Using a Hidden-Markov Model in Semi-Automatic Indexing of Historical Handwritten Records

Thomas Packer, Oliver Nina, Ilya Raykhel

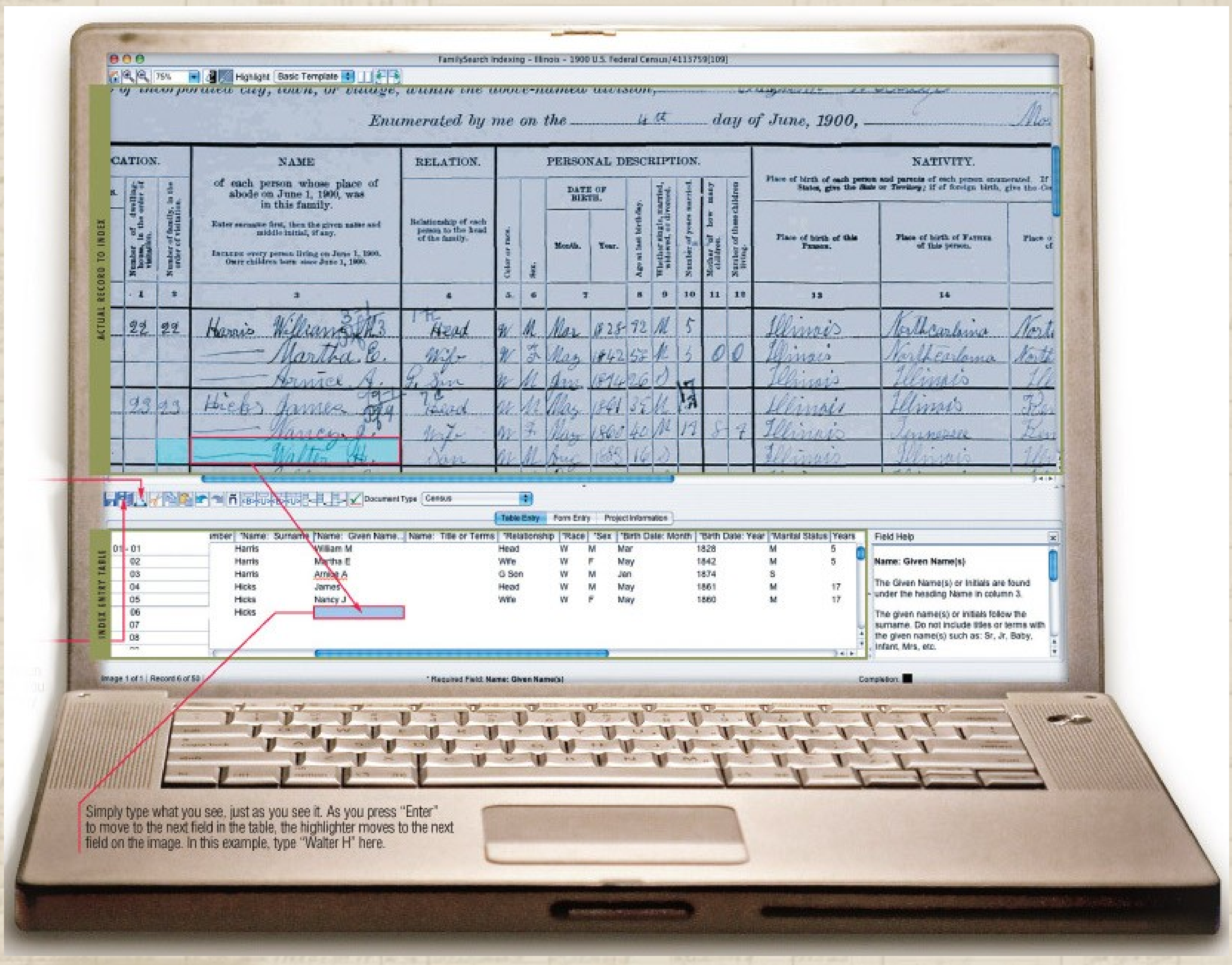
Computer Science

Brigham Young University

The Challenge: Indexing Handwriting

- Millions of historical documents.
- Many hours of manual indexing.
- Years to complete using hundreds of thousands of volunteers.
- Previous transcriptions not fully leveraged.

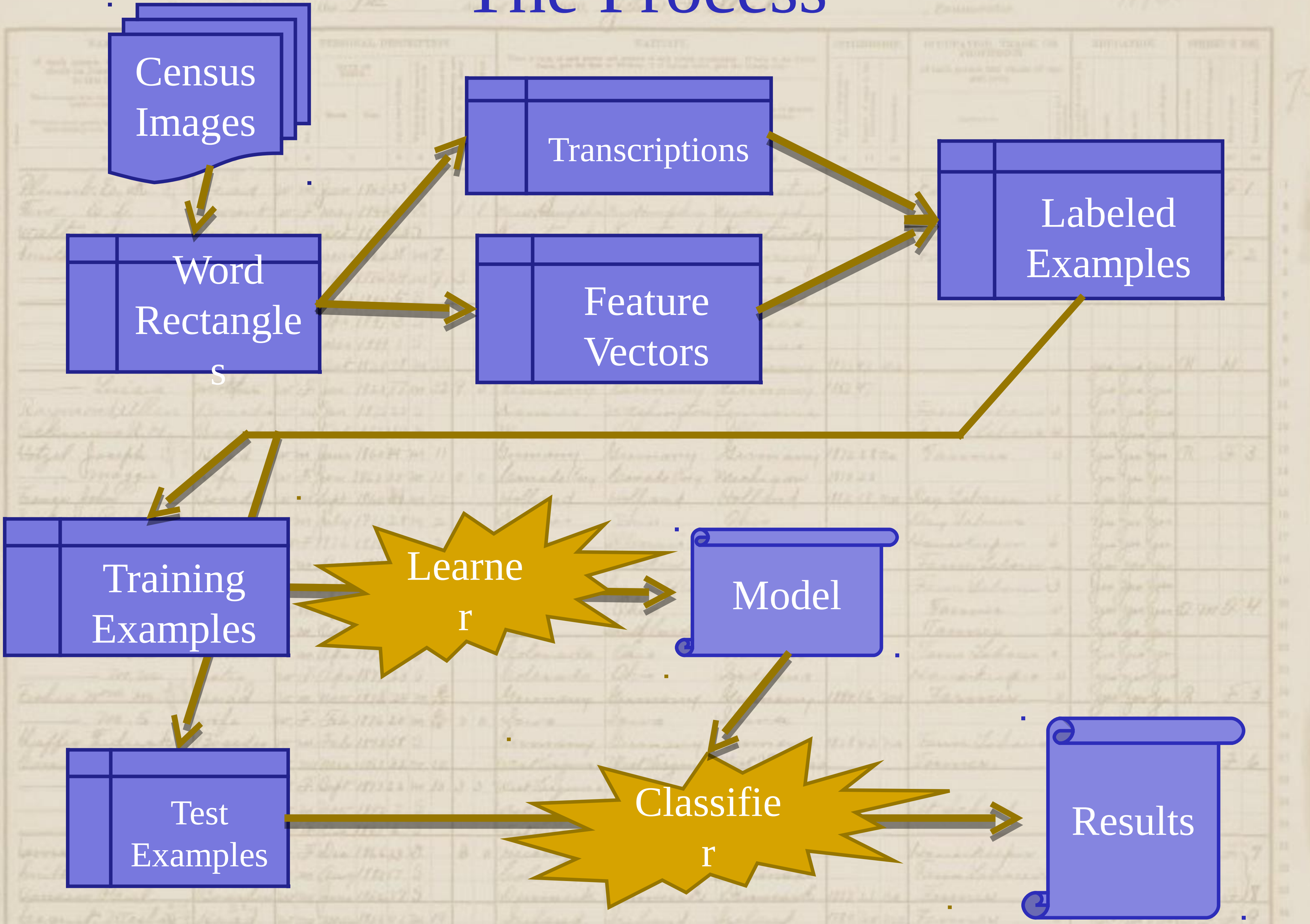
Family Search Indexing Tool



A Solution: On-Line Machine Learning

- Holistic handwritten word recognition using a Hidden Markov Model (HMM), based on Lavrenko et al. (2004).
- HMM selects words to maximize joint probability:
  - Word-feature probability model
  - Word-transition probability model
- Word-feature model predicts a word from its visual features.
- Word-transition model predicts a word from its neighboring word.

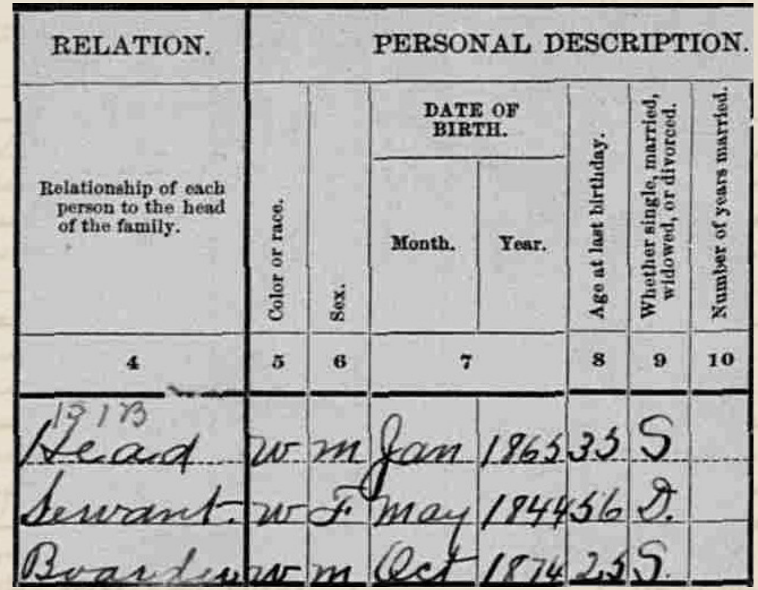
The Process

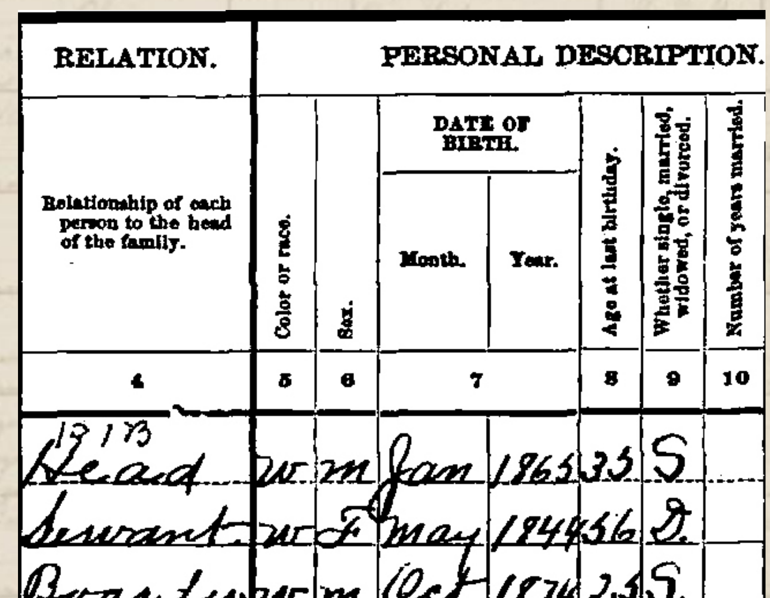


Census Images
Transcriptions
Labeled Examples
Word Rectangle s
Feature Vectors
Training Examples
Learne r
Model
Test Examples
Classifie r
Results

Census Images

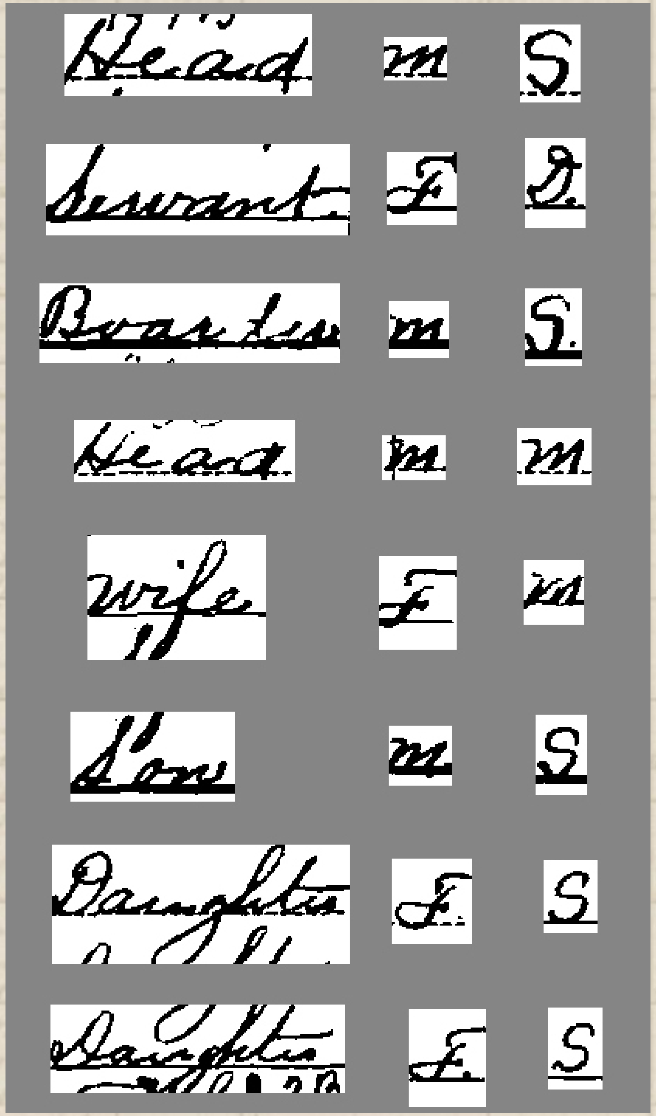
- 3 US Census images
- Same census taker
- Preprocessing: Kittler's algorithm to threshold images





Extracted Fields

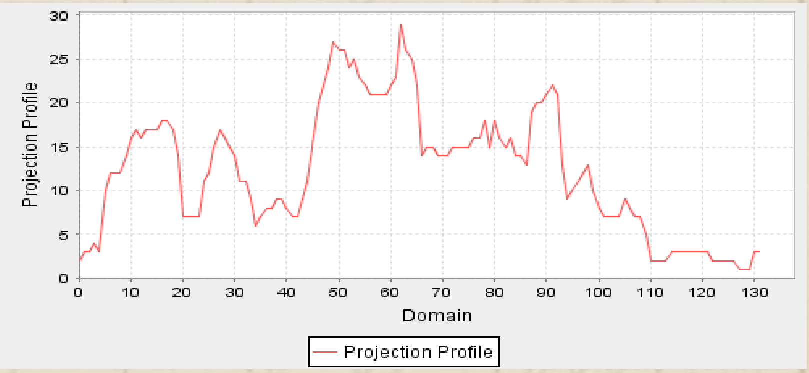
- Manually copied bounding rectangles
- 3 columns:
  1. Relationship to Head (14)
  2. Sex (2)
  3. Marital Status (4)
- 123 rows total
- N-fold cross validation
- N = 24 (5 rows to test)

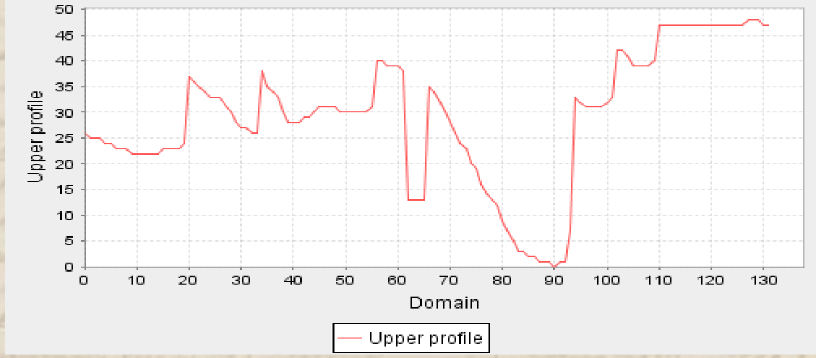


Examples to Feature Vectors

25 Numeric Features Extracted:

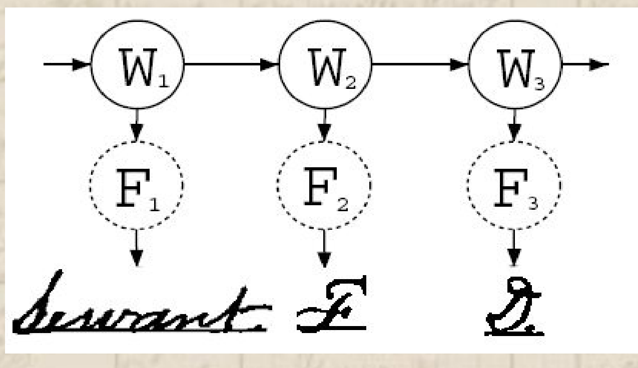
- Scalar Features:
  - height (h)
  - width (w)
  - aspect ratio (w / h)
  - area ($w * h$)
- Profile Features:
  - projection profile
  - upper/lower word profile
  - 7 lowest scalar values from DFT





HMM and Transition Probability Model

- **Probability Model:**
  - Hidden Markov Model



$$P(w_1 \ldots w_n, I_1 \ldots I_n) = \prod_{j=1}^{n} P(w_j|w_{j-1}) p(f_j|w_j)$$

  - State Transition Probabilities

$$P_T(w|w_0) = \left(\frac{\text{number of times } w \text{ occurs in } T}{\text{total number of words in } T}\right)$$

$$P_T(w|v) = \left(\frac{\text{number of times } v, w \text{ occurs in } T}{\text{number of times } v \text{ occurs in } T}\right)$$

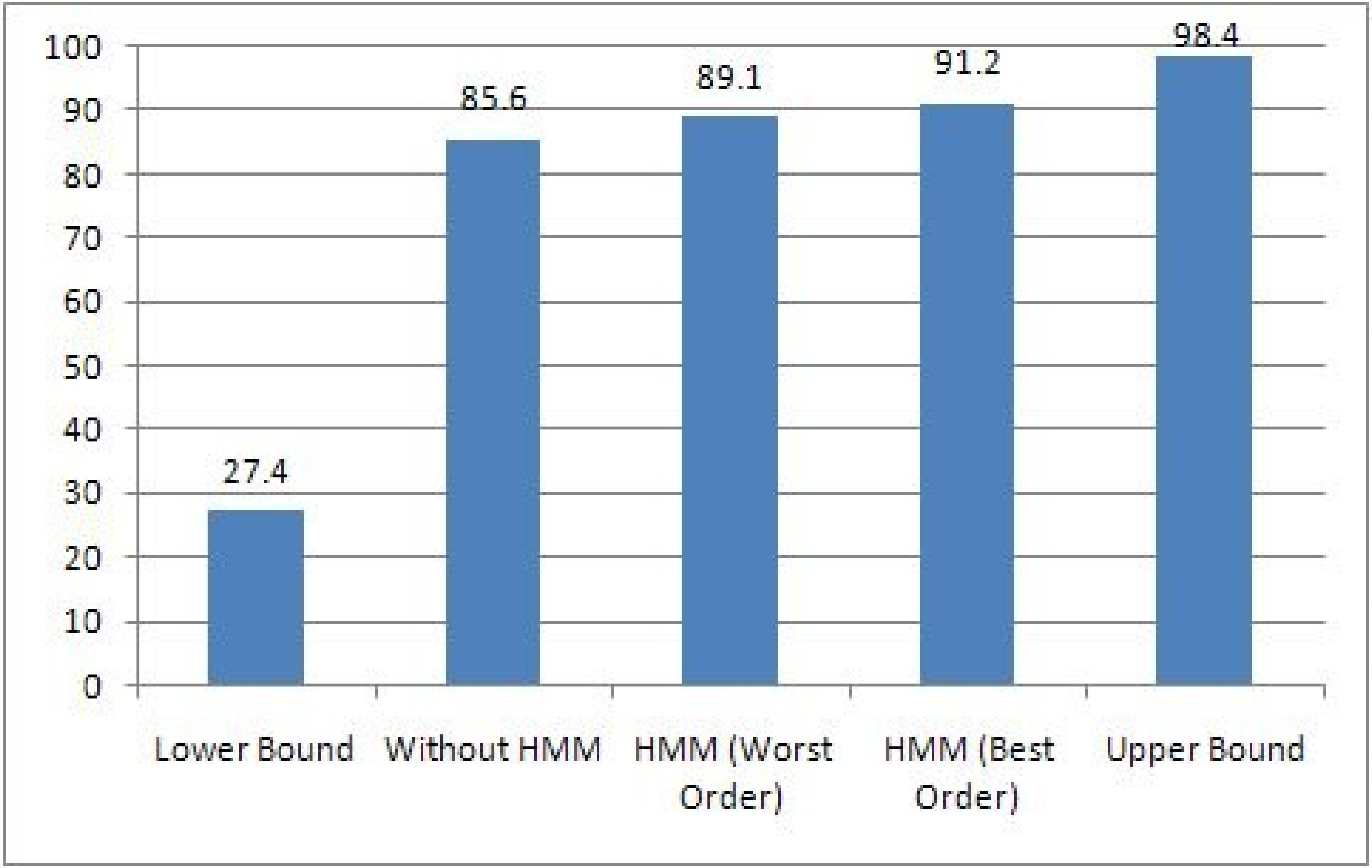
Observation Probability Model

- Multi-variate normal distribution:

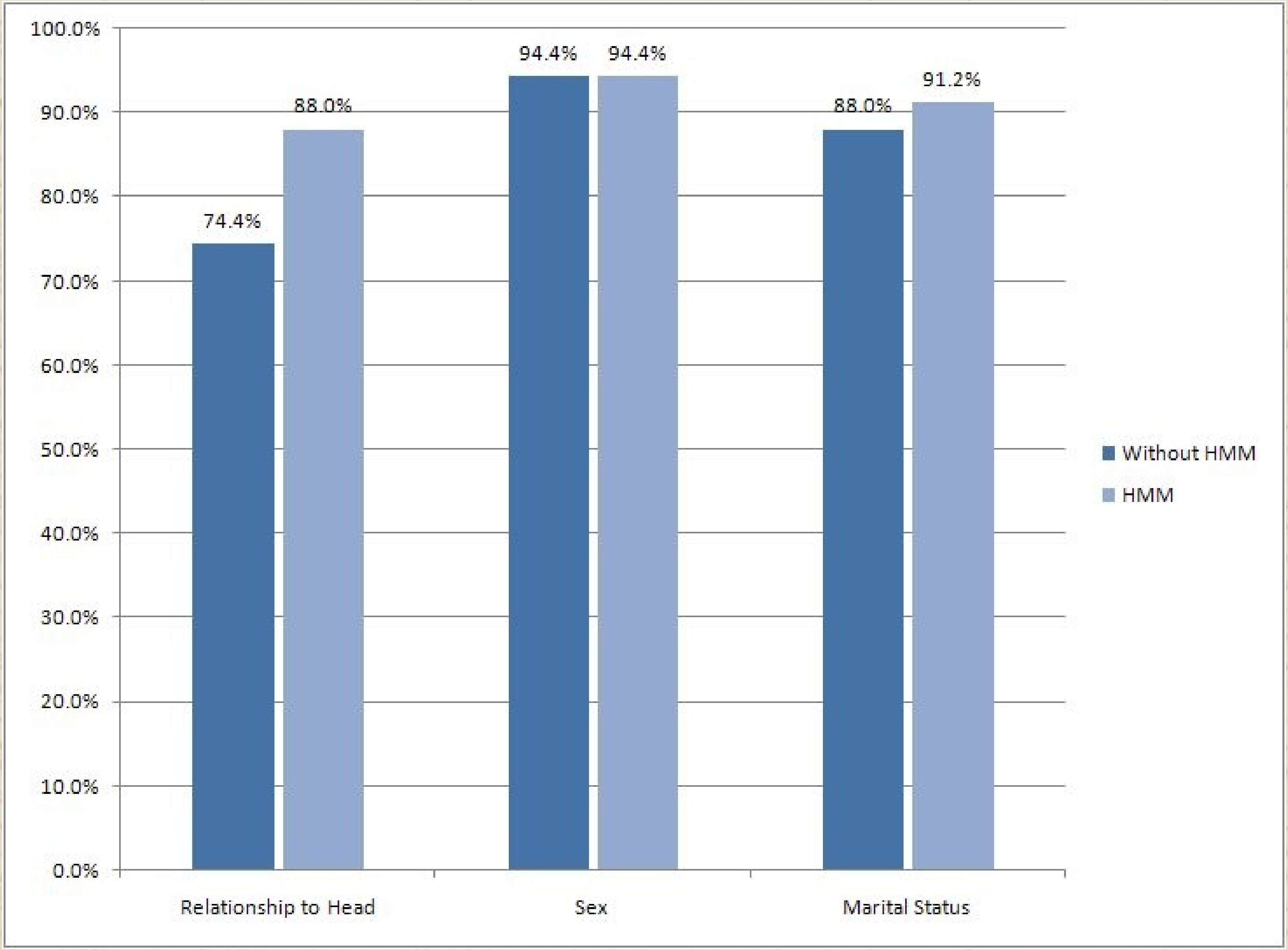
$$p(f|w) = \frac{\exp\left\{(f-\mu_w)^\top \Sigma_w^{-1}(f-\mu_w)\right\}}{\sqrt{2^D \pi^D |\Sigma_w|}}$$

$$p(\mathbf{x}|\omega_i) = (2\pi)^{-d/2}\,|\mathbf{C}_i|^{-1/2}\; e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T \mathbf{C}_i^{-1}(\mathbf{x}-\mathbf{m}_i)}$$

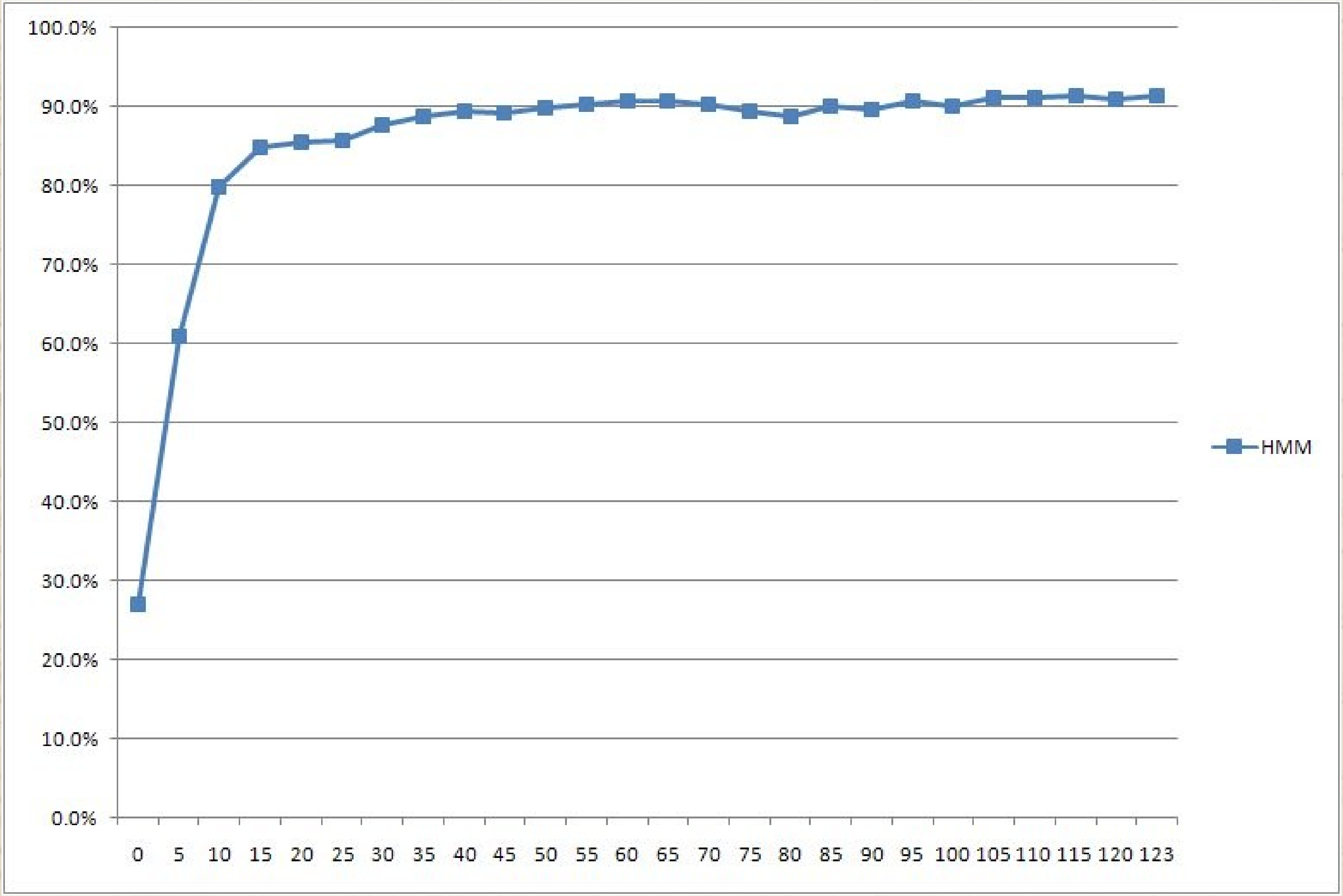
Accuracies with and without HMM



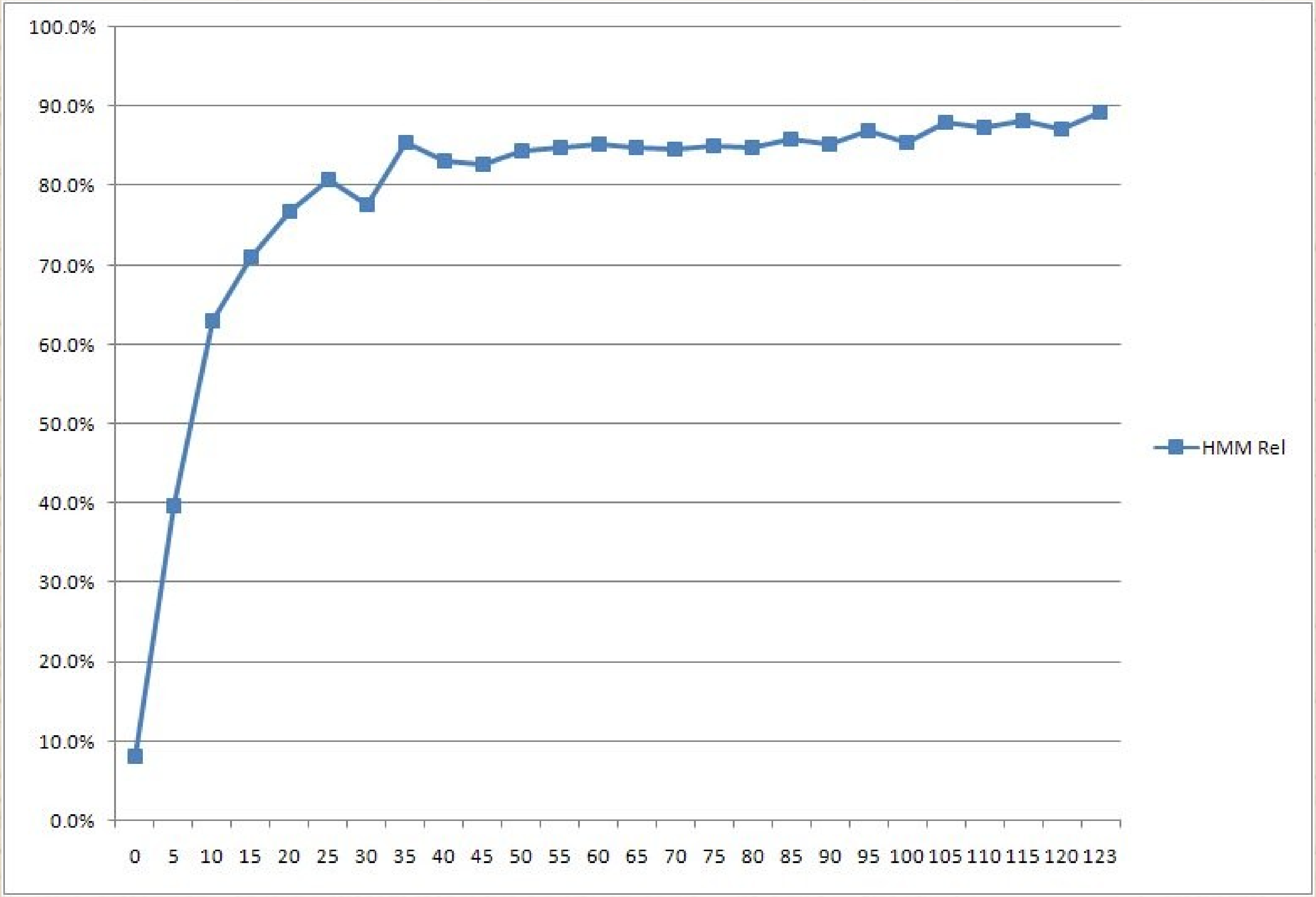
Accuracies for Separate Columns with and without HMM



Accuracies of HMM for Varying Numbers of Training Examples



Accuracies of “Relationship to Head” for Varying Numbers of Examples



Conclusions and Future Work

- 10% correction rate for chosen columns after one page.
- Measure indexing time.
- Update models in real-time.
- Columns with larger vocabularies.
- More image preprocessing.
- More visual features.
- More dependencies among words (in different rows).
- More training data.

Questions?