Lexicon Reduction in Handwriting Recognition Using Topic Categorization

Faisal Farooq

ffarooq2@cedar.buffalo.edu

Outline

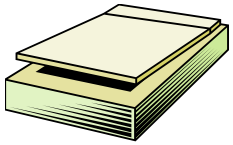
- Problem Statement
- Proposed Approach
- Results
- Conclusions

Outline

- Problem Statement
- Proposed Approach
- Results
- Conclusions

Problem Definition

Handwritten Documents



Images

OCR System

Raw Recognition Results

?

Improved Recognition Results

Lexicon Dependence

- Lexicon – restricted vocabulary of words
- Dependence on lexicon
  - Availability of lexicon
  - Size of lexicon

Lex. Size	10	100	1000	10000
Accuracy %	96.80	88.23	68.70	32.22

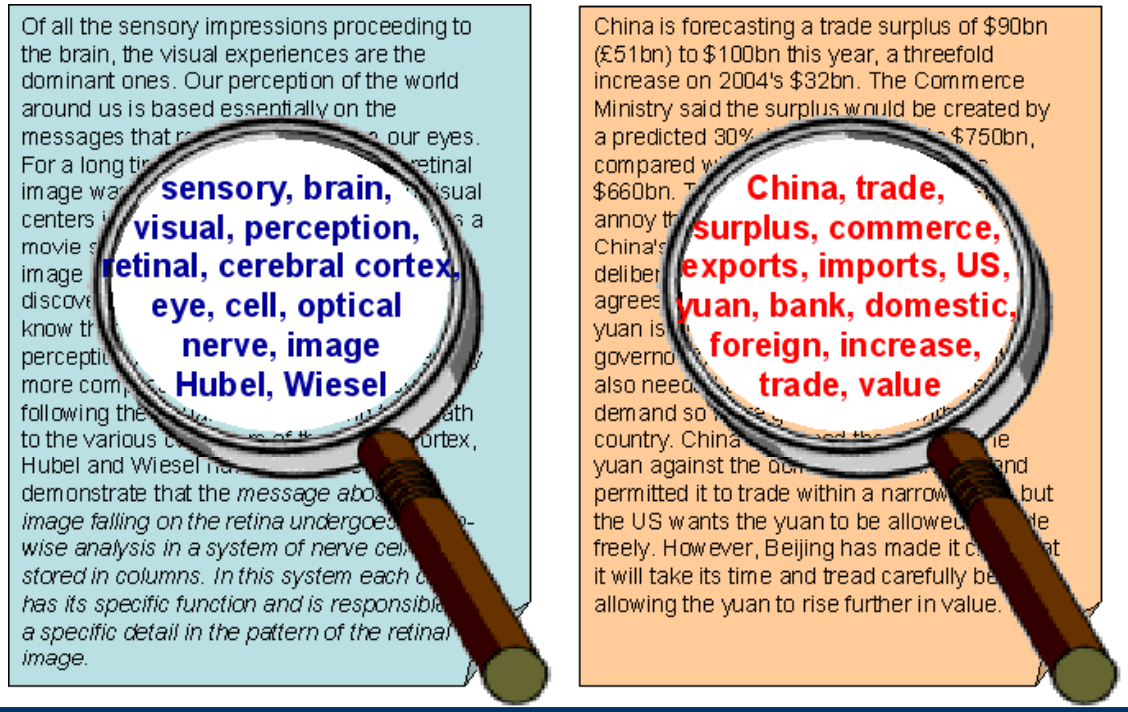
Effect of Lexicon Size on recognition accuracies [Results on IAM]

Outline

- Problem Statement
- Proposed Approach
- Results
- Conclusions

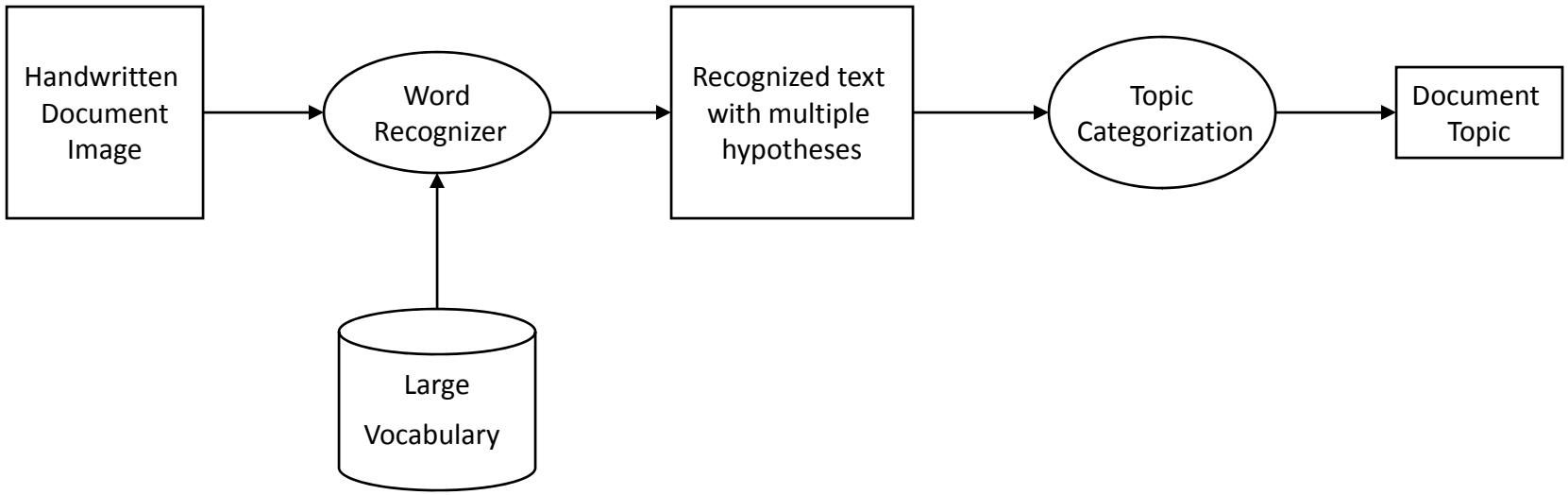
Topic Models

- Document consists of topic(s)
- Words used describe the topic
- Bag of words model [Rijsbergen 1979], latent dirichlet allocation [Blei et al 2003]



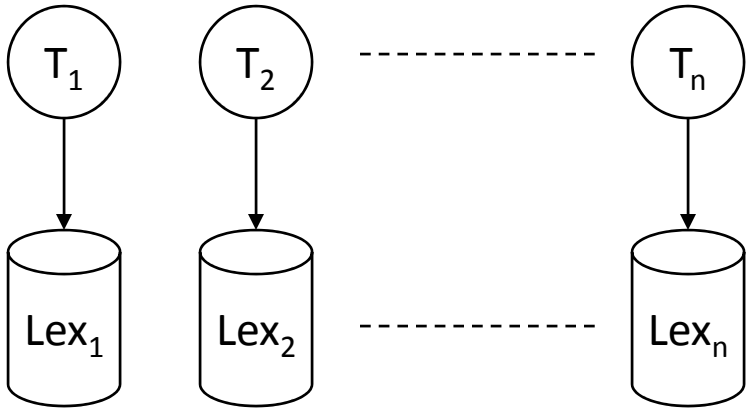
Topic Categorization

- Categorize a document into pre-defined classes based on raw results



Lexicon Reduction

- Current techniques word length [Madhvanath et al 1996,1999,2003] or word shape [Seni et al 1994, Leroy 1996, Hennig et al 2001] based
- Every topic has its key vocabulary
  - General - RELIGION , SPORTS, HOBBIES, POPULAR LORE, BIOGRAPHIES, SCIENTIFIC, MYSTERY
  - Medical - IMMUNE-SYSTEM, CIRCULATORY-CARDIOVASCULAR-SYSTEM, NERVOUS-SYSTEM, ENDOCRINE-SYSTEM
- Lexicon is *generated* by a topic



Naïve Bayes

- Document d_i assigned to category c_j

$$j_0 = \arg\max_{j=1,\ldots,m} P(c_j \mid d_i, \theta) = \arg\max_{j=1,\ldots,m} P(c_j \mid \theta) P(d_i \mid c_j, \theta)$$

- Class prior

$$P(c_j \mid \theta) = \frac{\sum_{i=1}^{|D|} P(c_j \mid d_i)}{|D|}$$

- Naïve independence assumption

$$P(d_i \mid c_j, \theta) = \prod_{k=1}^{|d_i|} P(w_k \mid c_j, \theta)$$

Naïve Bayes – Setting I

- Generative Model – Multivariate Bernoulli

$$P(w_t \mid c_j, \theta) = \frac{1 + \sum_{i=1}^{|D|} B_{it} P(c_j \mid d_i)}{2 + \sum_{i=1}^{|D|} P(c_j \mid d_i)}, B_{it} = \begin{cases} 1, w_t \in d_i \\ 0, w_t \notin d_i \end{cases}$$

- Topic is dependent only on occurrence and non-occurrence of words

Naïve Bayes – Setting II

- Generative Model – Multinomial

$$P(w_t \mid c_j, \theta) = \frac{1 + \sum_{i=1}^{|D|} N_{it} P(c_j \mid d_i)}{|V| + \sum_{s=1}^{|V|} \sum_{i=1}^{|D|} P(c_j \mid d_i)}, N_{it} = \sum w_t, w_t \in d_i$$

- Topic is dependent on count (scaled) of occurrences of words
- Including top-n choices in model possible
- Word positions exchangeable – Bag of Words

Maximum Entropy

- Prefer the most uniform model satisfying constraints

$$P(c \mid d) = \frac{e^{\sum_i \lambda_i f_i(d,c)}}{\sum_c e^{\sum_i \lambda_i f_i(d,c)}}$$

- Real-valued functions $f_i(d,c)$

$$f_{w,c'}(d,c) = \begin{cases} 0, c \neq c' \\ \frac{N(d,w)}{|d|}, otherwise \end{cases}$$

- ML Estimation of MaxEnt overfits
- MAP estimation using Gaussian priors over feature functions

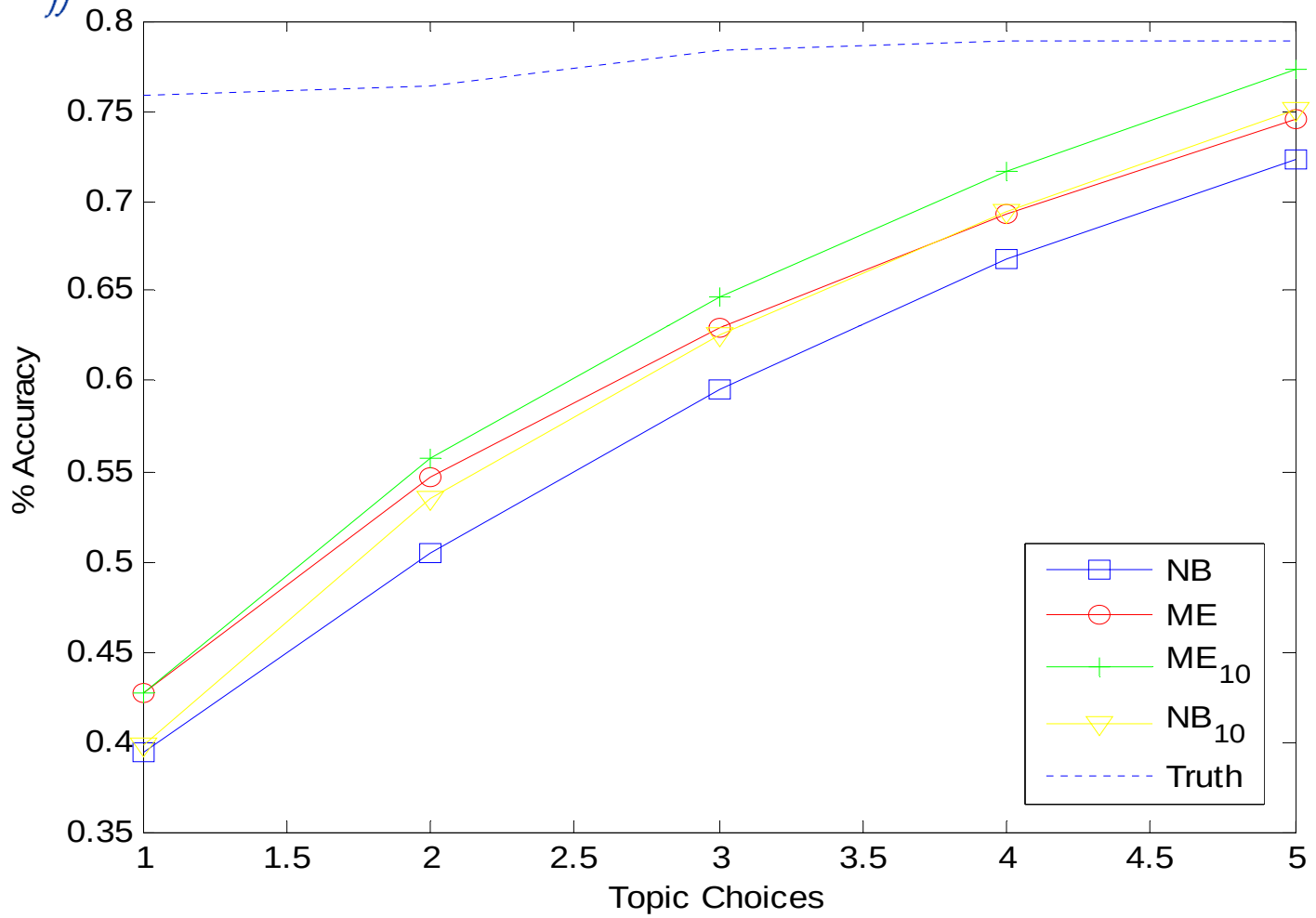
Lexicon Reduction

- $\text{Lexicon} = \underset{i}{\text{argmax}}\ \text{M.I.}(C, V_i)$
- Average Lexicon per topic ~1k
- Average Reduction in Lexicon size ~ 1/5

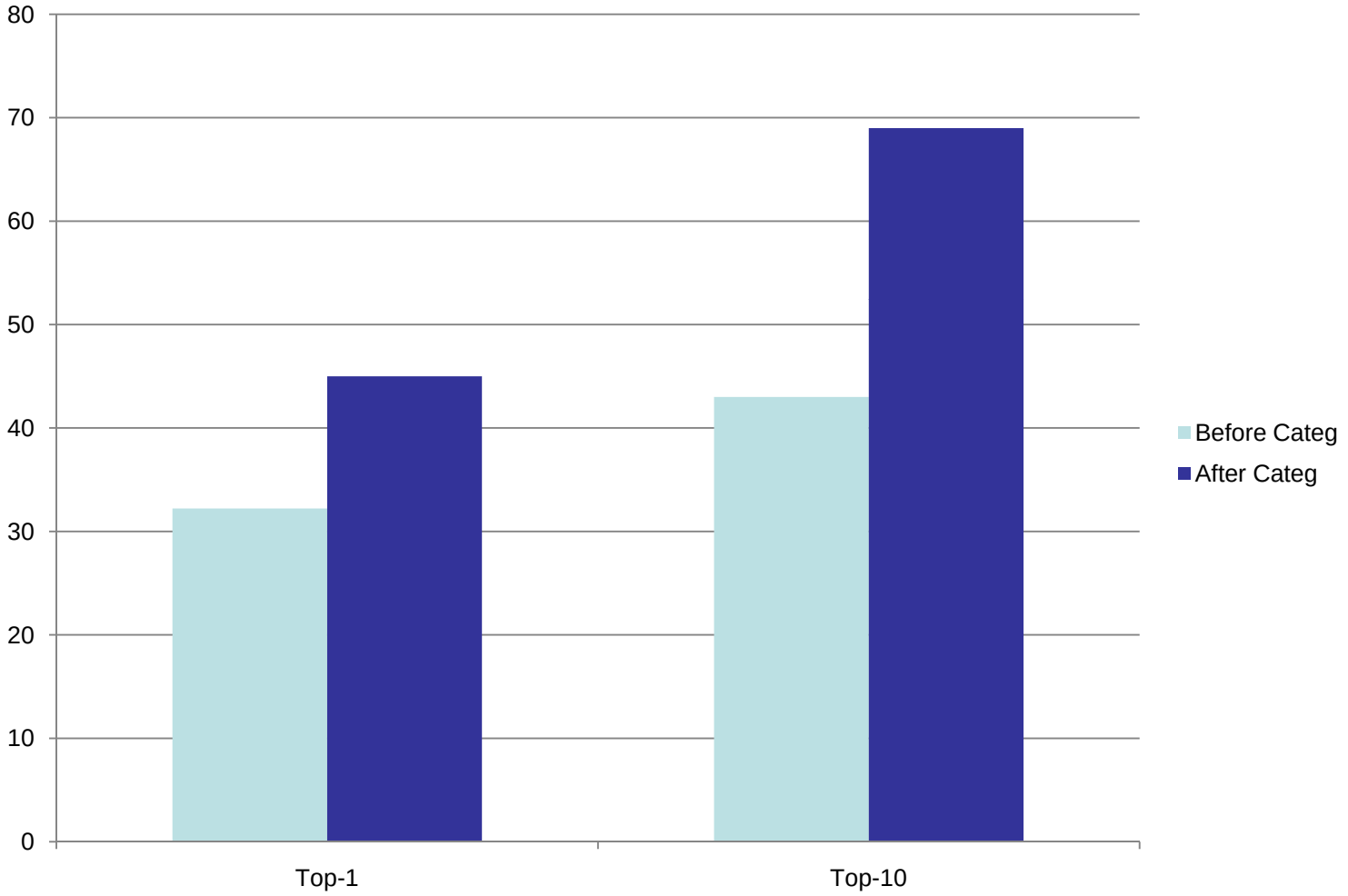
Outline

- Problem Statement
- Proposed Approach
- Results
- Conclusions

Categorization Results

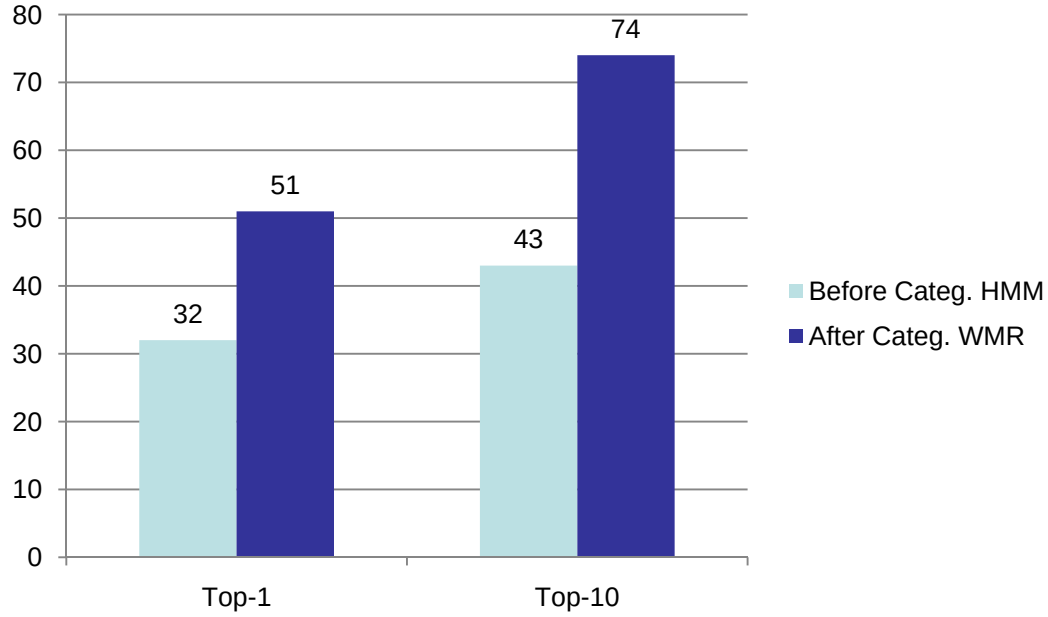


Recognition Accuracies



Combination

- HMM handles larger lexicons
- WMR better at smaller lexicons
- Reduce Lexicon HMM → WMR



Outline

- Problem Statement
- Proposed Approach
- Results
- Conclusions

Conclusions

- Lexicon Reduction – a principled way to improving recognition accuracies without retraining OCR
- Reduced Lexicon generated using Topic Models.
- Top-n output from OCR can be utilized

cse@buffalo

Thank You