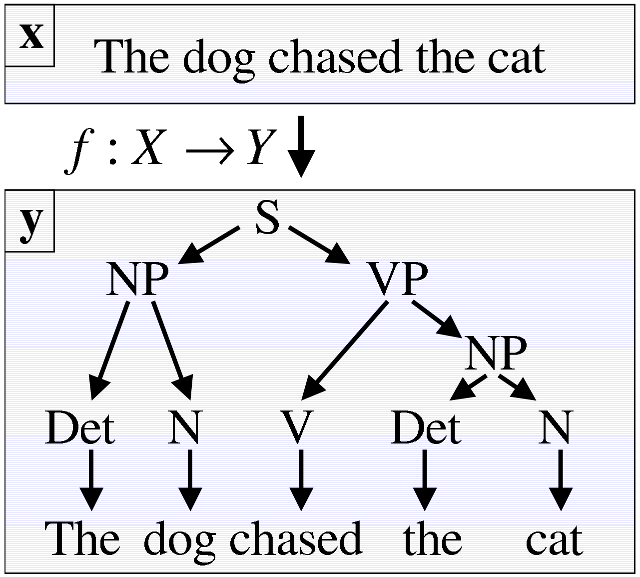
Search-Based Structured Prediction

by Harold C. Daumé III (Utah), John Langford (Yahoo),
and Daniel Marcu (USC)
Submitted to Machine Learning, 2007

Presented by:
Eugene Weinstein, NYU/Courant Institute
October 2nd, 2007

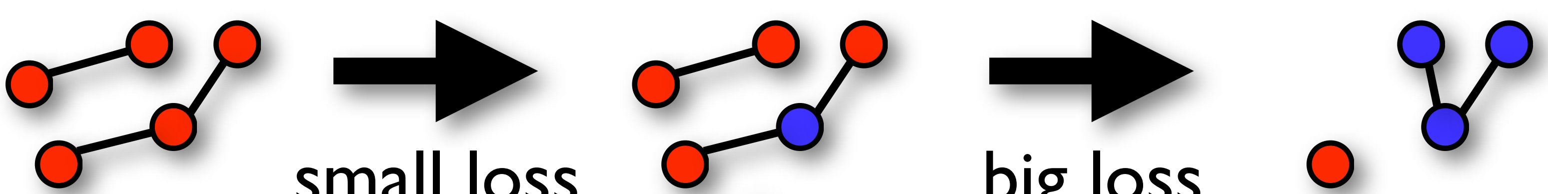
Structured Prediction Intro

- Given: labeled training data $(x_1, y_1), \ldots, (x_m, y_m) \in \mathcal{X} \times \mathcal{Y}$
- Task: learn mapping from inputs $x \in \mathcal{X}$ to outputs $y \in \mathcal{Y}$
- Special cases
    - Binary classification: $\mathcal{Y} = \{-1, 1\}$
    - Multiclass classification: $\mathcal{Y} = \{1, \ldots, k\}$
- Natural language parsing example:



Exploiting Structure

- Naive approach: treat each possible output in $\mathcal{Y}$ as discrete label, apply multiclass classification. But:
  - Enumerating all members of $\mathcal{Y}$ often intractable
  - Cannot model closeness of examples (changing one node of tree vs. changing the entire tree)
- Approach: try to exploit structure and dependencies within the output space
  - Represent closeness of outputs using loss function



SP Overview

- Discriminative structured prediction papers typically extend multiclass classification or regression techniques
- Most classification schemes use SVM-like max-margin linear classifications incorporating loss functions
  - [Taskar, Guestrin, Koller ’03], [Tsochantaridis, Hofmann, Joachims, Altun ’04] [Sha, Saul ’07]
- Regression formulation of SP: [Cortes, Mohri, Weston ’06]
- Searn is a meta-algorithm. Claim: given multiclass classifier achieving good generalization, Searn does the same for SP

Search-based SP

[Daumé '06] [Daumé, Langford, Marcu '07]

- Searn: view structured prediction as search problem
- SP: distribution $\mathcal{D}$ over inputs, output costs (x, c) $|c| = |\mathcal{Y}|$
  - e.g.: x_i is input, c_y is the loss for any y to the true label y_i
- Define loss of cost-sensitive classifier $h : \mathcal{X} \to \mathcal{Y}$ as

$$L(\mathcal{D}, h) = \mathbb{E}_{(x,\boldsymbol{c})\sim\mathcal{D}} \left\{ c_{h(x)} \right\}$$

- View outputs as vectors $y = [y^{(1)}, \ldots, y^{(l)}]$, but classification problems not limited to sequences
- A classifier defines a path through space of input/output pairs, and training process iteratively refines the classifier

Searn Specifics

- We need to provide:
  - Cost-sensitive multiclass learning algorithm
  - Initial classifier
  - Loss function
- Initial classifier should have low training error, but need not generalize well
  - Could be best path from any standard search algorithm
  - Each Searn iteration finds a classifier that is not as good on the training set, but generalizes a little better

Searn Training

- Search state space: (input, partial output): $s = (x, y^{(1)}, \ldots, y^{(l)})$
- Initial classifier: pick next label that minimizes cost, assuming that all future decisions are also optimal:

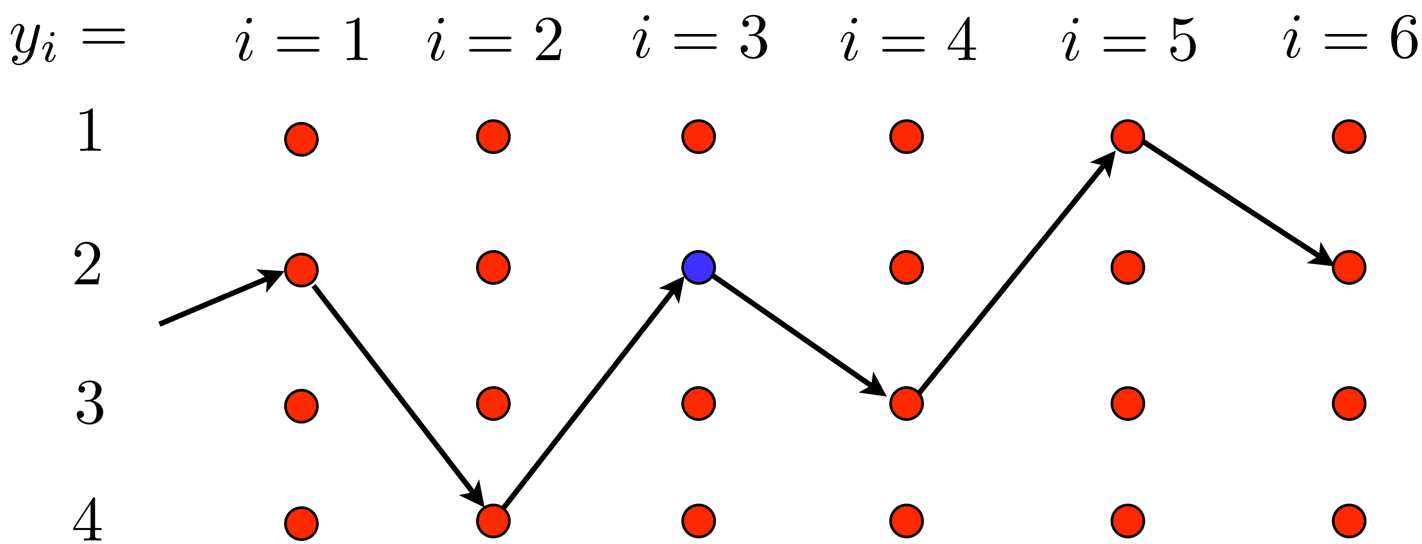
$$h_0(s, c) = \arg\min_{y^{(l+1)}} \min_{y^{(l+2)}, \ldots, y^{(L)}} c_{[(y^{(1)}, \ldots, y^{(L)})]}$$

- Iterative step: use current classifier h to construct a set of examples to train the next classifier; then interpolate
  - For each state, try every possible next output
  - Cost assigned to each output tried is loss difference

$$l_h(c, s, a) = \mathbb{E}_{y \sim (s,a,h)} c_y - \min_{a'} \mathbb{E}_{y \sim (s,a',h)} c_y$$

Searn Training Illustration

$$l_h(c, s, a) = \mathbb{E}_{y \sim (s,a,h)} c_y - \min_{a'} \mathbb{E}_{y \sim (s,a',h)} c_y$$



Prediction of current classifier h

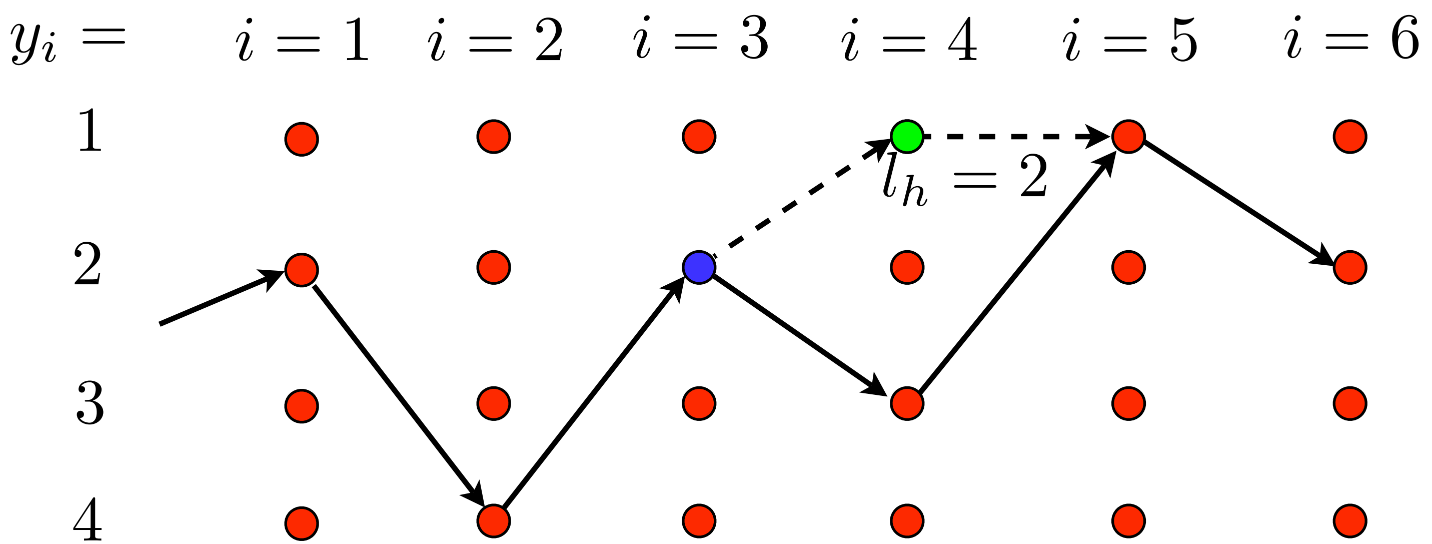
Other path being considered (s, a, h)

Current state s

Potential next state a

Searn Training Illustration

$$l_h(c, s, a) = \mathbb{E}_{y \sim (s,a,h)} c_y - \min_{a'} \mathbb{E}_{y \sim (s,a',h)} c_y$$



$\longrightarrow$ Prediction of current classifier h

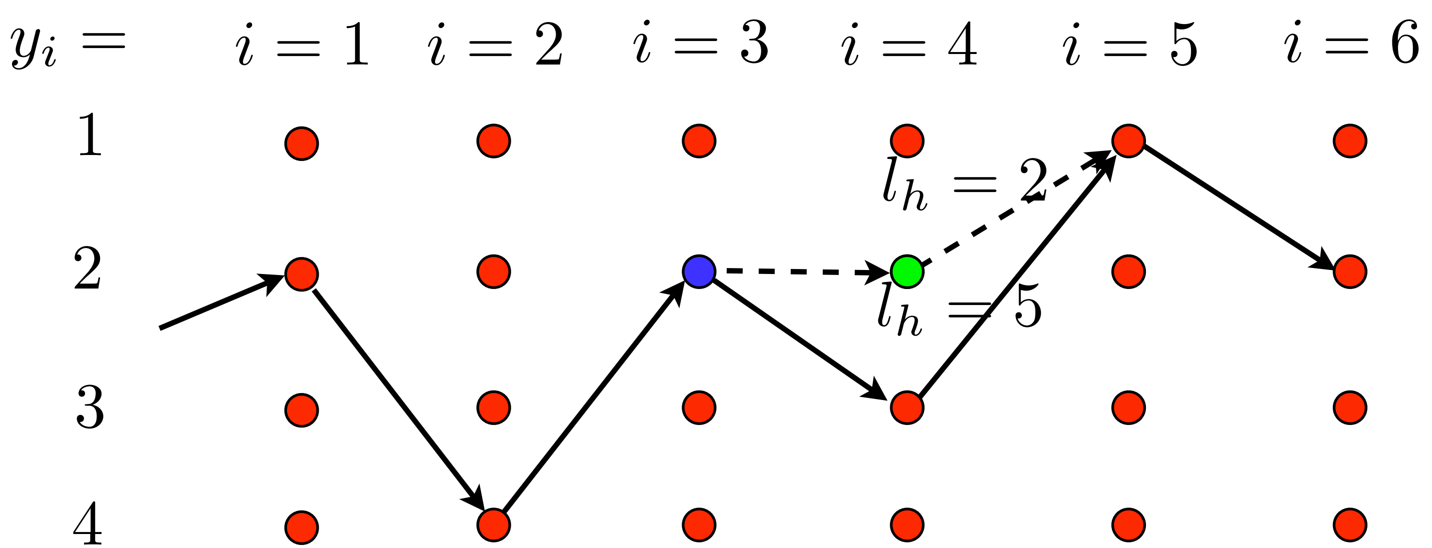
$\dashrightarrow$ Other path being considered (s, a, h)

Current state s

Potential next state a

Searn Training Illustration

$$l_h(c, s, a) = \mathbb{E}_{y \sim (s,a,h)} c_y - \min_{a'} \mathbb{E}_{y \sim (s,a',h)} c_y$$



Prediction of current classifier h

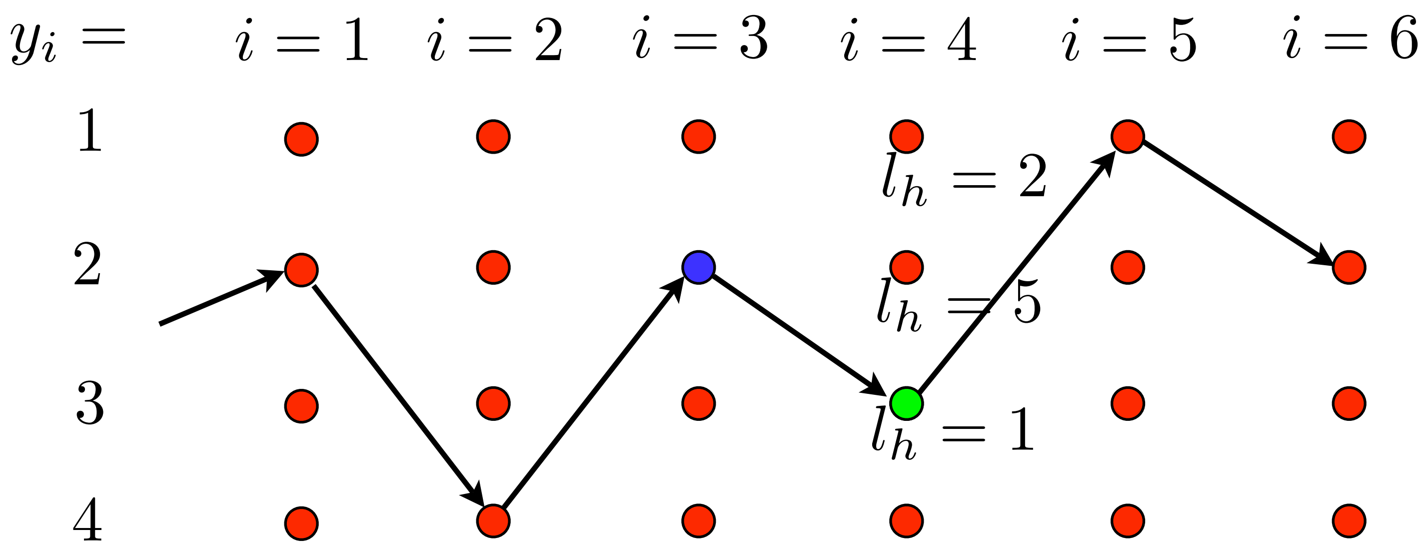
Other path being considered (s, a, h)

Current state s

Potential next state a

Searn Training Illustration

$$l_h(c, s, a) = \mathbb{E}_{y \sim (s,a,h)} c_y - \min_{a'} \mathbb{E}_{y \sim (s,a',h)} c_y$$



Prediction of current classifier h

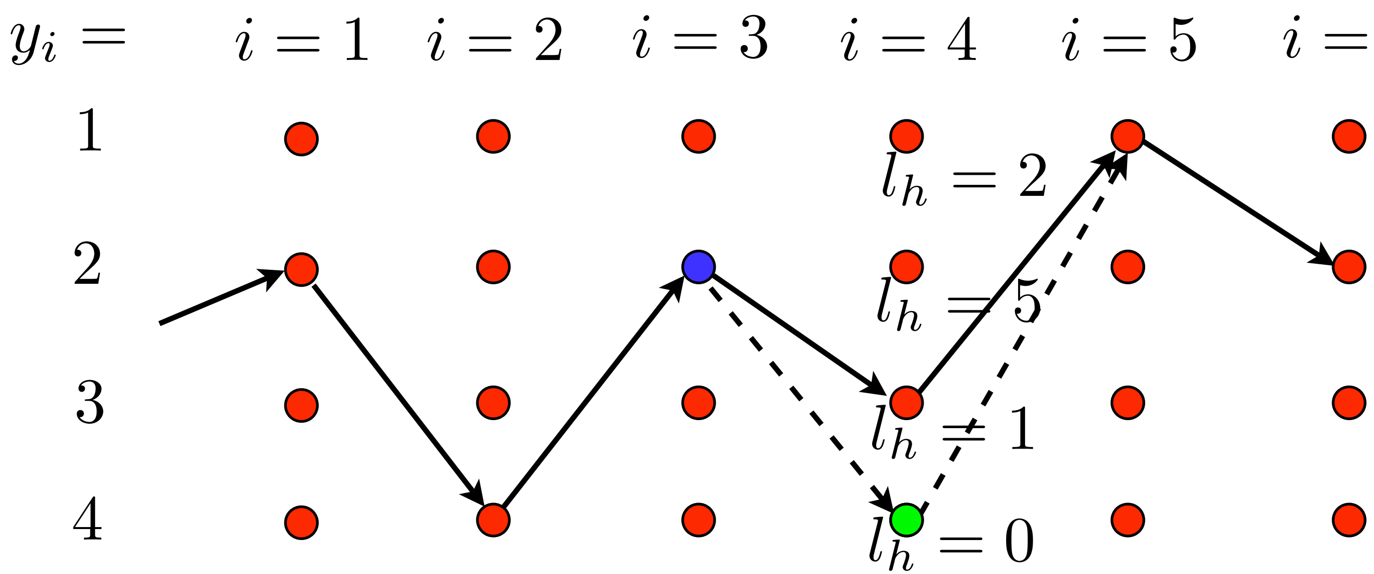
Other path being considered (s, a, h)

Current state s

Potential next state a

Searn Training Illustration

$$l_h(c, s, a) = \mathbb{E}_{y \sim (s,a,h)} c_y - \min_{a'} \mathbb{E}_{y \sim (s,a',h)} c_y$$



Prediction of current classifier h

Other path being considered (s, a, h)

Current state s

Potential next state a

Searn Meta-Algorithm

- Input: $(x_1, y_1), \ldots, (x_m, y_m), h_0, A$
- while h has a significant dependence on h_0:
  - Initialize set of cost-sensitive examples: $S \leftarrow \emptyset$
  - for $i \leftarrow 1, \ldots, m$
    - Compute prediction: $(y^{(1)}, \ldots, y^{(L)}) \leftarrow h(x_i)$
    - for $l \leftarrow 1, \ldots, L$
      - $s_l \leftarrow (x_i, y^{(1)}, \ldots, y^{(l)})$
      - for each next output a after s_l : $c'_{s_l,a} \leftarrow l_h(c, s_l, a)$
      - Compute features and add example: $S \leftarrow f(s_l, c')$
      - Learn and interpolate: $h' \leftarrow A(S); h \leftarrow \beta h' + (1 - \beta) h$
- Return h with h_0 removed

State consists of input and

Use losses to build up training examples for next iteration

Algorithm Analysis

- h_i is the classifier trained up to the ith iteration and $l_{h_i}(h'_i)$ is the loss of h'_i on this iteration's training examples
- T is the maximum length of any output sequence
- Theorem: If $c_{max} = \mathbb{E}_{(x,c)\sim\mathcal{D}} \max_y c_y$ and $l_{avg} = \frac{1}{I}\sum_{i=1}^{I} l_{h_i}(h'_i)$ (average loss over I iterations) then total loss with $\beta = 1/T^3$ and $2T^3 \ln T$ iterations is bounded as

$$L(\mathcal{D}, h_{last}) \leq L(\mathcal{D}, h_0) + 2T l_{avg} \log T + (1 + \log T) c_{max}/T$$

- Proof analyses the mixture of old and new classifiers
- In practice, β can be larger (more aggressive learning)

Proof

- Lemma 1: For classifier h^{new} learned by interpolating h and h' as $h^{new} \leftarrow \beta h' + (1-\beta)h$, if $c_{max} = \mathbb{E}_{(x,c)\sim\mathcal{D}} \max_y c_y$, we have
$$L(\mathcal{D}, h^{new}) \leq L(\mathcal{D}, h) + T\beta l_h^{CS}(h') + \frac{1}{2}\beta^2 T^2 c_{max}$$
- Proof: Consider 3 cases: h' is never called ($c=0$), is called exactly once ($c=1$), and is called more than once ($c \geq 2$)
- Then loss of h_{new} is bounded as
$$\begin{aligned}
L(\mathcal{D}, h^{\text{new}}) =& Pr(c=0)L(\mathcal{D}, h^{\text{new}} \mid c=0) \\
&+ Pr(c=1)L(\mathcal{D}, h^{\text{new}} \mid c=1) \\
&+ Pr(c\geq 2)L(\mathcal{D}, h^{\text{new}} \mid c\geq 2) \\
\leq& (1-\beta)^T L(\mathcal{D}, h) + T\beta(1-\beta)^{T-1}\Big[L(\mathcal{D}, h) + \ell_h^{\text{CS}}(h')\Big] \\
&+ \Big[1-(1-\beta)^T - T\beta(1-\beta)^{T-1}\Big]c_{\max}
\end{aligned}$$

Proof Cont'd

$$
\begin{aligned}
L(\mathcal{D}, h^{\mathrm{new}}) \leq & (1-\beta)^T L(\mathcal{D}, h) + T\beta(1-\beta)^{T-1}\left[L(\mathcal{D}, h) + \ell_h^{\mathrm{CS}}(h')\right] \\
& + \left[1-(1-\beta)^T - T\beta(1-\beta)^{T-1}\right] c_{\max} \\
= & L(\mathcal{D}, h) + T\beta(1-\beta)^{T-1}\ell_h^{\mathrm{CS}}(h') + \left(\sum_{i=2}^{T}(-1)^i\beta^i\binom{T}{i}\right) L(\mathcal{D}, h) \\
& + \left[1-(1-\beta)^T - T\beta(1-\beta)^{T-1}\right] c_{\max} \\
\leq & L(\mathcal{D}, h) + T\beta\ell_h^{\mathrm{CS}}(h') \\
& + \left[1-(1-\beta)^T - T\beta(1-\beta)^{T-1}\right] (c_{\max} - L(\mathcal{D}, h)) \\
\leq & L(\mathcal{D}, h) + T\beta\ell_h^{\mathrm{CS}}(h') \\
& + \left[1-(1-\beta)^T - T\beta(1-\beta)^{T-1}\right] c_{\max} \\
= & L(\mathcal{D}, h) + T\beta\ell_h^{\mathrm{CS}}(h') + \left(\sum_{i=2}^{T}(-1)^i\beta^i\binom{T}{i}\right) c_{\max} \\
\leq & L(\mathcal{D}, h) + T\beta\ell_h^{\mathrm{CS}}(h') + \frac{1}{2}T^2\beta^2 c_{\max}
\end{aligned}
$$

[Binomial Expansion]

[Binomial Expansion]

[Keep first term and $\beta < T/2$]

Proof Cont'd

- Lemma 2: After C/β iterations of Searn, the loss of the final classifier learned is bounded as

$$L(\mathcal{D}, h^{last}) \leq L(\mathcal{D}, h_0) + CTl_{avg} + c_{max}\left(\frac{1}{2}CT^2\beta + T\exp(-C)\right)$$

- Proof: Invoking Lemma 1 repeatedly, we get

$$L(\mathcal{D}, h) \leq L(\mathcal{D}, h_0) + CTl_{avg} + \left(\frac{1}{2}CT^2\beta\right)$$

- If we remove the initial (optimal) classifier, might incur a loss of $c_{\max}$; probability of failing after C/β iterations

$$T(1-\beta)^{C/\beta} \leq T\exp[-C]$$

Experiments

- Handwriting recognition [Kassel '95]
- Named entity recognition

  El presidente de la [Junta de Extremadura]$_{\text{ORG}}$, [Juan Carlos Rodríguez Ibarra]$_{\text{PER}}$, recibirá en la sede de la [Presidencia del Gobierno]$_{\text{ORG}}$ extremeño a familiares de varios de los condenados por el proceso “ [Lasa-Zabala]$_{\text{MISC}}$ ” , entre ellos a [Lourdes Díez Urraca]$_{\text{PER}}$, esposa del ex gobernador civil de [Guipúzcoa]$_{\text{LOC}}$ [Julen Elgorriaga]$_{\text{PER}}$; y a [Antonio Rodríguez Galindo]$_{\text{PER}}$, hermano del general [Enrique Rodríguez Galindo]$_{\text{PER}}$.

- Syntactic chunking and part-of-speech (POS) tagging

  [Great American]$_{\text{NP}}$ [said]$_{\text{VP}}$ [it]$_{\text{NP}}$ [increased]$_{\text{VP}}$ [its loan-loss reserves]$_{\text{NP}}$ [by]$_{\text{PP}}$ [$ 93 million]$_{\text{NP}}$ [after]$_{\text{PP}}$ [reviewing]$_{\text{VP}}$ [its loan portfolio]$_{\text{NP}}$, [raising]$_{\text{VP}}$ [its total loan and real estate reserves]$_{\text{NP}}$ [to]$_{\text{PP}}$ [$ 217 million]$_{\text{NP}}$.

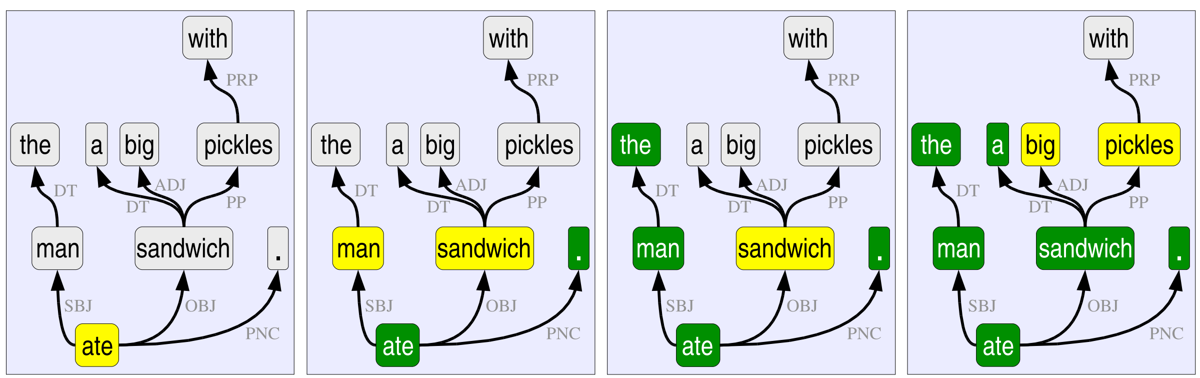
  $\text{Great}^{\text{NNP}}_{\text{B-NP}}$ $\text{American}^{\text{NNP}}_{\text{I-NP}}$ $\text{said}^{\text{VBD}}_{\text{B-VP}}$ $\text{it}^{\text{PRP}}_{\text{B-NP}}$ $\text{increased}^{\text{VBD}}_{\text{B-VP}}$ $\text{its}^{\text{PRP\$}}_{\text{B-NP}}$ $\text{loan-loss}^{\text{NN}}_{\text{I-NP}}$ $\text{reserves}^{\text{NNS}}_{\text{I-NP}}$ $\text{by}^{\text{IN}}_{\text{B-PP}}$ $\$^{\$}_{\text{B-NP}}$ $93^{\text{CD}}_{\text{I-NP}}$ $\text{million}^{\text{CD}}_{\text{I-NP}}$ $\text{after}^{\text{IN}}_{\text{B-PP}}$ $\text{reviewing}^{\text{VBG}}_{\text{B-VP}}$ $\text{its}^{\text{PRP\$}}_{\text{B-NP}}$ $\text{loan}^{\text{NN}}_{\text{I-NP}}$ $\text{portfolio}^{\text{NN}}_{\text{I-NP}}$ $.^{.}_{\text{O}}$

Experiments

ALGORITHM	Handwriting		NER		Chunk	C+T
	Small	Large	Small	Large		
CLASSIFICATION						
Perceptron	65.56	70.05	91.11	94.37	83.12	87.88
Log Reg	68.65	72.10	93.62	96.09	85.40	90.39
SVM-Lin	75.75	82.42	93.74	97.31	86.09	93.94
SVM-Quad	82.63	82.52	85.49	85.49	$\sim$	$\sim$
STRUCTURED						
Str. Perc.	69.74	74.12	93.18	95.32	92.44	93.12
CRF	–	–	94.94	$\sim$	94.77	96.48
SVM$^{\text{struct}}$	–	–	94.90	$\sim$	–	–
M^{3}N-Lin	81.00	$\sim$	–	–	–	–
M^{3}N-Quad	87.00	$\sim$	–	–	–	–
SEARN						
Perceptron	70.17	76.88	95.01	97.67	94.36	96.81
Log Reg	73.81	79.28	95.90	98.17	94.47	96.95
SVM-Lin	82.12	90.58	95.91	98.11	94.44	96.98
SVM-Quad	87.55	90.91	89.31	90.01	$\sim$	$\sim$

Experiments

- New "vine-growth" model for sentence summarization
- DUC 2005 data set: 50 sets of 25 documents each
- Evaluation: Rouge (n-gram overlap) vs. human summaries



" "   " ate . "   " the man ate . "   " the man ate a sandwich . "

	ORACLE		SEARN		BAYESUM			
	Vine	Extr	Vine	Extr	D05	D03	Base	Best
100 w	.0729	.0362	.0415	.0345	.0340	.0316	.0181	-
250 w	.1351	.0809	.0824	.0767	.0762	.0698	.0403	.0725

Bibliography

- Harold C. Daumé III, Practical structured learning for natural language processing, Ph.D. Thesis, University of Southern California, 2006.
- Harold C. Daumé III, John Langford, and Daniel Marcu. Search-Based Structured Prediction, Submitted to Machine Learning, 2007
- Robert Kassel. A Comparison of Approaches to On-line Handwritten Character Recognition. PhD thesis, Massachusetts Institute of Technology, Spoken Language Systems Group, 1995.
- Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. Journal of Machine Learning Research, 2(5):265-292, 2001.
- Ben Taskar, Carlos Guestrin, and Daphne Koller. Max-margin Markov networks. Neural Information Processing Systems (NIPS) 16, 2003.
- Ioannis Tsochantaridis, Thomas Hofmann, Thorsten Joachims, and Yasemin Altun. Support vector machine learning for interdependent and structured output spaces. Proceedings ICML, 2004.
- Fei Sha and Lawrence K. Saul. Large margin hidden Markov models for automatic speech recognition, Neural Information Processing Systems (NIPS) 19, 2007.
- William W. Cohen and Vitor Carvalho. Stacked sequential learning. In Proceedings of the International Joint Conference on Artificial Intelligence (IJ-CAI), 2005.
- Michael Collins and Brian Roark. Incremental parsing with the perceptron algorithm. In Proceedings of the Conference of the Association for Computational Linguistics (ACL), 2004.
- Alina Beygelzimer, Varsha Dani, Tom Hayes, John Langford, and Bianca Zadrozny. Error limiting reductions between classification tasks. In Proceedings of the International Conference on Machine Learning (ICML), 2005.