Improving Context-Aware Query Classification via Adaptive Self-Training

Minmin Chen[1], Jian-Tao Sun[2], Xiaochuan Ni[2], Yixin Chen[1]

1. Washington University, Saint Louis, MO
2. Microsoft Research Asia

Background

- Query classification: classify query to predefined categories
  - E.g., query: “cikm 2011” → category: “Computer Science Conferences”
- Challenges
  - Lack of training data (human labeling is expensive & time consuming)
  - Sparseness of query features (e.g., query: “up” → category: “movie”)
    - Query enrichment (Shen et. al 2006, Broder et.al 2007)

Context-aware query classification

- Context is users action within one session
  - E.g., submitting query, viewing result pages, click on URLs

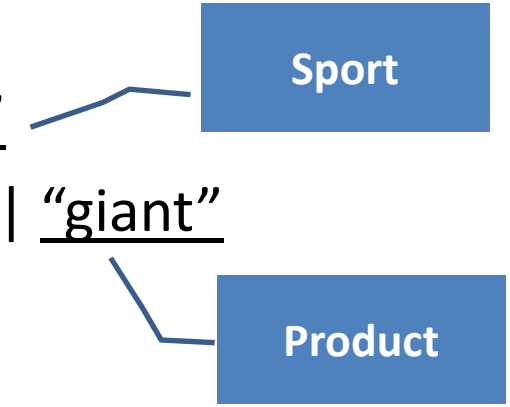
- Search session: a series of interactions by the user toward addressing a single information need (Jansen, 2007)

query = "giant" → category = "**Sport**" ? (*San Francisco baseball team*)

category = "**Product**"? (*Giant bicycle*)

Session A: "Major League Baseball" | "giant"

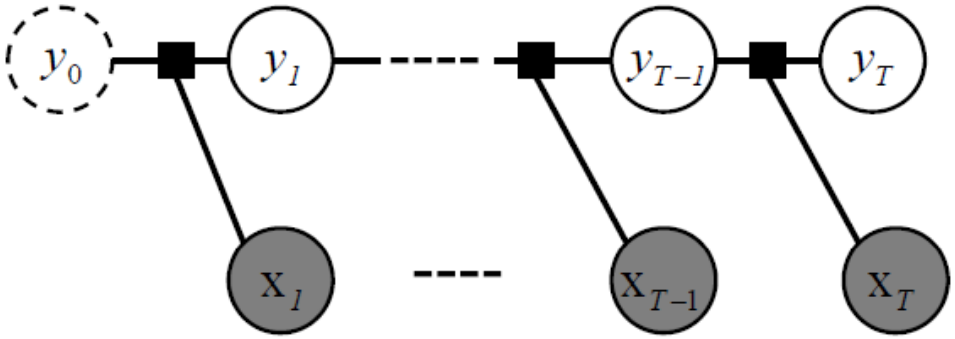
Session B: "bike review" | "mountain bikes" | "giant"



Model search context using Conditional Random Fields (CRF)

- $\mathbf{x} = \langle x_1, x_2, \cdots, x_T \rangle$, observations of query session (queries, clicked URLs, etc.)

  $\mathbf{y} = \langle y_1, y_2, \cdots, y_T \rangle$, category sequence



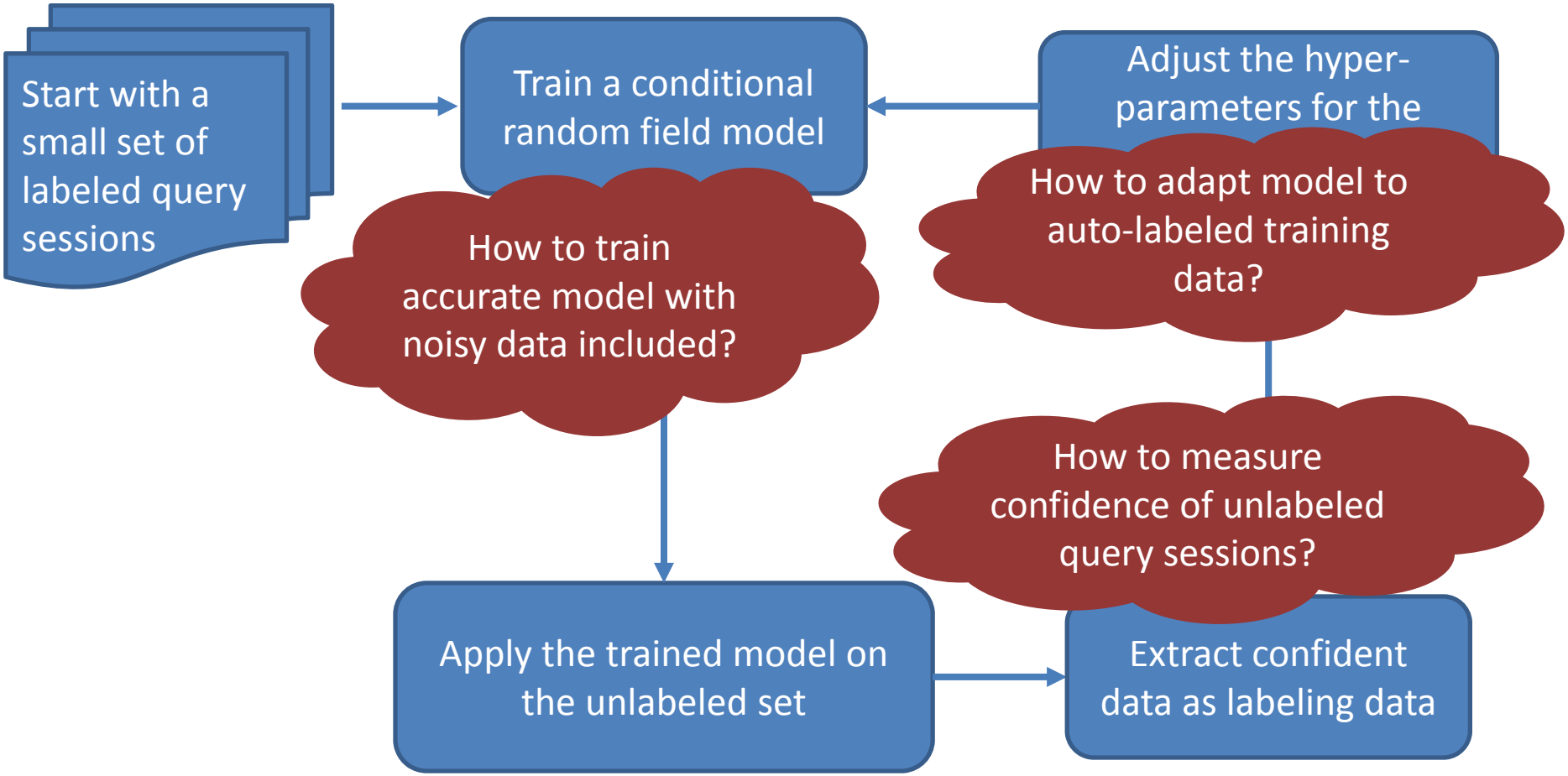
- Maximum likelihood training

$$\max_{\mathbf{w}} \sum_{n=1}^{N} \log p(\mathbf{y}^{(n)}|\mathbf{x}^{(n)}), \quad \mathcal{D} = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N}$$

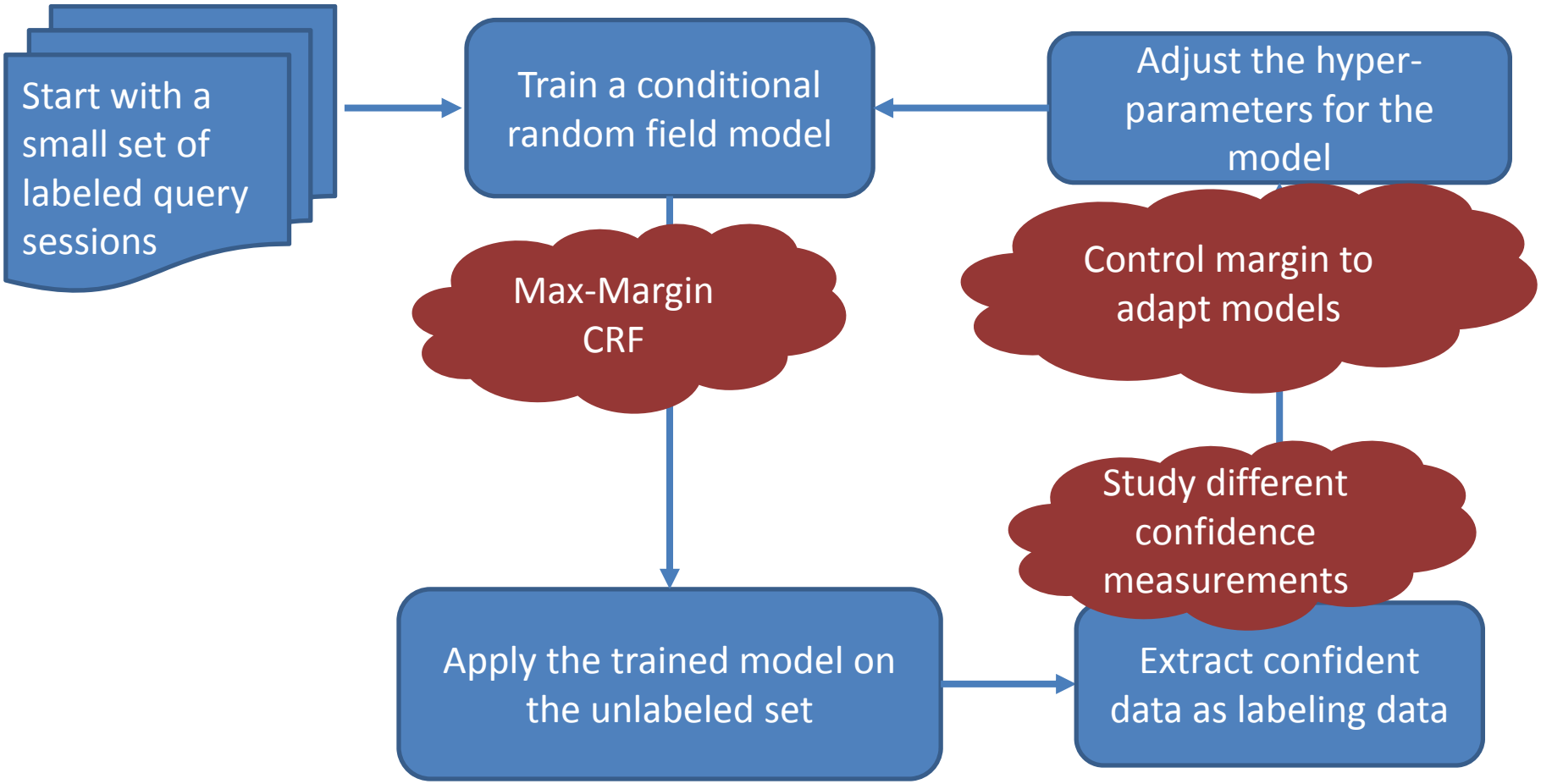
H. Cao, et, al. SIGIR, 2009

Our work: utilize unlabeled query sessions for context-aware query classification

- Combine self-training with Conditional Random Fields



Adaptive self-training CRF (ASCRF)



SoftMax margin Conditional Random Field (SMMCRF)

- Maximum likelihood training: maximizes the likelihood of observing the labeling sequences given input sequences

- Margin maximization training: finds a hyperplane that not only separates the training data, but also maximizes the score difference between the true label and other labels (Taskar 2003, Tsochantaridis 2004, Gimpel 2010)

$$\min_{\mathbf{w}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum \xi_n$$

$$\text{s.t.} \quad -\mathbf{w}^\top \mathbf{f}_{\mathbf{x}^{(n)}}(\mathbf{y}^{(n)}) + \max_{\mathbf{y}} \left(\mathbf{w}^\top \mathbf{f}_{\mathbf{x}^{(n)}}(\mathbf{y}) + \Delta_{\mathbf{y}^{(n)}}(\mathbf{y})\right) \leq \xi_n, \forall n$$

  - More robust to noise than maximum likelihood training
  - Better generalization performance

SMMCRF generalization performance

Empirical risk

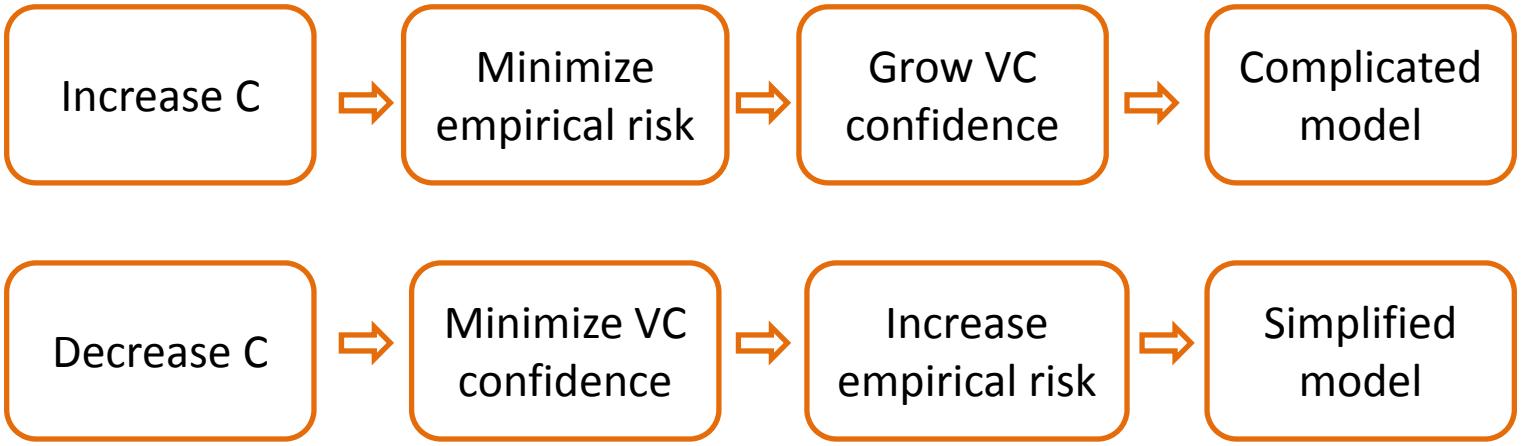
VC confidence

$$R(\mathbf{w}) \leq R_{emp}(\mathbf{w}) + \Omega(N, d, \eta)$$

Expected risk

Trade-off parameter

$$\min_{\mathbf{w}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum \xi_n$$



Model adaption in self-training

- Infeasible to tune C parameter of SMMCRF by validation data
  - Training data is limited
  - Auto-labeled data contains noise
- Dynamically update C in self-training
  - Decrease C as unlabeled query sessions are included as training data
  - Avoid complicated model and over-fitting to noises

Model adjustment in self-training (Cont'd)

- Control the margin according to the noise rate estimated using new auto-labeled data

$$\epsilon \approx \frac{\sum_{num}(1 - p_{\mathbf{w}}(\hat{\mathbf{y}}|\mathbf{x}))}{num}$$

$$C \leftarrow C/(1+\epsilon)$$

- Drop unconfident instances included in previous runs

$$\tau_{\mathbf{w}}(\mathbf{x}, \hat{\mathbf{y}}) = \mathbf{w}^{\top}\mathbf{f}_{\mathbf{x}}(\hat{\mathbf{y}}) - \max_{\mathbf{y}'}\left(\mathbf{w}^{\top}\mathbf{f}_{\mathbf{x}}(\mathbf{y}') + \Delta_{\hat{\mathbf{y}}}(\mathbf{y}')\right) < 0.$$

Confidence measurement

- c1: margin based confidence measurement

$$c_{\mathbf{w}}^{(1)}(\mathbf{x}, \hat{\mathbf{y}}) = \mathbf{w}^\top \mathbf{f}_{\mathbf{x}}(\hat{\mathbf{y}}) - \log \sum_{\mathbf{y}' \neq \hat{\mathbf{y}}} e^{\mathbf{w}^\top \mathbf{f}_{\mathbf{x}}(\mathbf{y}') + \Delta_{\hat{\mathbf{y}}}(\mathbf{y}')}$$

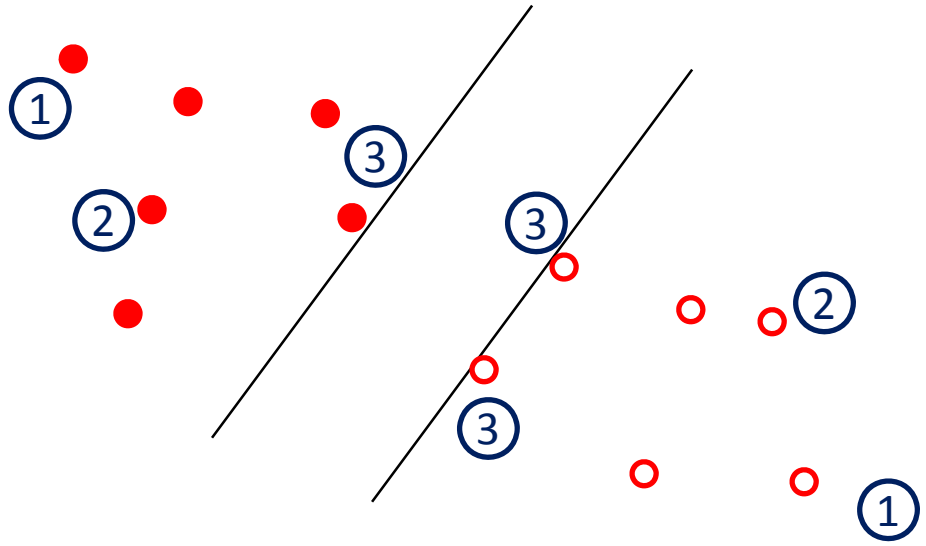
- c2: conditional probability

$$c_{\mathbf{w}}^{(2)}(\mathbf{x}, \hat{\mathbf{y}}) = p(\hat{\mathbf{y}}|\mathbf{x}) = \frac{e^{\mathbf{w}^\top \mathbf{f}_{\mathbf{x}}(\hat{\mathbf{y}})}}{\sum_{\mathbf{y}'} e^{\mathbf{w}^\top \mathbf{f}_{\mathbf{x}}(\mathbf{y}')}}$$

- c3: distance to decision boundary

$$c_{\mathbf{w}}^{(3)}(\mathbf{x}, \hat{\mathbf{y}}) = -c_{\mathbf{w}}^{(1)}(\mathbf{x}, \hat{\mathbf{y}}) \cdot I\left(c_{\mathbf{w}}^{(1)}(\mathbf{x}, \hat{\mathbf{y}}) \geq 0\right)$$

Confidence measurement (Cont'd)



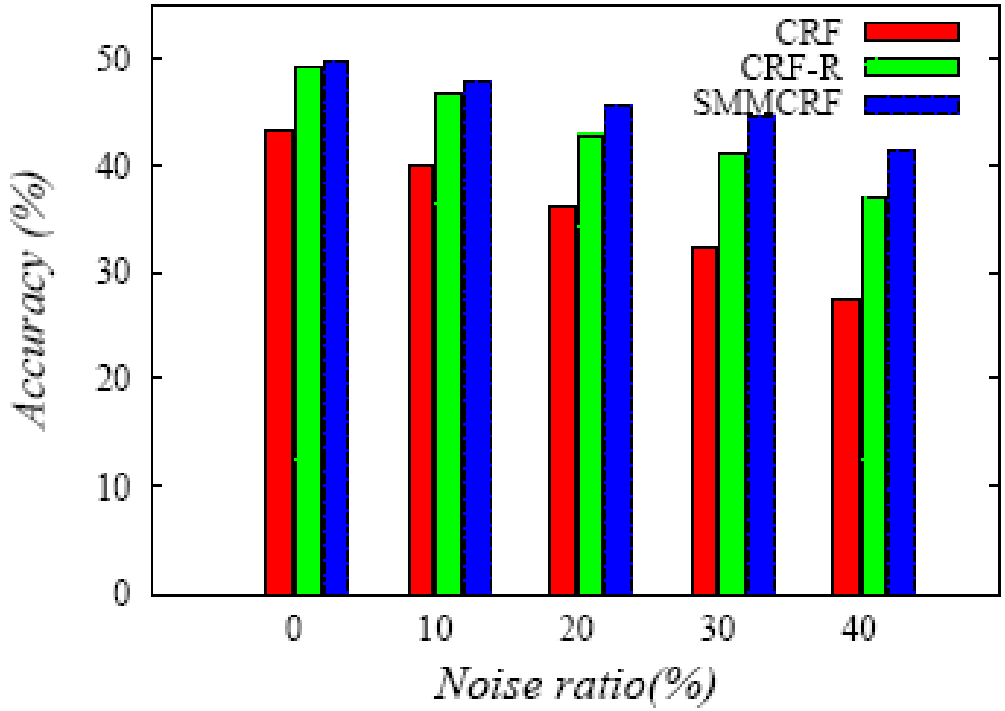
	c1: margin gap	c2: conditional probability	c3: support vector
Noisy	Low	Medium	High
Useful	Low	Medium	High

Experiment datasets

- Dataset 1
  - 3,500 query sessions (8,988 queries)
  - Manually labeled using KDD Cup'05 categories
  - Used by Cao, SIGIR 2009

- Dataset 2
  - 1,727 query sessions (39, 565 queries)
  - Manually labeled by search intent type, e.g., "Compare products, services, or activities for use"

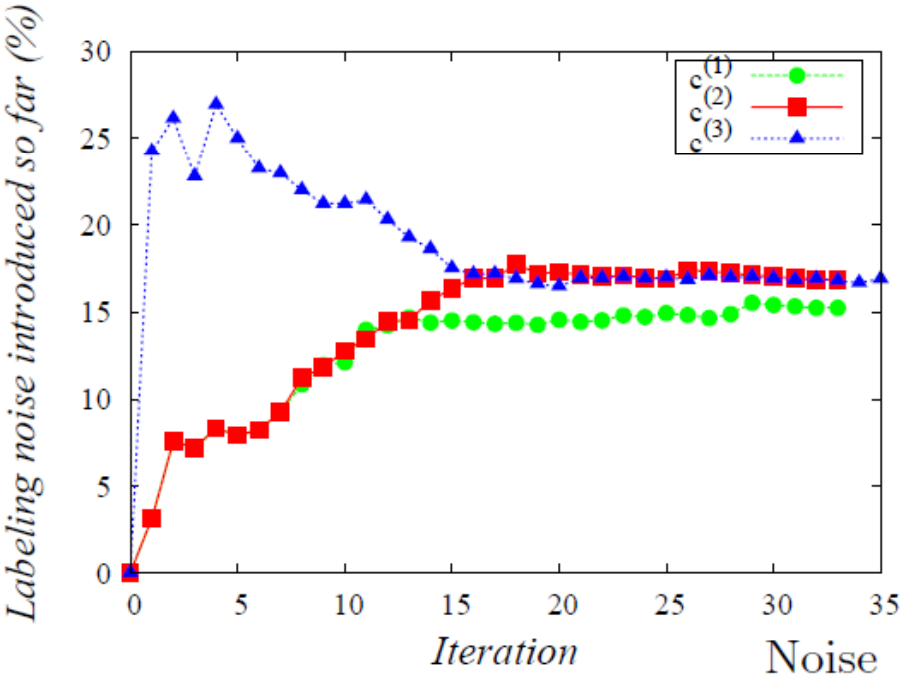
Experiment 1: algorithm robustness to training data noise

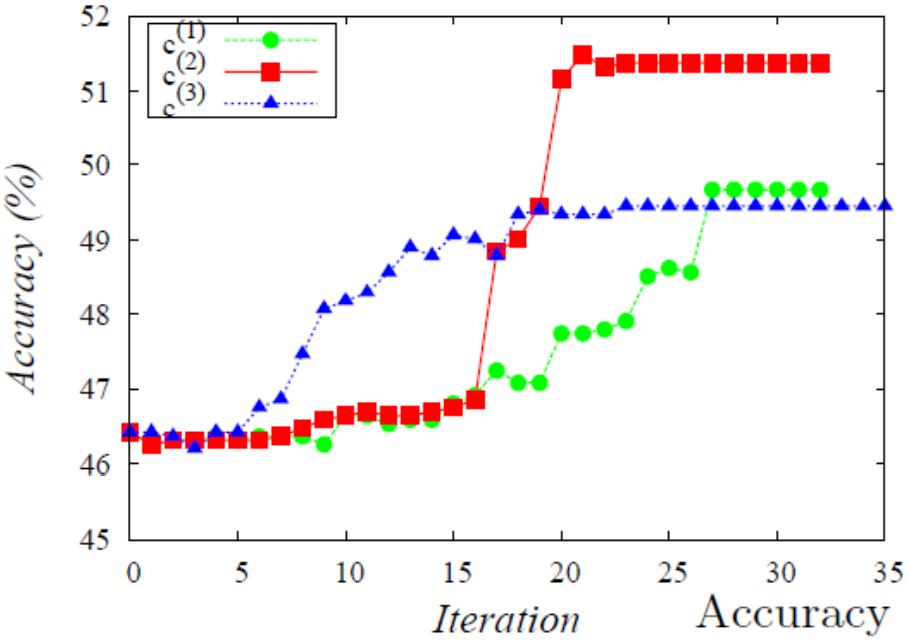


CRF: maximum likelihood
CRF-R: maximum likelihood with regularization
SMMCRF: margin maximization

- Run experiments with Dataset 1
- 80% data used for training, 20% for testing
- Average results of 5 fold experiments

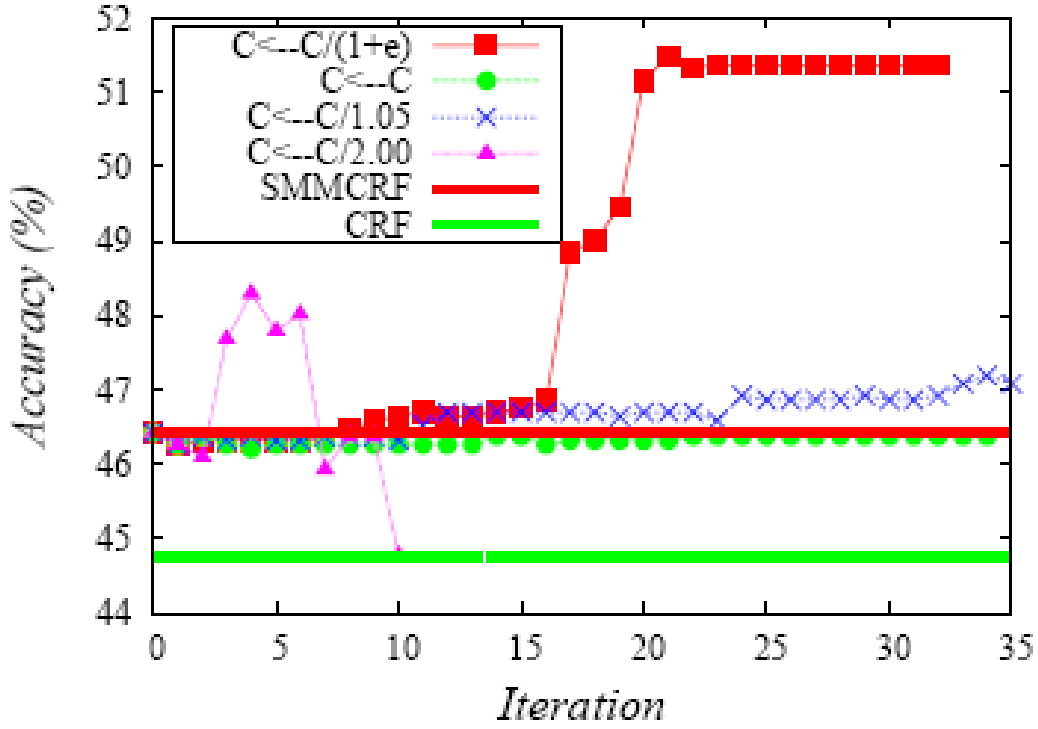
Experiment 2: confidence measurement





- Run experiment on Dataset 1
- Adopt adaptive self-training with SMMCRF
- 80% as training (10% as labeled data, others as unlabeled), 20% for testing

Experiment 3: margin control methods



- C<--C : no change to C value
- C<--C/(1+e): change C according to error rate
- C<--C/1.05, C<--C/2: constant update

Overall Results

Accuracy	Baseline	ASCRF	Improvement
Dataset 1	$44.67\%(1.48 \times 10^{0})$	$49.94\%(8.07 \times 10^{-1})$	11.7%
Dataset 2	$35.25\%(4.95 \times 10^{-2})$	$43.53\%(4.44 \times 10^{-3})$	23.4%

- SMMCRF used as baseline algorithm
- Dataset 1: 80% as training (10% as labeled data, others as unlabeled), 20% for testing
- Dataset 2: 80% as training (5% as labeled data, others as unlabeled), 20% for testing
- Average results of 10 runs experiments

Conclusion and future work

- Conclusion
  - Search context can be used to improve query classification
  - Adaptive self-training is able to utilize unlabeled data
    - Max margin CRF
    - Control margin to adapt model training
    - Effective confidence measurement method

- Future Work
  - Include more information as search contexts to further improve performance

Thanks!