A Boosting Approach to Topic Spotting on Subdialogues

Kary Myers

Department of Statistics
Carnegie Mellon University

Michael Kearns
Satinder Singh
Marilyn A. Walker

AT&T Labs

Eyes and Ears Project

The dream scenario

⋆ Listen in on conversations

⋆ Look for keywords, salient phrases

⋆ Enrich state space with estimates of current topic

⋆ Take relevant actions

Our acoustical nightmare

⋆ Open, multi-directional “live” microphone

⋆ Multiple speakers at varying distances from mike

⋆ Unconstrained conversational speech

⋆ Varying levels and sources of background noise

Switchboard Speech Corpus

⋆ 2340 spontaneous 5-minute telephone conversations

⋆ 70 topics

⋆ Short audio files by speaker

⋆ Full text transcriptions of each conversation

“Verbatim” Transcript Excerpt

```
 A.1: [Laughter].
 B.2: Uh-huh.
 A.3: Um, I guess we're supposed to talk about music.
 B.4: Okay.
 A.5: And, uh, let me go ahead and push one here.
      [tone] Uh, do you ha-, are you a musician yourself?
 B.6: Uh, well, I sing.
 A.7: Uh-huh.
 B.8: I don't play an instrument.
 A.9: Uh-huh.  Where, do you sing in, in a choir or
      a choral group?
B.10: Oh, not right now.
```

Our Constraints

- ⋆ Want quick estimates of topic—study of **subdialogues**
- ⋆ Prefer to build on “off-the-shelf” technologies:

  WATSON—automatic speech recognizer
  (Sharp, et al., 1997)

  BOOSTEXTER—system for text classification
  (Schapire & Singer, 2000)

- ⋆ Interested in fast performance even at the expense of classification and recognition accuracy
- ⋆ Want topics of relevance to Eyes and Ears
  e.g., books, movies, sports, weather

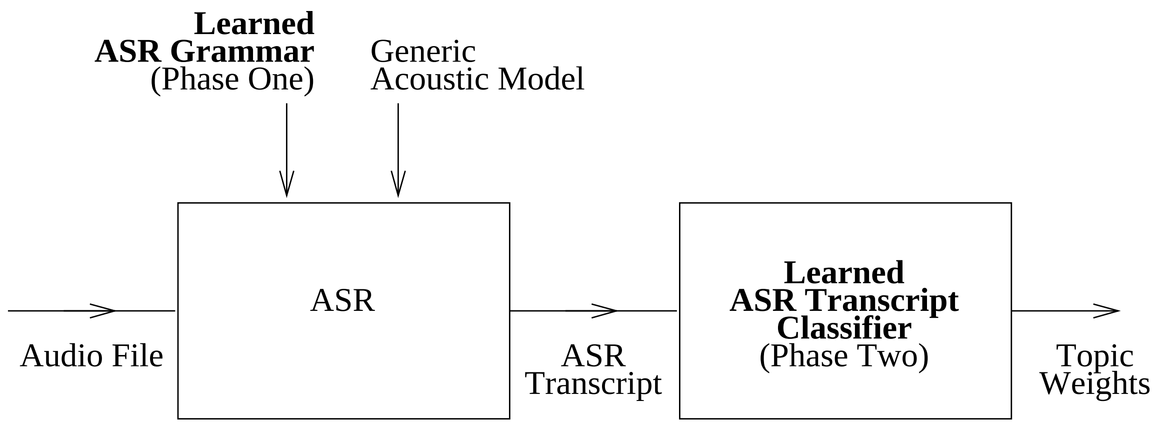
Conversation Data

Topic	**Train**	**Valid**	**Test**
Books and Literature	15	3	0
Exercise and Fitness	26	2	4
Family Life	27	3	1
Movies	26	4	7
Music	25	3	3
Pets	25	5	2
Recipes, Food, Cooking	27	3	3
Restaurants	15	1	3
Weather and Climate	17	0	4
Sports (majority class, 13%)	31	5	2
Total	**234**	**29**	**29**

Divided validation and test sets into six **subdialogue lengths**: full, half, fourth, eighth, sixteenth, thirty-second

Two-Phase Learning Algorithm

Learned Classifier:



⋆ **Phase One:** Learning the ASR grammar
(2 parameters)

⋆ **Phase Two:** Mapping ASR transcripts to topics
(1 parameter)

Phase 1: Learn the ASR Grammar

Steps

1. Train BoosTexter to classify the *verbatim* transcripts.

2. Extract the "best" BoosTexter rules to form a grammar of word sequences.

Selecting Word Sequences from Rules

Two rules learned by BoosTexter:

Topic	**'listen to'**	**'uh yeah'**
Books and Literature	-0.183	-0.474
Exercise and Fitness	-0.972	-0.548
Family Life	-0.815	-0.077
Movies	-1.604	0.530
Music	2.044	-1.122
Pets	-1.022	-0.976
Recipes, Food, Cooking	-1.089	-1.542
Restaurants	-2.371	-0.998
Weather and Climate	-0.605	0.456
Sports	-0.639	-0.138
Variance	**1.306**	**0.470**

Grammar is the OR of sequences with largest variances.

Phase 1: Learn the ASR Grammar

Parameters

- **Grammar Size:**
  20, 100, 200, or 400 word sequences

- **Verbatim Transcript Length:**
  Full conversations, eighths, or sixteenths

Phase 2: Map ASR Transcripts to Topics

Steps

1. Run the audio files through the ASR using the grammar learned in Phase 1.

2. Train BoosTexter to classify the noisy ASR transcripts.

   `...east any uh...the deer the grass dip`

   `...that i dip...sing ankle cook`

   `...the fish...cat the sing yeah and um`

Parameter

⋆ **ASR Transcript Length:**
Full conversations, halves, fourths, eighths, sixteenths, or thirty-seconds

Overview of Training and Selection

Parameters Considered: $3 \times 4 \times 6 = 72$ distinct classifiers

Verbatim Transcript Length	Grammar Size	ASR Transcript Length
Full	20	Full
Eighth	100	Half
Sixteenth	200	Fourth
	400	Eighth
		Sixteenth
		Thirty-second

1. Select the classifiers that perform best on the six subdialogue lengths within the validation set.
2. Consider the performance of these fixed classifiers on the data in the test set.

Experimental Results

⋆ Overview of classification accuracy

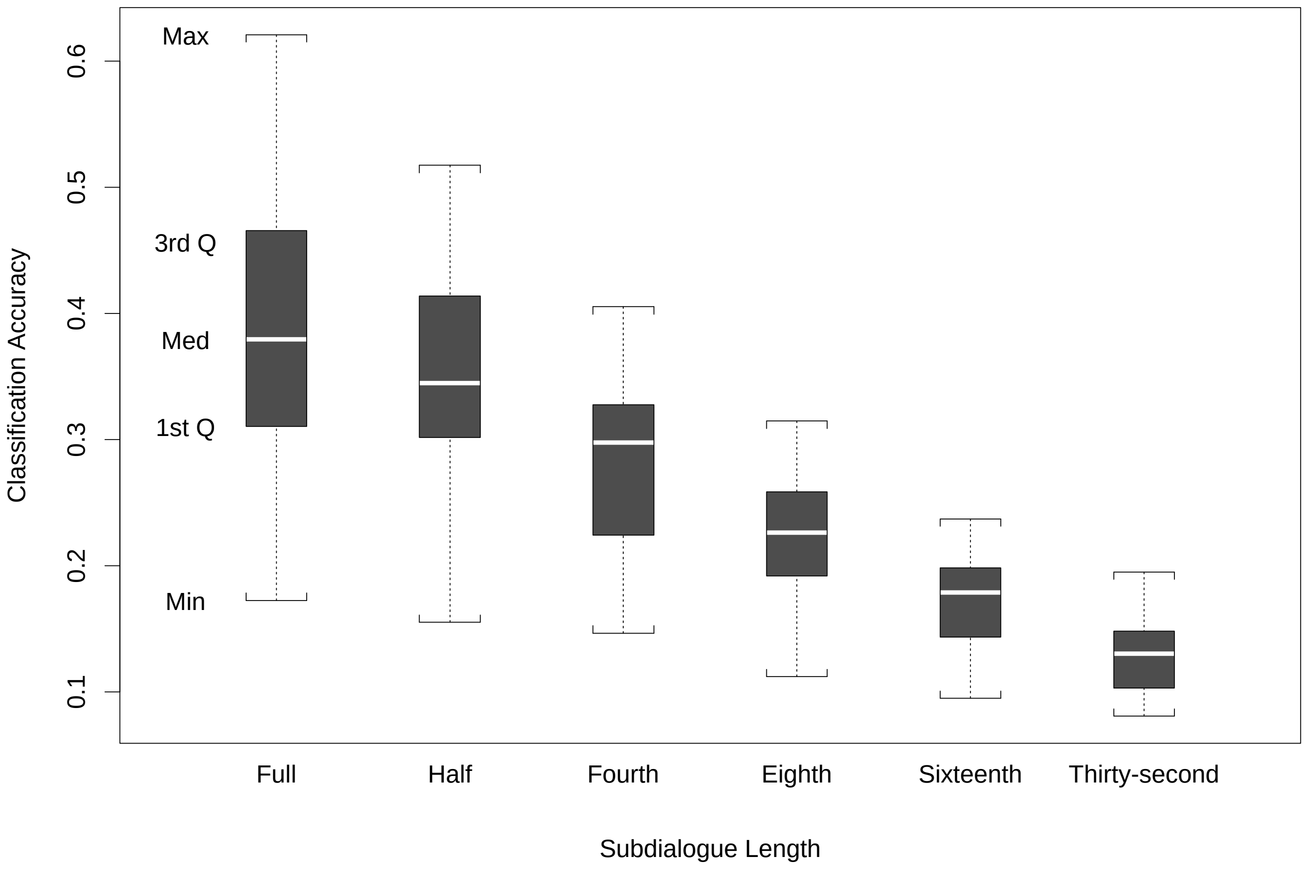
⋆ “Anytime” classification

⋆ Classification with “perfect” ASR

⋆ Incorporating confidence measures

⋆ Classification accuracy by topic

Effect of Subdialogue Length on Accuracy



Classifiers Selected on Validation Set

	Best Parameter Settings (Validation Set)			Performance Measures (Test Set)		
Subdialogue Length	Verbatim Trscrpt Length	Grammar Size	ASR Trscrpt Length	Classification Accuracy	Mean Regret	Max Regret
Full	Full	200	Full	0.448	0.081	0.121
Half	Sixteenth	400	Full	0.483	0.042	0.086
Fourth	Eighth	200	Fourth	0.353	0.053	0.103
Eighth	Sixteenth	400	Thirty-second	0.302	0.020	0.086
Sixteenth	Sixteenth	400	Thirty-second	0.237	0.020	0.086
Thirty-second	Full	200	Sixteenth	0.154	0.052	0.103

Best "anytime" classifier:

The classifier chosen for eighth- and sixteenth-length subdialogues jointly minimizes the mean and maximum regrets.

How Does ASR Impact Classification?

The verbatim transcripts are like having perfect ASR.

Subdialogue Length	Verbatim Transcripts	ASR Transcripts
Full	1.0000	0.4483
Half	0.9483	0.4828
Fourth	0.7845	0.3534
Eighth	0.7056	0.3017
Sixteenth	0.6442	0.2371
Thirty-second	0.5089	0.1541

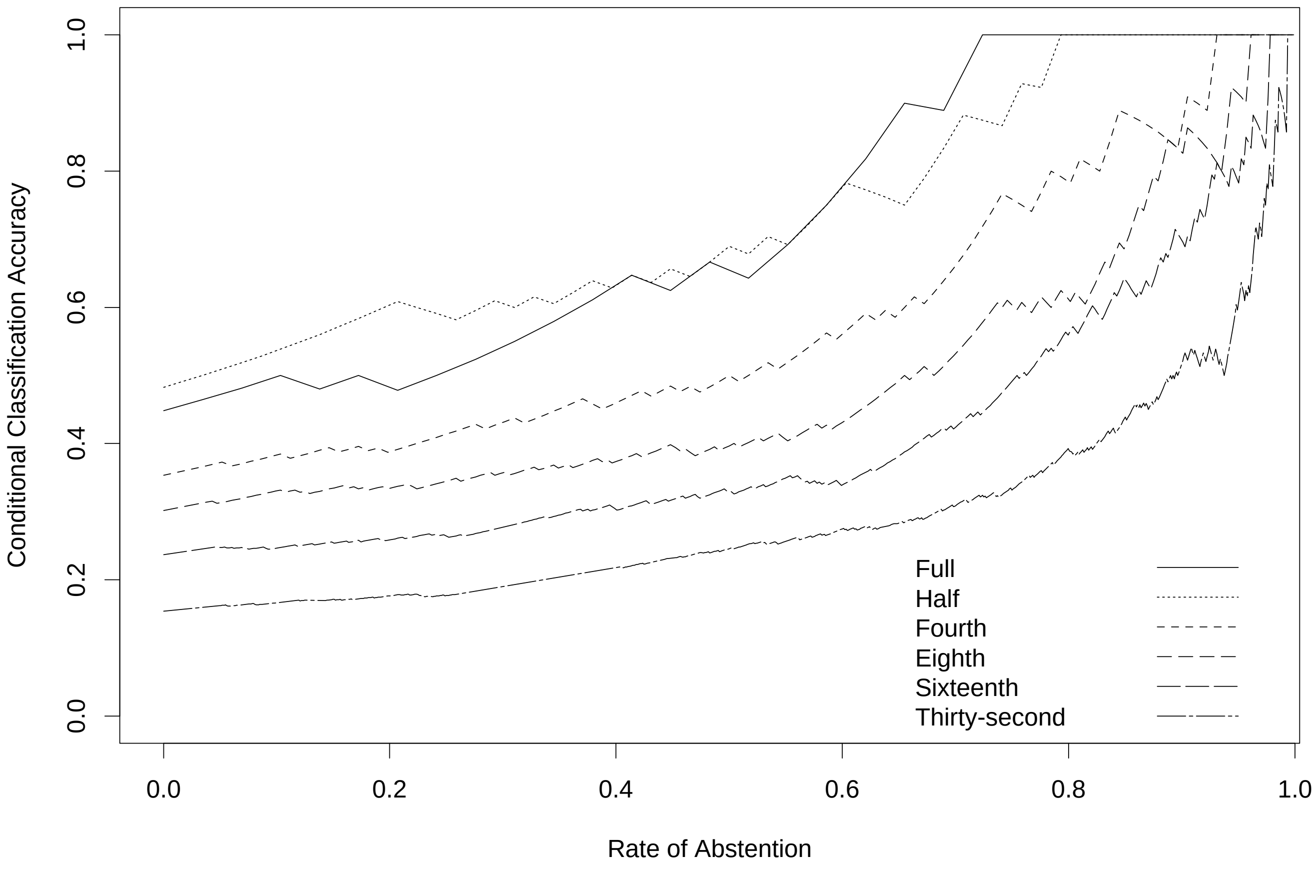
Incorporating Confidence Measures

- $\star$ γ: difference between two largest weights (margin)
- $\star$ A, B: number of correctly, incorrectly classified examples.
- $\star$ τ: a threshold on the margin—if $\gamma > \tau$, make a prediction; else abstain
- $\star$ a_τ, b_τ: number of correctly, incorrectly classified examples with $\gamma > \tau$

$$\frac{a_\tau}{a_\tau + b_\tau} \qquad \text{conditional classification accuracy}$$

$$1 - \frac{a_\tau + b_\tau}{A + B} \qquad \text{proportion of examples on which we abstain}$$

Conditional Accuracy vs. Abstention



Confusion Matrix

Our anytime classifier on eighth-length subdialogues.

Topics	Bks	Exer	Fam	Mov	Mus	Pets	Rec	Rest	Wthr	Sprt	% of Predictions
Books	—	0.062	0.120	0.110	0.083	0.062	0	0.042	0.120	0.062	0.078
Exercise	—	**0.310**	0.380	0.036	0.083	0.062	0.170	0	0.062	0.120	0.110
Family	—	0.190	**0.120**	0.140	0.120	0.062	0.083	0.120	0.031	0.310	0.130
Movies	—	0.031	0.120	**0.250**	0.083	0	0	0.120	0.062	0.062	0.100
Music	—	0.031	0	0.071	**0.290**	0.062	0	0	0	0	0.056
Pets	—	0.220	0.120	0.110	0.250	**0.560**	0	0	0.094	0.120	0.150
Recipes	—	0.031	0.120	0.089	0.042	0	**0.540**	0.120	0.094	0	0.120
Restaurants	—	0.031	0	0.054	0.042	0.120	0.042	**0.210**	0.062	0.062	0.069
Weather	—	0.031	0	0.036	0	0	0.083	0	**0.220**	0	0.052
Sports	—	0.062	0	0.110	0	0.062	0.083	0.380	0.250	**0.250**	0.140
Counts	0	32	8	56	24	16	24	24	32	16	232

Conclusions

⋆ Established a baseline for the subdialogue problem

⋆ Found a competitive "anytime" classifier

⋆ Sure, we could do even better with sophisticated ASR...

⋆ ...but things look good for our "acoustical nightmare"!

Word Error Rate

$$\mathrm{WER} = \frac{\#\ \text{insertions} + \#\ \text{deletions} + \#\ \text{substitutions}}{\text{total number of words in input}}$$

Bayes Theorem in Speech Recognition

Given acoustic data A, what is the probability that the target word or phrase W was spoken?

$$\Pr(W|A) = \frac{\Pr(W)\Pr(A|W)}{\Pr(A)}, \text{ where}$$

- $\star$ $\Pr(W)$ is the **language model**
- $\star$ $\Pr(A|W)$ is the **acoustic model**
- $\star$ $\Pr(A)$ is the average probability that A will be observed (constant)

Find $\hat{W}$ such that

$$\hat{W} = \arg\max_{W} \Pr(W)\Pr(A|W)$$