Max. Expected F-Measure Training of Logistic Regression Models

Martin Jansche
jansche@acm.org

Center for Computational Learning Systems
Columbia University

1 Introduction

Log-linear models are widely used in speech and language processing and IR, sometimes under the guise of "maximum entropy". When used for binary classification, they are known as logistic regression models.

In information extraction or retrieval tasks (e.g. Ittycheriah et al. 2003, Greiff and Ponte 2000), logistic regression classifiers are evaluated in terms of precision, recall, and F-measure.

Classifier training should be informed by the evaluation criterion. This paper describes a procedure that maximizes expected F-measure.

2 Logistic Regression

Binary response variable Y over $\{-1, +1\}$. Vector $\vec{X} = (X_1, \ldots, X_k)$ of k explanatory variables.

$$Y \sim \mathrm{Bernoulli}(p)$$

$$\text{i.e. } \Pr(Y = +1 \mid \vec{X} = (x_1, \ldots, x_k), \vec{\theta}) = p$$

where $\mathrm{logit}(p) = \theta_0 + x_1\,\theta_1 + \cdots + x_k\,\theta_k$

let $\vec{x} = (1, x_1, \ldots, x_k)$

then $\Pr(+1 \mid \vec{x}, \vec{\theta}) = \dfrac{1}{1 + \exp(-\vec{x} \cdot \vec{\theta})}$

Here, $\vec{\theta}$ is a $k+1$-dimensional vector of parameters.

3 F-Measure

Maximum *a posteriori* (MAP) decision rule:

$$y_{\mathrm{map}}(\vec{x} \mid \vec{\theta}) = \operatorname*{argmax}_{y} \Pr(y \mid \vec{x}, \vec{\theta}) = \operatorname{sgn}\left(\vec{x} \cdot \vec{\theta}\right)$$

		y_{map} $+1$	y_{map} -1	total
true	$+1$	A	B	n_{pos}
true	-1	C	D	n_{neg}
total		m_{pos}	m_{neg}	n

Recall/sensitivity:

$$R = \frac{A}{A+B}$$

Precision:

$$P = \frac{A}{A+C}$$

$$F_\alpha(R, P) = \left(\alpha \frac{1}{R} + (1-\alpha)\frac{1}{P}\right)^{-1}$$

Let $[\![\phi]\!] = 1$ if Boolean expression ϕ is true, and 0 otherwise. Given an evaluation data set $(\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n)$:

$$\text{hits } A(\vec{\theta}) = \sum_{i=1}^{n} \left[\!\left[y_{\text{map}}(\vec{x}_i \mid \vec{\theta}) = +1 \right]\!\right] [\![y_i = +1]\!]$$

$$\text{misses } B(\vec{\theta}) = \sum_{i=1}^{n} \left[\!\left[y_{\text{map}}(\vec{x}_i \mid \vec{\theta}) = -1 \right]\!\right] [\![y_i = +1]\!]$$

$$\text{false alarms } C(\vec{\theta}) = \sum_{i=1}^{n} \left[\!\left[y_{\text{map}}(\vec{x}_i \mid \vec{\theta}) = +1 \right]\!\right] [\![y_i = -1]\!]$$

$$F_\alpha(\vec{\theta}) = \frac{A(\vec{\theta})}{\alpha\,[A(\vec{\theta}) + B(\vec{\theta})] + (1-\alpha)\,[A(\vec{\theta}) + C(\vec{\theta})]}$$

4 Relation to Expected Utility

It exists. See the paper for details.

5 Discriminative Estimation

The goal is to estimate $\vec{\theta}$ as

$$\vec{\theta}^{\star} = \underset{\vec{\theta}}{\operatorname{argmax}}\, F_{\alpha}(\vec{\theta})$$

Problem: $F_{\alpha}(\vec{\theta})$ is defined in terms of $A(\vec{\theta})$ (etc.), which depends on $\vec{\theta}$ via the step function $[\![\cdot]\!]$ (Kronecker delta). This means that the gradient of $F_{\alpha}(\vec{\theta})$ is zero almost everywhere.

The key idea of our solution is to replace the discontinuous step function $[\![\cdot]\!]$ in

$$\left[\!\!\left[y_{\mathrm{map}}(\vec{x}_i \mid \vec{\theta}) = +1 \right]\!\!\right] = \left[\!\!\left[\Pr(+1 \mid \vec{x}, \vec{\theta}) > 0.5 \right]\!\!\right]$$

with a continuous approximation

$$\left[\!\!\left[\Pr(+1 \mid \vec{x}, \vec{\theta}) > 0.5 \right]\!\!\right] \approx \Pr(+1 \mid \vec{x}, \vec{\theta})$$

In the case of logistic regression, this amounts to approximating the limit

$$\lim_{\gamma \to \infty} \frac{1}{1 + \exp(-\gamma\, \vec{x} \cdot \vec{\theta})} = \left[\!\!\left[\Pr(+1 \mid \vec{x}, \vec{\theta}) > 0.5 \right]\!\!\right]$$

with a term where $\gamma = 1$ (different values of γ could be used as well).

In particular, approximate

$$A(\vec{\theta}) \approx \tilde{A}(\vec{\theta}) = \sum_{\substack{i=1\\ y_i=+1}}^{n} \frac{1}{1+\exp(-\vec{x}_i \cdot \vec{\theta})} \quad \text{(expected hits)}$$

and

$$A(\vec{\theta}) + C(\vec{\theta}) = m_{\text{pos}}(\vec{\theta}) \approx \tilde{m}_{\text{pos}}(\vec{\theta}) = \sum_{i=1}^{n} \frac{1}{1+\exp(-\vec{x}_i \cdot \vec{\theta})}$$

to obtain the relaxed optimization objective:

$$\tilde{F}_\alpha(\vec{\theta}) = \frac{\tilde{A}(\vec{\theta})}{\alpha\, n_{\text{pos}} + (1-\alpha)\, \tilde{m}_{\text{pos}}(\vec{\theta})}$$

Maximization of $\tilde{F}_\alpha$ as can be carried out numerically using conjugate gradient search or quasi-Newton methods such as the BFGS algorithm. This requires the evaluation of partial derivatives.

One can compute the value of $\tilde{F}_\alpha(\vec{\theta})$ and its gradient $\nabla\tilde{F}_\alpha(\vec{\theta})$ simultaneously at a given point $\vec{\theta}$ in $\mathrm{O}(n\,k)$ time and $\mathrm{O}(k)$ space. Pseudo-code for such an algorithm and formulas for the gradient can be found in the paper.

6 Comparison with MLE

A graphical comparison with maximum likelihood estimation (MLE) is instructive. Consider the toy dataset shown on the right. The logistic regression model simplifies to:

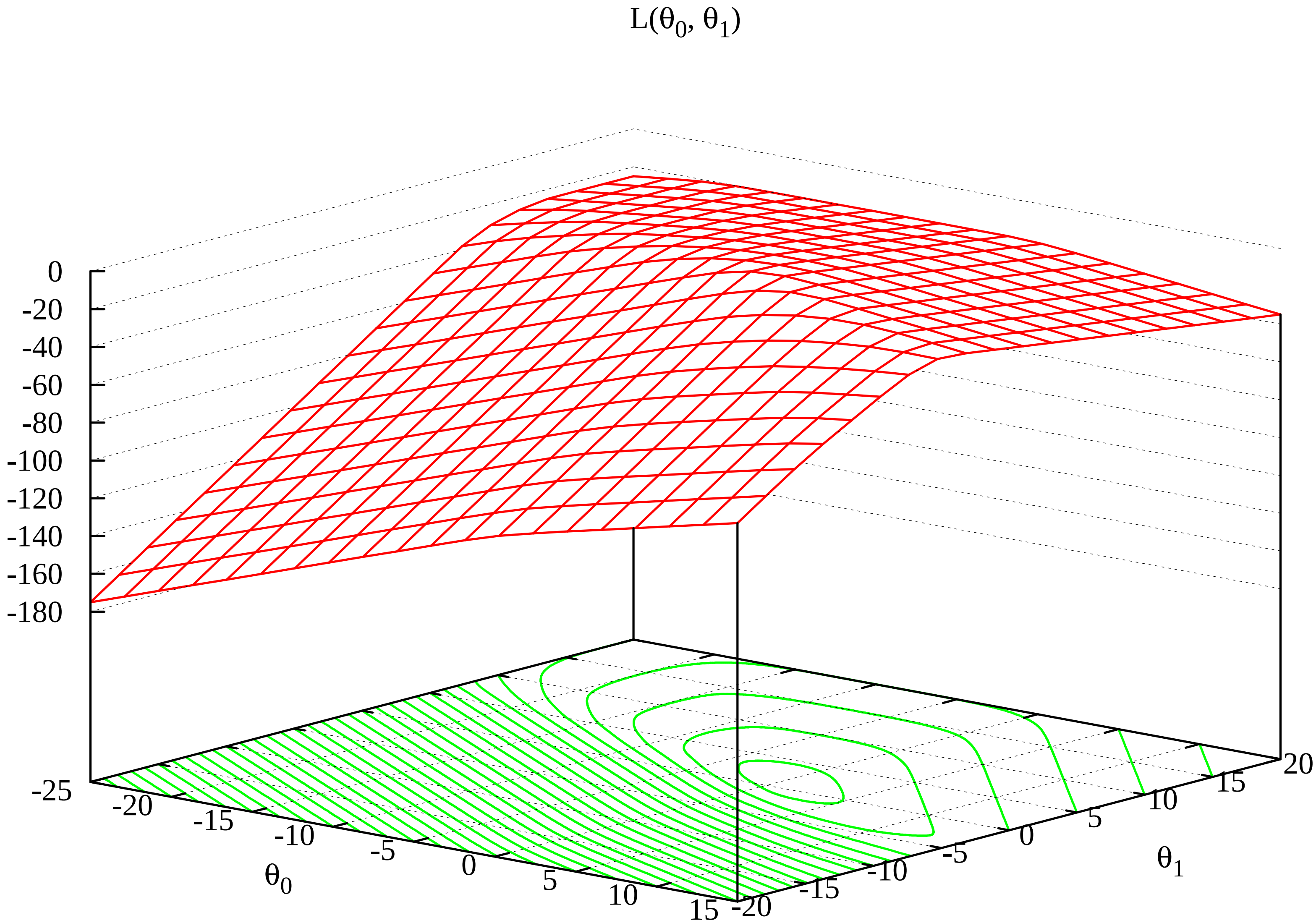
x	y
0	$+1$
1	-1
2	$+1$
3	$+1$

$$\Pr(+1 \mid x, \theta_0, \theta_1) = \frac{1}{1 + \exp(-\theta_0 - x\,\theta_1)}$$

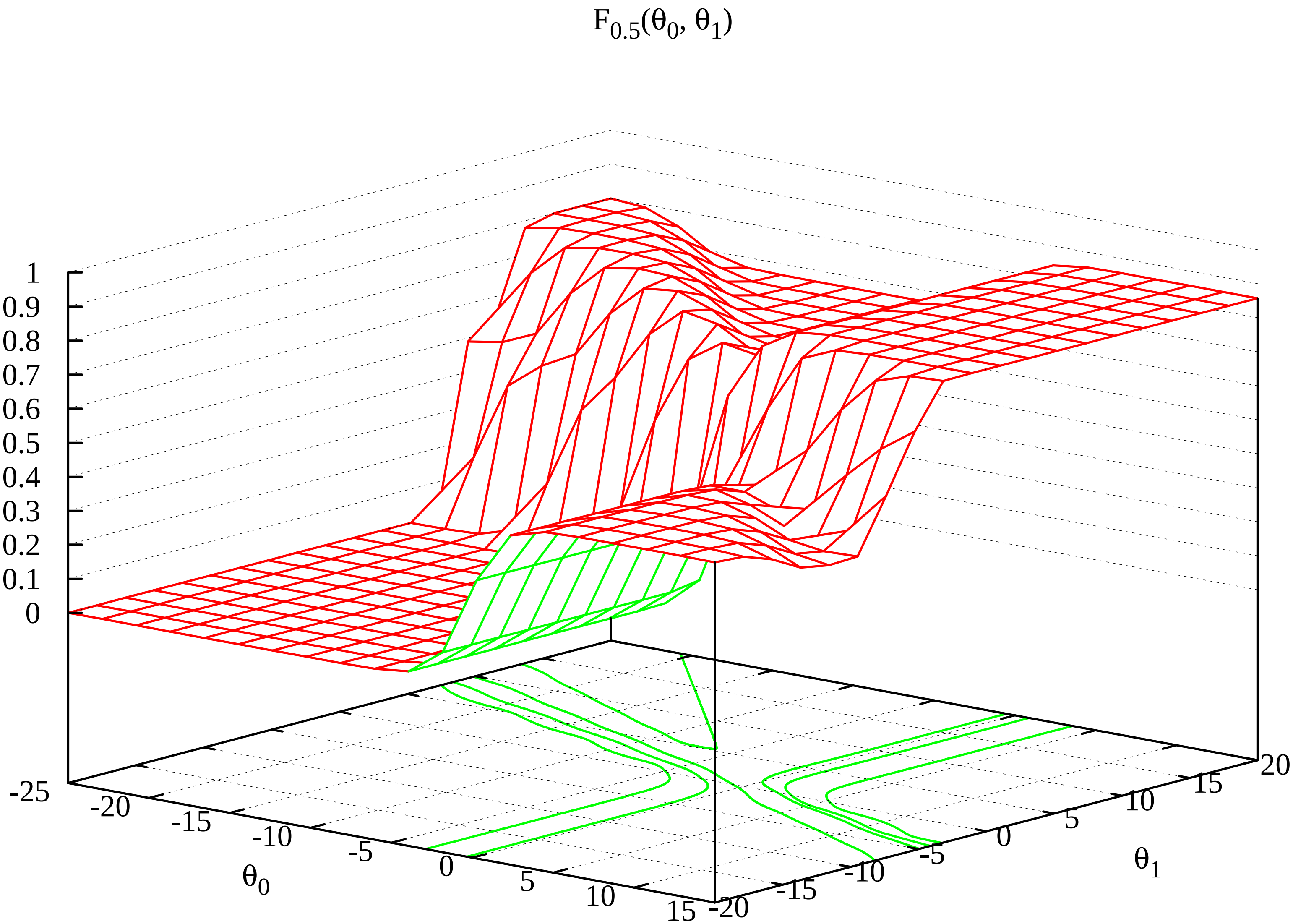
The log-likelihood function L is

$$\begin{aligned} L(\theta_0, \theta_1) = \log \Pr(+1 \mid 0, \theta_0, \theta_1) + \log \Pr(-1 \mid 1, \theta_0, \theta_1) \\ + \log \Pr(+1 \mid 2, \theta_0, \theta_1) + \log \Pr(+1 \mid 3, \theta_0, \theta_1) \end{aligned}$$

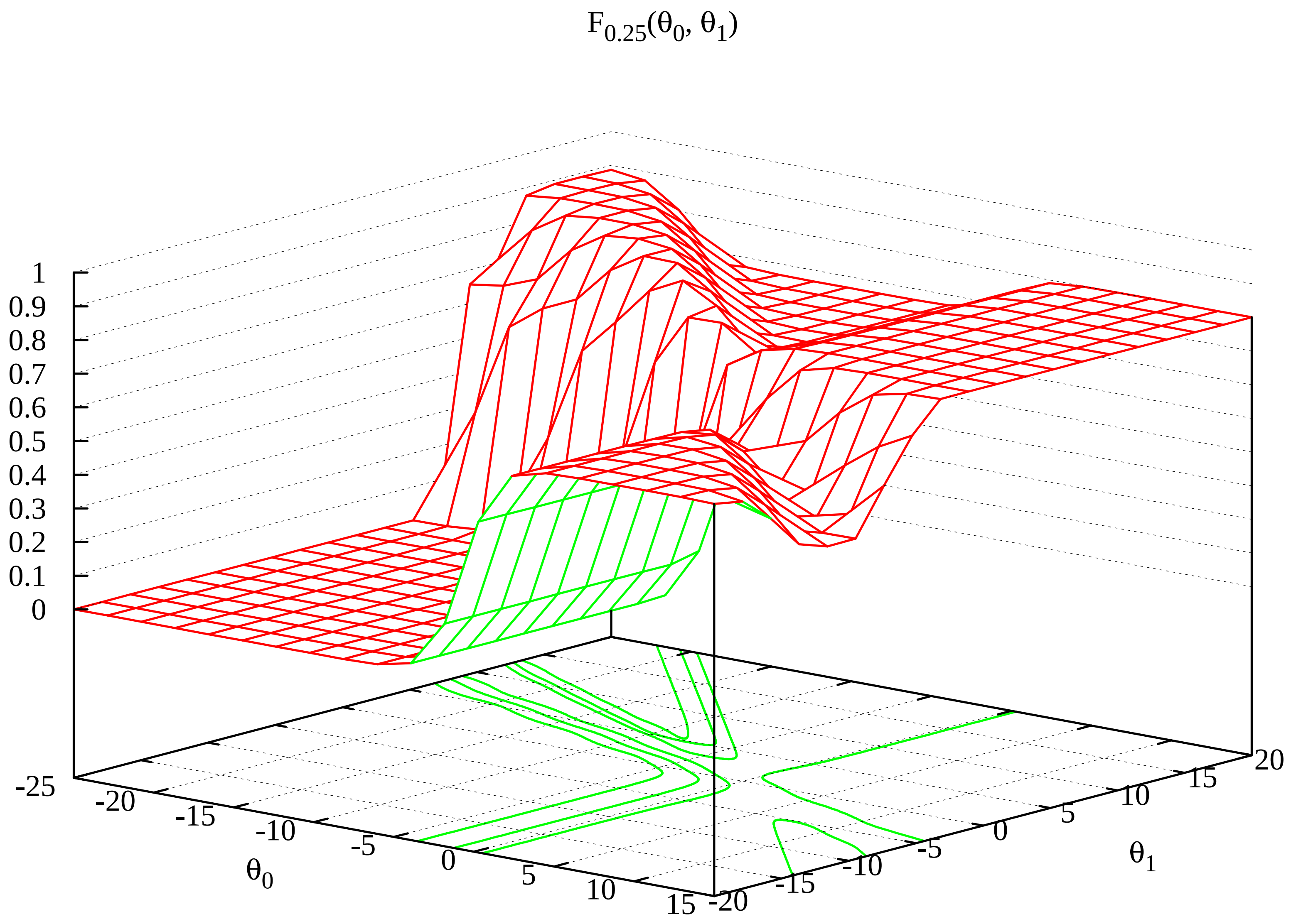
Log-likelihood is a concave function:



But $\tilde{F}_\alpha$ is, in general, not concave:



Notice that $\alpha = 0.25$ gives a different maximum:



7 Evaluation

Evaluation on a speech summarization task (Maskey & Hirschberg 2005): extractive summarization by classifying each sentence of a broadcast with an *include* or *exclude* label. Dataset with 29 mostly integer- or real-valued explanatory variables. Trained on 3,535 instances, evaluated on 408 instances.

Method	R	P	$F_{\alpha=0.5}$
MLE	24/99	24/33	0.3636
$\tilde{F}_{\alpha=0.5}$	85/99	85/211	0.5484

Further details can be found in the paper.

8 Conclusions

This presentation describes discriminative training of logistic regression classifiers by maximizing a relaxed version of F-measure expressed in terms of the expectations of hits, misses, and false alarms. The assumption about the class of models (logistic regression) is not crucial: the same technique applies to many other kinds of models.

Maximizing F-measure during training seems especially well suited for dealing with skewed classes, where predicting the majority class would result in high accuracy, but low F-measure.

Acknowledgements

This work was partly supported by Columbia University's Fu Foundation School of Engineering and Applied Science. All opinions expressed here are those of the author.

I would like to thank Julia Hirschberg, Phil Long, Sameer Maskey, and the three anonymous reviewers for helpful comments and discussions. I am especially grateful to Sameer Maskey for allowing me to use his speech summarization dataset. The usual disclaimers apply.