

Experiments with Detector-based Conditional Random Fields in Phonetic Recognition

Jeremy Morris & Eric Fosler-Lussier
04/19/2007

speech &
language
technologies
@osu



Outline

- Background
- Previous Work
- Feature Combination Experiments
- Viterbi Realignment Experiments
- Conclusions and Future Work

Background

- Goal: Integrate outputs of speech attribute detectors together for recognition
 - e.g. Phone classifiers, phonological feature classifiers
- Attribute detector outputs highly correlated
 - Stop detector vs. phone classifier for /t/ or /d/
- Accounting for correlations in HMM
 - Ignore them (decreased performance)
 - Full covariance matrices (increased parameters)
 - Explicit decorrelation (e.g. Karhunen-Loeve transform)

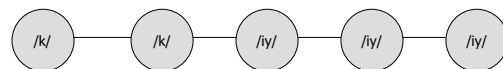
Background

- Conditional Random Fields (CRFs)
 - Discriminative probabilistic model
 - Used successfully in various domains such as part of speech tagging and named entity recognition
 - Directly defines a posterior probability of a sequence Y given an input sequence X
 - e.g. $P(Y|X)$
 - Does not make independence assumptions about correlations among input features

Background

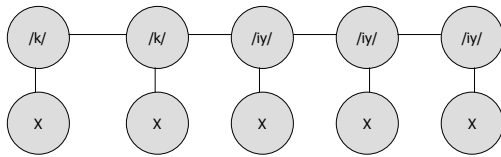
- CRFs for ASR
 - Phone Classification (Gunawardana et al., 2005)
 - Uses sufficient statistics to define feature functions
 - Different approach than NLP tasks using CRFs
 - Define binary feature functions to characterize observations
 - Our approach follows the latter method
 - Use neural networks to provide "soft binary" feature functions (e.g. posterior phone outputs)

Conditional Random Fields



- Based on the framework of Markov Random Fields

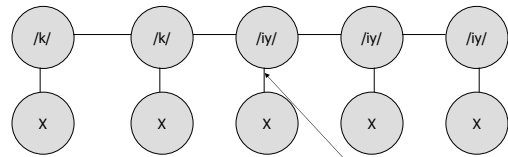
Conditional Random Fields



- Based on the framework of Markov Random Fields
 - A CRF iff the graph of the label sequence is an MRF when conditioned on a set of input observations (Lafferty et al., 2001)

7

Conditional Random Fields

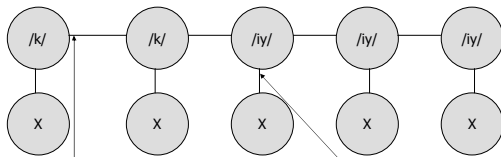


- Based on the framework of Markov Random Fields
 - A CRF iff the graph of the label sequence is an MRF when conditioned on the input observations

State functions help determine the identity of the state

8

Conditional Random Fields



- Based on the framework of Markov Random Fields
 - A CRF iff the graph of the label sequence is an MRF when conditioned on the input observations

Transition functions add associations between transitions from one label to another

State functions help determine the identity of the state

9

Conditional Random Fields

$$P(Y | X) = \frac{\exp \left(\sum_k \left(\sum_i \lambda_i s_i(x, y_k) + \sum_j \mu_j t_j(x, y_k, y_{k-1}) \right) \right)}{Z(x)}$$

- CRF defined by a weighted sum of state and transition functions
 - Both types of functions can be defined to incorporate observed inputs
 - Weights are trained by maximizing the likelihood function via gradient descent methods

10

Previous Work

- Implemented CRF models on data from phonetic attribute detectors
 - Performed phone recognition
 - Compared results to Tandem/HMM system on same data
- Experimental Data
 - TIMIT corpus of read speech

11

Attribute Selection

- Attribute Detectors
 - ICSI QuickNet Neural Networks
- Two different types of attributes
 - Phonological feature detectors
 - Place, Manner, Voicing, Vowel Height, Backness, etc.
 - N-ary features in eight different classes
 - Posterior outputs -- P(Place=dental | X)
 - Phone detectors
 - Neural networks output based on the phone labels
 - Trained using PLP 12+deltas

12

Experimental Setup

- CRF code
 - Built on the Java CRF toolkit from Sourceforge
 - <http://crf.sourceforge.net>
 - Performs maximum log-likelihood training
 - Uses Limited Memory BGFS algorithm to perform minimization of the log-likelihood gradient

13

Experimental Setup

$$s_{/t/,f}(y, x) = \begin{cases} NN_f(x), & \text{if } y = /t/ \\ 0, & \text{otherwise} \end{cases}$$

- Feature functions built using the neural net output
 - Each attribute/label combination gives one feature function
 - Phone class: $S_{N/N}$ or $S_{N/S}$
 - Feature class: $S_{N,stop}$ or $S_{N,dental}$

14

Experimental Setup

- Baseline system for comparison
 - Tandem/HMM baseline (Hermansky et al., 2000)
 - Use outputs from neural networks as inputs to gaussian-based HMM system
 - Built using HTK HMM toolkit
- Linear inputs
 - Better performance for Tandem with linear outputs from neural network
 - Decorrelated using a Karhunen-Loeve (KL) transform

15

Initial Results (Morris & Fosler-Lussier, 06)

Model	Params	Phone Accuracy
Tandem [1] (phones)	20,000+	60.82%
Tandem [3] (phones) 4mix	420,000+	68.07%*
CRF [1] (phones)	5280	67.32%*
Tandem [1] (feas)	14,000+	61.85%
Tandem [3] (feas) 4mix	360,000+	68.30%*
CRF [1] (feas)	4464	65.45%*
Tandem [1] (phones/feas)	34,000+	61.72%
Tandem [3] (phones/feas) 4mix	774,000+	68.46%
CRF (phones/feas)	7392	68.43%*

* Significantly ($p > 0.05$) better than comparable Tandem monophone system
 * Significantly ($p > 0.05$) better than comparable CRF monophone system

Feature Combinations

- CRF model supposedly robust to highly correlated features
 - Makes no assumptions about feature independence
- Tested this claim with combinations of correlated features
 - Phone class outputs + Phono. Feature outputs
 - Posterior outputs + transformed linear outputs
- Also tested whether linear, decorrelated outputs improve CRF performance

17

Feature Combinations - Results

Model	Phone Accuracy
CRF (phone posteriors)	67.32%
CRF (phone linear KL)	66.80%
CRF (phone post+linear KL)	68.13%*
CRF (phono. feature post.)	65.45%
CRF (phono. feature linear KL)	66.37%
CRF (phono. feature post+linear KL)	67.36%*

* Significantly ($p > 0.05$) better than comparable posterior or linear KL systems

18

Viterbi Realignment

- Hypothesis: CRF results obtained by using only pre-defined boundaries
 - HMM allows “boundaries” to shift during training
 - Basic CRF training process does not
- Modify training to allow for better boundaries
 - Train CRF with fixed boundaries
 - Force align training labels using CRF
 - Adapt CRF weights using new boundaries

19

Viterbi Realignment - Results

Model	Accuracy
CRF (phone posteriors)	67.32%
CRF (phone posteriors – realigned)	69.92%**
Tandem[3] 4mix (phones)	68.07%
Tandem[3] 16mix (phones)	69.34%
CRF (phono. fea. linear KL)	66.37%
CRF (phono. fea. lin-KL – realigned)	68.99%**
Tandem[3] 4mix (phono fea.)	68.30%
Tandem[3] 16mix (phono fea.)	69.13%
CRF (phones+feas)	68.43%
CRF (phones+feas – realigned)	70.63%**
Tandem[3] 16mix (phones+feas)	69.40%

* Significantly (p>0.05) better than comparable CRF monophone system

** Significantly (p>0.05) better than comparable Tandem 4mix triphone system

*** Significantly (p>0.05) better than comparable Tandem 16mix triphone system

Conclusions

- Using correlated features in the CRF model did not degrade performance
 - Extra features improved performance for the CRF model across the board
- Viterbi realignment training significantly improved CRF results
 - Improvement did not occur when best HMM-aligned transcript was used for training

21

Future Work

- Recently implemented stochastic gradient training for CRFs
 - Faster training, improved results
- Work currently being done to extend the model to word recognition
- Also examining the use of transition functions that use the observation data

22