Structured Prediction 1

Joseph Keshet
Dept. of Computer Science
Bar-Ilan University
Binary problems
Structured problems
Structured problems

- Structured output rather than yes/no or \{-1, +1\}.

- Exponentially large output space (exponential number of configurations)

- The evaluation is unique to the task and usually not 0-1 loss

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, 'and what is the use of a book,' thought Alice 'without pictures or conversation?' So she was considering in her own mind (as well as she could, for the hot day made her feel very sleepy and stupid), whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies, when suddenly a White Rabbit with pink eyes ran close by her.
Evaluation

- measure of performance
- evaluation metric
- task loss
- cost
- loss
Example:

Automatic Voice Onset Time Measurement
Does the Queen speak the Queen’s English?

Phonetic imitation, convergence, accommodation

Giles et al., 1991; Goldinger, 1998; Pardo, 2006; Babel, 2009, ...
Big Brother
Season 9, 2008, UK
Voice Onset Time (VOT) is the time between the onset of a stop’s burst and the onset of voicing.
SOME EFFECTS OF CONTEXT ON VOICE ONSET TIME IN ENGLISH STOPS*

Leigh Lisker** and Arthur S. Abramson***
Haskins Laboratories, New York

Recent work has led us to the conclusion that the English stop categories /bdg/ and /ptk/ are distinguished by the timing of changes in glottal aperture relative to

In our study of word-initial stops in English the finding which deserves primary emphasis is that the two categories /ptk/ and /bdg/ are characterized by significantly different distributions of VOT values. This relation holds true quite independently of several contextual factors investigated. At the same time, however, the timing of

Lisker & Abramson, 1964
Eimas et al. (1971) tested infants at two different ages, 1 and 4 months, for their ability to discriminate the English voicing contrast that is observed in the syllables [ba] and [pa]. In articulatory terms, the voicing

Jusczyk, *The discovery of spoken language*, p. 50, 2000
Eimas et al. (1971) tested infants at two different ages, 1 and 4 months, for their ability to discriminate the English voicing contrast that is observed in the syllables [ba] and [pa]. In articulatory terms, the voicing

Jusczyk, *The discovery of spoken language*, p. 50, 2000

---

**Speech Perception by the Chinchilla:**

**Voiced-Voiceless Distinction in Alveolar Plosive Consonants**

Abstract. Four chinchillas were trained to respond differently to /t/ and /d/ consonant-vowel syllables produced by four talkers in three vowel contexts. This training generalized to novel instances, including synthetically produced /da/ and /ta/ (voice-on-
Voice Onset Time (VOT)

VOT begins at $t_b$

VOT ends at $t_v$

$x = (x_1, \ldots, \ldots, x_T)$
How good is the predicted VOT?

prediction VOT  \[ \hat{y} = \hat{t}_v - \hat{t}_b \]

manual VOT  \[ y = t_v - t_b \]

\[
L(y, \hat{y}) = \max\{|y - \hat{y}| - \epsilon, 0\}
\]
Goal

Find VOT predictor such that

\[ E[L(y, \hat{y})] \]

is minimized.

- \( \hat{y} \) is the predicted VOT (burst time minus voicing time)
- the expectation is taken over a random draw of \( x \) and \( y \)
Model and Inference

\[ \hat{y} = \arg \max_{\hat{y}} \ w \cdot \phi(x, \hat{y}) \]

weight vector
\[ w \in \mathbb{R}^n \]

feature maps
Acoustic Features

Each $x_t$ is a 7 components taken every 1 ms with 5 ms window:
Acoustic Features

Each $x_t$ is a 7 components taken every 1 ms with 5 ms window:
Acoustic Features

Each $x_t$ is a 7 components taken every 1 ms with 5 ms window:

\[
\begin{align*}
E_{\text{low}} & \quad E_{\text{high}} \\
H_{\text{wiener}} & \\
\text{ACF, Voicing} & \\
\end{align*}
\]

\[
\log \int E(f) - \int \log E(f)
\]
Model and Inference

\[ \hat{y} = \arg \max_{\hat{y}} w \cdot \phi(x, \hat{y}) \]

1. if \( \hat{y} \) \( (t_b \text{ and } t_v) \) makes sense, then the value of the feature map should be high

2. take into account different lengths of the input \( x \)
Model and Inference

\[ \hat{y} = \arg \max_{\hat{y}} w \cdot \phi(x, \hat{y}) \]

1. if \( \hat{y} \) (\( t_b \) and \( t_v \)) makes sense, then the value of the feature map should be high

2. take into account different lengths of the input \( x \)
1. if $\hat{y}$ ($t_b$ and $t_v$) makes sense, then the value of the feature map should be high
2. take into account different lengths of the input $x$
Typical duration of \((t_v - t_b)\)
Typical duration of \((t_v - t_b)\)
mean of $\log(E_{\text{high}})$ in $(t_b, t_v)$ – mean in $(1, t_b)$
mean of $\log(E_{\text{high}})$ in $(t_b, t_v)$ – mean in $(1, t_b)$
local differences of \textit{Wiener entropy} around $t_b$ and around $t_v$
local differences of Wiener entropy around $t_b$ and around $t_v$
\[ \hat{y} = \arg \max_{\hat{y}} \ w \cdot \phi(x, \hat{y}) \]

weight vector
\[ w \in \mathbb{R}^n \]
\[
\hat{y} = \arg\max_{\hat{y}} \left( w \right) \phi(x, \hat{y})
\]
\[ \hat{y} = \arg \max_{\hat{y}} \ w \cdot \phi(x, \hat{y}) \]

- **weight vector**
- \( w \in \mathbb{R}^n \)
Empirical results

- Conversational speech
- British accent.
- TV studio.
- 704 stops, 4 speakers.
- Hand-annotated.

(Sonderegger and Keshet, 2011; Sonderegger and Keshet, 2012)
System of choice in phonology labs

[Image of a phonology software interface showing waveforms and time stamps]
Example: Keyword Spotting
"a telltale sign [of autism] is the lack of attention the child pays to the people interacting with him."
Given a speech signal

and a word

**Goal**: find if the word is uttered in the speech signal and where

bought
Given a speech signal

and a word

bought

Goal: find if the word is uttered in the speech signal and where
\[ \bar{t} = (t_{\text{start}}, t_{\text{end}}) \]

\[ k = \] bought

\[ \bar{x} = (x_1, x_2, \ldots, x_T) \]
Keyword Spotter

\[ f(\bar{x}, k) \]

\( k = /bought/ \)

\( \bar{x} \)

\( \bar{t}' \)

speech input signal

predicted decision

predicted timing

input keyword
\[ k = /bought/ \]

\[ \bar{x} \rightarrow \text{Keyword Spotter} \rightarrow \max_{\bar{t}} f(\bar{x}, k, \bar{t}) \rightarrow \bar{t}' \]

- speech input signal
- predicted decision
- predicted timing
- input keyword
The task loss

The performance of keyword spotting system is measured by **Receiver operating Characteristics (ROC) curve**.

true positive =

\[
\frac{\text{detected utterances with keywords}}{\text{total utterances with keywords}}
\]

false positive =

\[
\frac{\text{detected utterances without keywords}}{\text{total utterances without keywords}}
\]
The task loss

The performance of keyword spotting system is measured by Receiver operating Characteristics (ROC) curve.

true positive =
\[
\frac{\text{detected utterances with keywords}}{\text{total utterances with keywords}}
\]

false positive =
\[
\frac{\text{detected utterances without keywords}}{\text{total utterances without keywords}}
\]
The task loss

The performance of keyword spotting system is measured by Receiver operating Characteristics (ROC) curve.

true positive =
\[
\frac{\text{detected utterances with keywords}}{\text{total utterances with keywords}}
\]

false positive =
\[
\frac{\text{detected utterances without keywords}}{\text{total utterances without keywords}}
\]

\[
A = 1
\]
The task loss

The performance of keyword spotting system is measured by **Receiver operating Characteristics (ROC)** curve.

**true positive =**
\[
\frac{\text{detected utterances with keywords}}{\text{total utterances with keywords}}
\]

**false positive =**
\[
\frac{\text{detected utterances without keywords}}{\text{total utterances without keywords}}
\]
The task loss

The performance of keyword spotting system is measured by Receiver operating Characteristics (ROC) curve.

true positive =
\[
\frac{\text{detected utterances with keywords}}{\text{total utterances with keywords}}
\]

false positive =
\[
\frac{\text{detected utterances without keywords}}{\text{total utterances without keywords}}
\]
Model and Inference

\[ \hat{y} = \arg \max_\hat{y} \ w \cdot \phi(x, p, \hat{y}) \]

- Speech signal (input)
- Sequence of phonemes (input)
- Weight vector \( w \in \mathbb{R}^n \)
- Feature maps
- Locations of the phonemes in the speech signal
Feature Map I

Cumulative spectral change around the boundaries
Feature Map I

Cumulative spectral change around the boundaries
Feature Map I

Cumulative spectral change around the boundaries
Feature Map I

Cumulative spectral change around the boundaries
Feature Map II

Cumulative confidence in the phoneme sequence

frame based phoneme classification score
(Dekel, Keshet, Singer, 2004)
Feature Map II

Cumulative confidence in the phoneme sequence

Frame based phoneme classification score (Dekel, Keshet, Singer, 2004)
Feature Map III

Phoneme duration model
Feature Map IV

Speaking-rate modeling ("dynamics")

Spectrogram at different rates of articulation (after Pickett, 1980)
The General Case
Goal

Find a function such that

$$\mathbb{E}[L(y, \hat{y})]$$

is minimized.

- $\hat{y}$ is the predicted sequence (label)
- the expectation is taken over a random draw of $(x, y)$
\( \hat{y}_w = \arg\max_y \ w \cdot \phi(x, y) \)

\( w^* = \arg\min_w \ E [L(y, \hat{y}_w)] \)

- \( \hat{y} \) is the predicted sequence (label)
- the expectation is taken over a random draw of \((x, y)\)
Structured Perceptron
Structured Perceptron

\[ \hat{y} = \arg \max_{\hat{y}} w \cdot \phi(x_i, \hat{y}) \]

\[ \Delta w \propto \phi(x_i, y_i) - \phi(x_i, \hat{y}) \]
Structured Support Vector Machines (SVMs)
Structured SVM

\[ w^* = \arg \min_w \mathbb{E}_{(x,y) \sim D} [L(y, \hat{y}_w(x))] \]
Structured SVM

\[ w^* = \arg \min_w \mathbb{E}_{(x,y) \sim D} [L(y, \hat{y}_w(x))] \]
Structured SVM

\[ w^* = \arg \min_w \mathbb{E}_{(x,y) \sim D} [L(y, \hat{y}_w(x))] \]

\[ w^* = \arg \min_w \frac{1}{M} \sum_{i=1}^{M} L(y_i, \hat{y}_w(x_i)) + \frac{\lambda}{2} \| w \|^2 \]
Structured SVM

\[ w^* = \arg \min_w \mathbb{E}_{(x,y) \sim D} [L(y, \hat{y}_w(x))] \]

\[ w^* = \arg \min_w \frac{1}{M} \sum_{i=1}^{M} L(y_i, \hat{y}_w(x_i)) + \frac{\lambda}{2} \| w \|^2 \]

\[ \leq \arg \min_w \frac{1}{M} \sum_{i=1}^{M} \hat{L}_{\text{hinge}}(y_i, \hat{y}_w(x_i)) + \frac{\lambda}{2} \| w \|^2 \]
Structured SVM

\[ w^* = \arg\min_w \mathbb{E}_{(x,y) \sim D} [L(y, \hat{y}_w(x))] \]

\[ w^* = \arg\min_w \frac{1}{M} \sum_{i=1}^{M} L(y_i, \hat{y}_w(x_i)) + \frac{\lambda}{2} \| w \|^2 \]

\[ \leq \arg\min_w \frac{1}{M} \sum_{i=1}^{M} \hat{L}_{\text{hinge}}(y_i, \hat{y}_w(x_i)) + \frac{\lambda}{2} \| w \|^2 \]

surrogate hinge-loss
Structured SVM

\[ w^* = \arg\min_w \mathbb{E}_{(x, y) \sim D}[L(y, \hat{y}_w(x))]] \]

\[ w^* = \arg\min_w \frac{1}{M} \sum_{i=1}^{M} L(y_i, \hat{y}_w(x_i)) + \frac{\lambda}{2} \|w\|^2 \]

\[ \leq \arg\min_w \frac{1}{M} \sum_{i=1}^{M} \hat{L}_{\text{hinge}}(y_i, \hat{y}_w(x_i)) + \frac{\lambda}{2} \|w\|^2 \]

**surrogate hinge-loss**

\[ \hat{L}_{\text{hinge}}(y_i, \hat{y}_w(x_i)) = \max_{\hat{y}} [L(y_i, \hat{y}) - w \cdot \phi(x_i, y_i) + w \cdot \phi(x_i, \hat{y})] \]
Structured SVM

\[ \hat{y}_L = \arg \max_{\hat{y}} \ w \cdot \phi(x_i, \hat{y}) + L(y_i, \hat{y}) \]

\[ \Delta w \propto \phi(x_i, y_i) - \phi(x_i, \hat{y}_L) - \lambda w \]
Structured SVM

\[ \hat{y}_L = \arg \max_{\hat{y}} w \cdot \phi(x_i, \hat{y}) + L(y_i, \hat{y}) \]

\[ \Delta w \propto \phi(x_i, y_i) - \phi(x_i, \hat{y}_L) - \lambda w \]
Structured SVM

Margin generalization bound

\[ \mathbb{E}[L(y, \hat{y})] \preceq \frac{1}{M} \sum_{i} \max_{y: w \cdot \phi(x_i, y_i) - w \cdot \phi(x_i, y) \leq H(y, y_i)} L(y_i, y) + O\left(\frac{1}{\sqrt{M}}, \sqrt{|w|}, \frac{1}{\sqrt{\delta}}, \ldots\right) \]

This involves both the Hamming distance (as a margin requirement) and the evaluation metric function (Taskar et al. 2003; McAllester 2007)

No guarantee for the risk (expected measure of performance)
Conditional Random Fields (CRFs)
\[ w^* = \arg \min_w \mathbb{E}_{(x,y) \sim D} [L(y, \hat{y}_w(x))] \]
\[ w^* = \arg \min_w \mathbb{E}_{(x,y) \sim \mathcal{D}} [L(y, \hat{y}_w(x))] \]
\[ w^* = \arg \min_w \mathbb{E}_{(x, y) \sim \mathcal{D}} \left[ L(y, \hat{y}_w(x)) \right] \]

\[ w^* = \arg \min_w \frac{1}{M} \sum_{i=1}^{M} - \log P_w(y_i | x_i) + \frac{\lambda}{2} \| w \|^2 \]

where \( P_w(y | x) = \frac{1}{Z} \exp(w \cdot \phi(x, y)) \)
\[ \Delta w \propto \phi(x_i, y_i) - \mathbb{E}_{\hat{y} \sim \mathcal{P}_w(y|x_i)} [\phi(x_i, \hat{y})] - \lambda w \]
CRF ignores the measure of performance