

**Physiological Data Modeling Contest
July 8, 2004, Banff, CANADA**

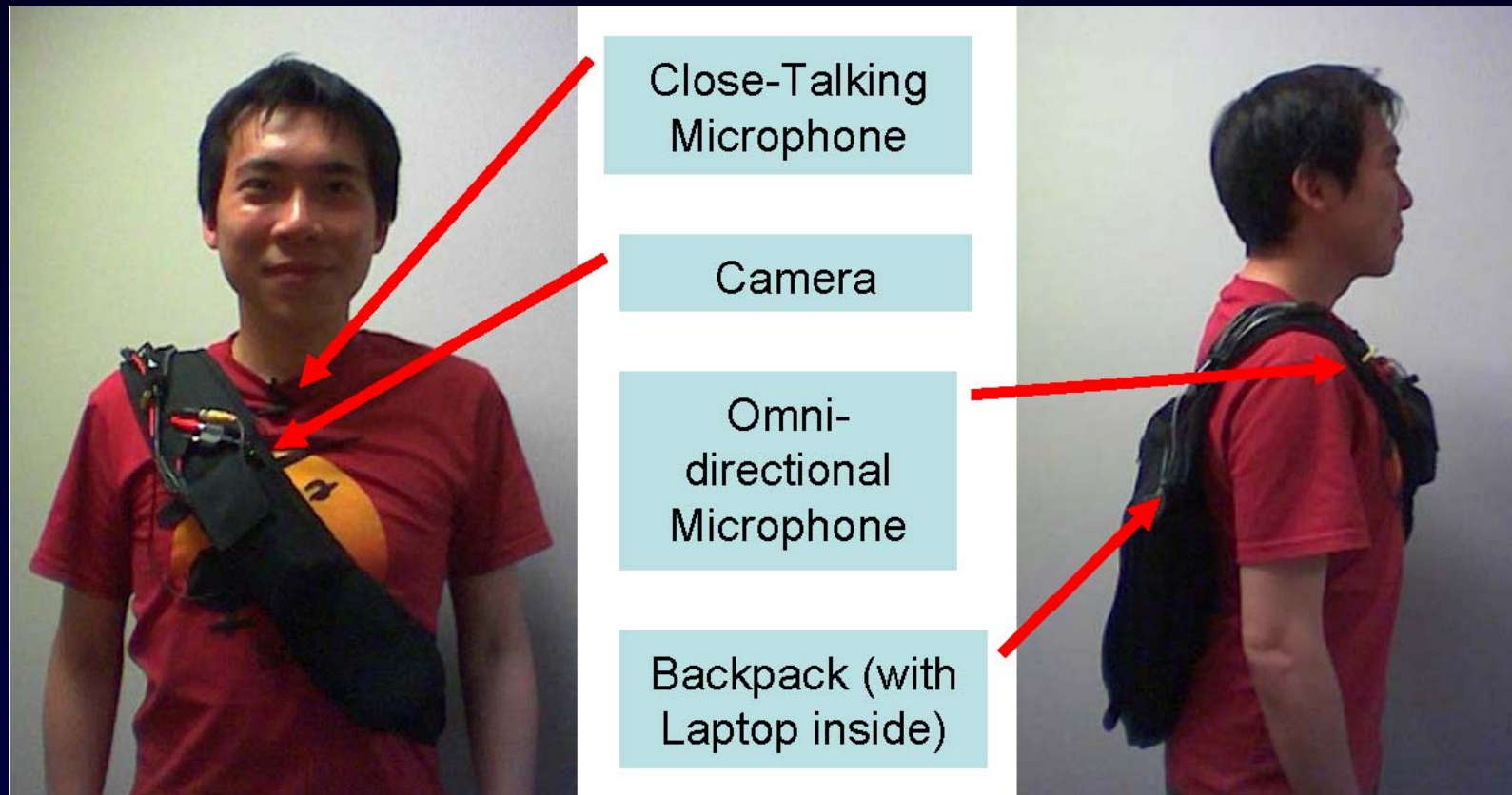
Informedia at PDMC 2004

Wei-Hao Lin and Alexander Hauptmann
Language Technologies Institute
School of Computer Science
Carnegie Mellon University

Digital Human Memory

- Record and index multiple aspects of daily human experiences in digital form
 - Visual experiences from a spy camera
 - Auditory experiences from microphones
 - Physiological experiences from a BodyMedia armband
- Physiological readings play an important role in identifying user contexts, and thus facilitate indexing and retrieval of past events.

DHM Capturing Device



Informedia at PDMC 2004

- Build a baseline system based on Support Vector Machine (SVM)
- Disambiguate noisily-labeled and unlabeled data in the multiple-label framework
- Construct SVM-based conditional model to incorporate sequential information

Baseline System

- Treat both gender and context tasks as simple binary classification
 - Features x : 9 sensor readings + 2 characteristics
 - Labels y : unambiguous annotations
 - Gender: 0 vs. 1
 - Context 1: Positive (3004) vs. Negative (\neq 3004, 0, 3003, 5199, 5101)
 - Context 2: Positive (5102) vs. Negative (\neq 5102, 0, 5103, 2901, 2902)

Baseline Performance

- Classifier: Support Vector Machine with RBF kernel
- Performance of 10-fold cross validation on the training set

	Gender	Context 1	Context 2
SVM Baseline	0.9572	0.7548	0.8711
Random Baseline	0.5	0.7	0.7

Noisily-Labeled and Unlabeled Data

- Different types of annotations
 - Unambiguous
 - positive, negative
 - Ambiguous: could be positive or negative
 - Noisily Labeled
 - Context 1: 3003, 5199, 5101
 - Context 2: 5103, 2901, 2902
 - Unlabeled
 - annotation 0
 - ~70% of training data are unlabeled.
- Goal: utilize noisily-labeled and unlabeled data

Disambiguation Strategies

- Strategy 1: ambiguous examples are all assumed to be negative.
 - $P(y = \text{"pos"} \mid x) = 0.0$
 - $P(y = \text{"neg"} \mid x) = 1.0$
- Strategy 2: Ambiguous examples are randomly assigned as positive or negative with equal probabilities.
 - $P(y = \text{"pos"} \mid x) = 0.5$
 - $P(y = \text{"neg"} \mid x) = 0.5$
 - Equivalent to duplicate ambiguous data with opposite labels.

Multi-label Framework

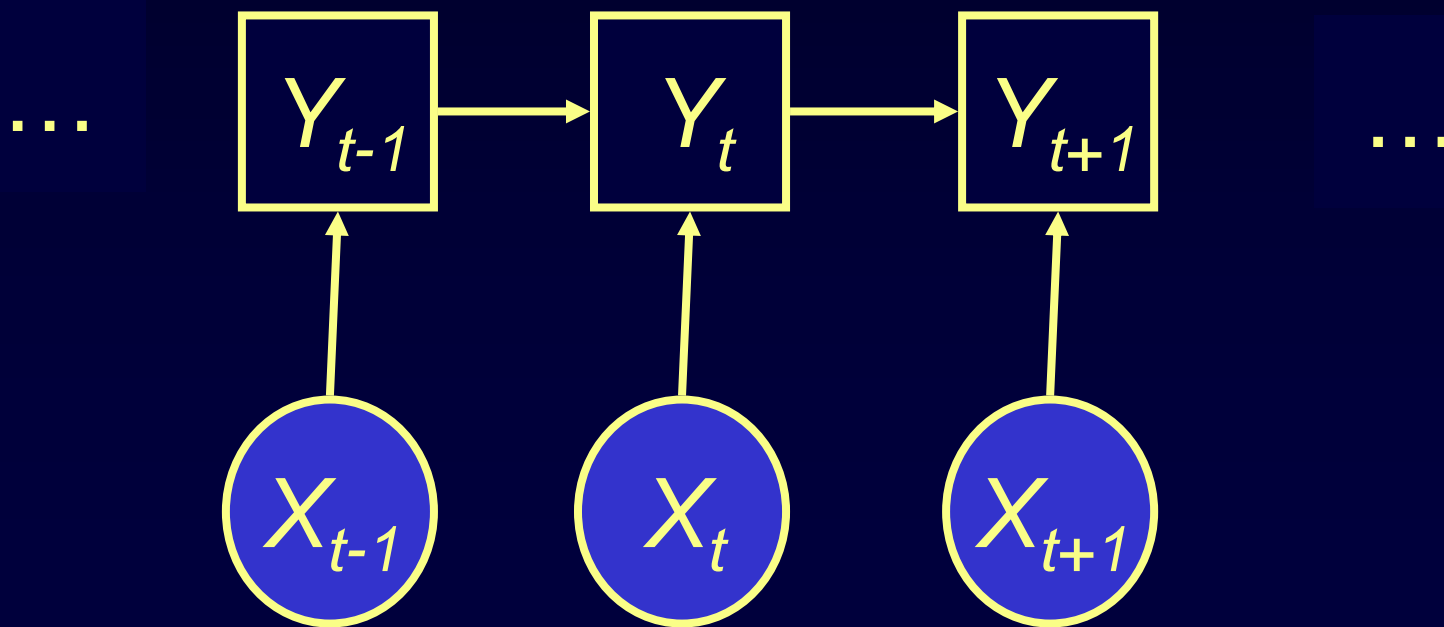
- Noisily-Labeled and unlabeled data can be seen as carrying both positive and negative annotations, i.e. multiple labels [Jin and Ghahramani 03], but only one of them is correct.
- Strategy 3: Iteratively estimate label distribution of ambiguous examples.
 - Estimate new label distribution $P(y^t | x)$ based on model M^{t-1}
 - Train a new model M^t based on $P(y^t | x)$

Disambiguation Performance

	Context 1	Context 2
Strategy 1	0.7625	0.8834
Strategy 2	0.6957	0.8559
Strategy 3	0.7613	0.8707
SVM Baseline	0.7548	0.8711
Random Baseline	0.7	0.7

SVM-based Conditional Models

- Aims to exploit sequential relationship
- Based on Maximum-Entropy Markov Models [McCallum, et. al. 00]
- Use SVM to construct $P(Y_t | X_t, Y_{t-1})$
- Viterbi-like decoding algorithm



Scarcity of Positive Examples

- Take context 1 as an example,

	$y_{t-1} = \text{neg}$	$y_{t-1} = \text{pos}$
$y_t = \text{neg}$	575776	75
$y_t = \text{pos}$	75	4338

- The model suffers greatly from the scarcity of positive sequences and perform poorly.

- Informedia

- <http://www.informedia.cs.cmu.edu/>