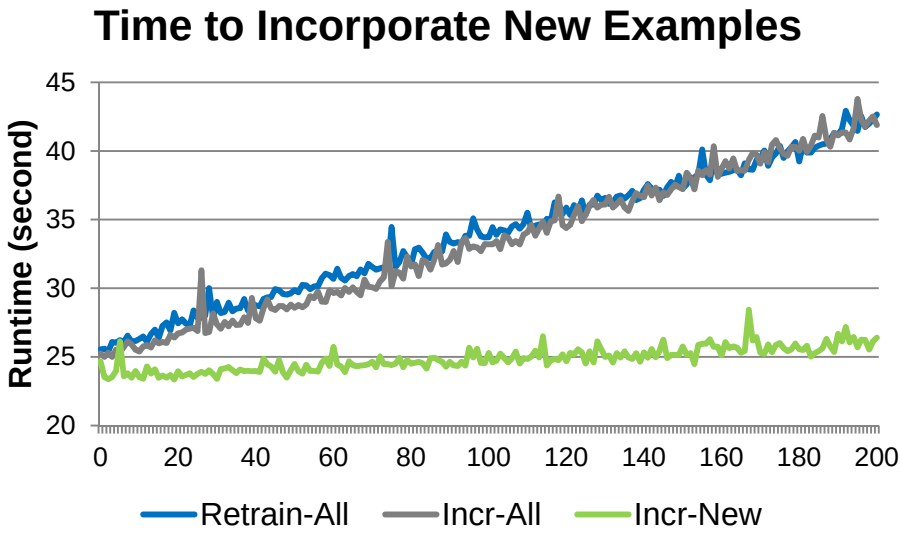
Incr-New

- $\mathbf{1.}\ \boldsymbol{w} \leftarrow \boldsymbol{w}_{previous}$
- $\mathbf{2.}\ \boldsymbol{w} \leftarrow \boldsymbol{w} - \nabla \boldsymbol{l}_{\boldsymbol{new}}(\boldsymbol{w})$
- $\mathbf{3.}$ If $\boldsymbol{w}$ is not converged GOTO 2

Increment Instead of Retrain

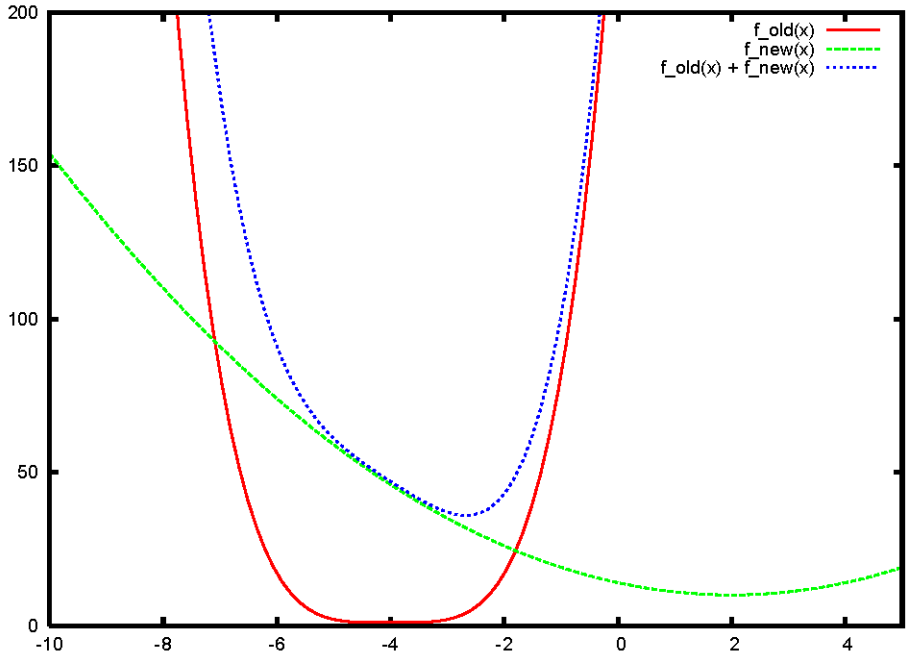
- **Keep the model as initial guess when new examples come**
  - We expected Incr-New gives bad quality results
  - **But we were wrong!**





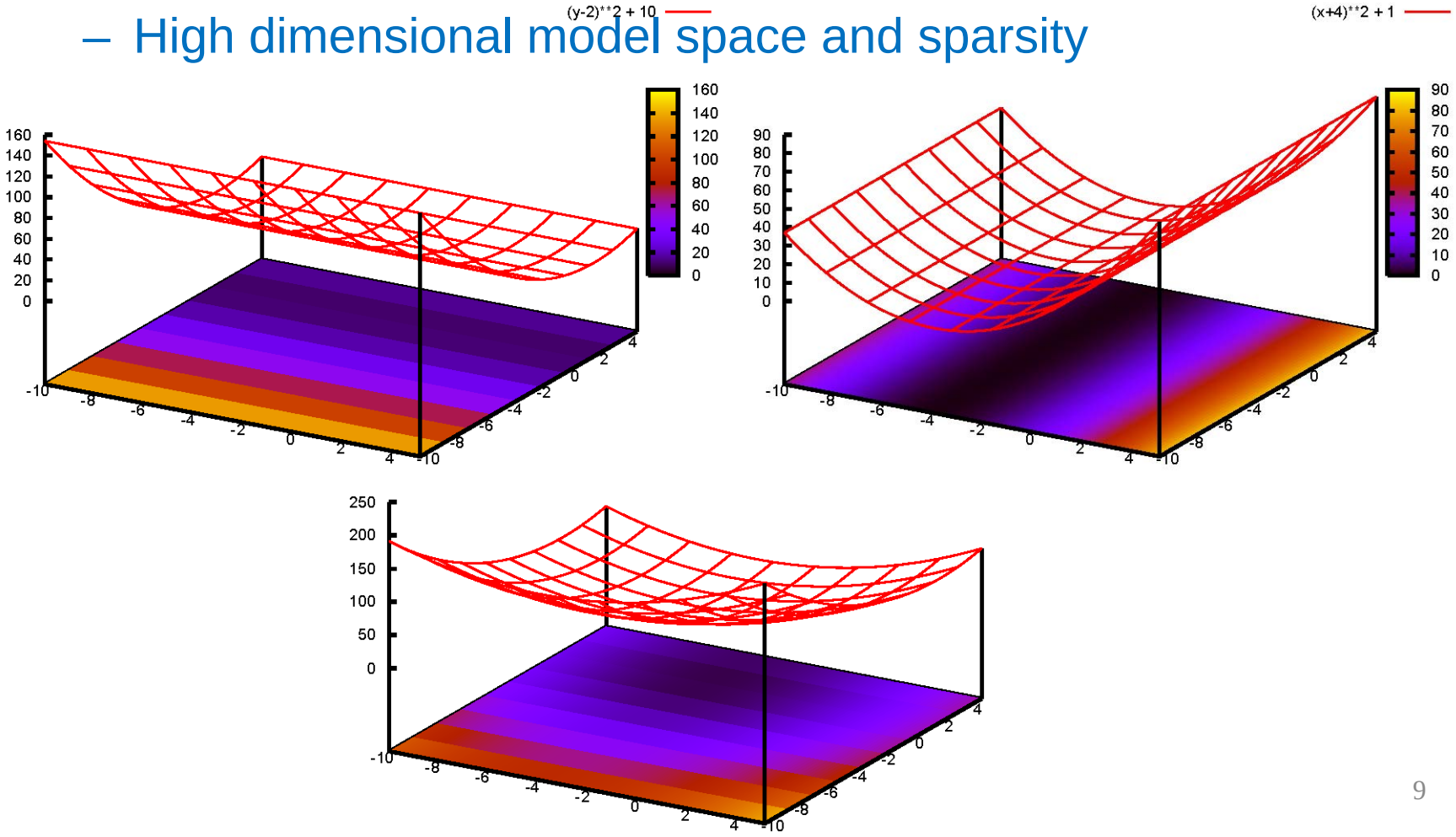
Analysis

- **Incr-New is not always good**
  - In general, gradient descent is robust to initial value



Analysis (Cont.)

- **Cases that Incr-New is good**
  - High dimensional model space and sparsity



Conclusion and Future Work

- We have evaluated and analyzed several strategies to incorporate new examples for CRF
  - Retrain-All, Retrain-New, Incr-All, Incr-New
- Next Steps
  - Evaluation with regularization
  - Formal proof of cases that Incr-New are good
  - Other data sets
- Thanks & Questions