Learning the Semantic Words and Pictures

Barnard *et al*.

Presented by Michael S. Ryoo

Annotation Problem

- We want to predict 'text' information, given an image.
- Primitive method:
  - Perform 'recognition' for each region.
  - Tiger in a sky?
- Joint probability!

Input

Image processing*

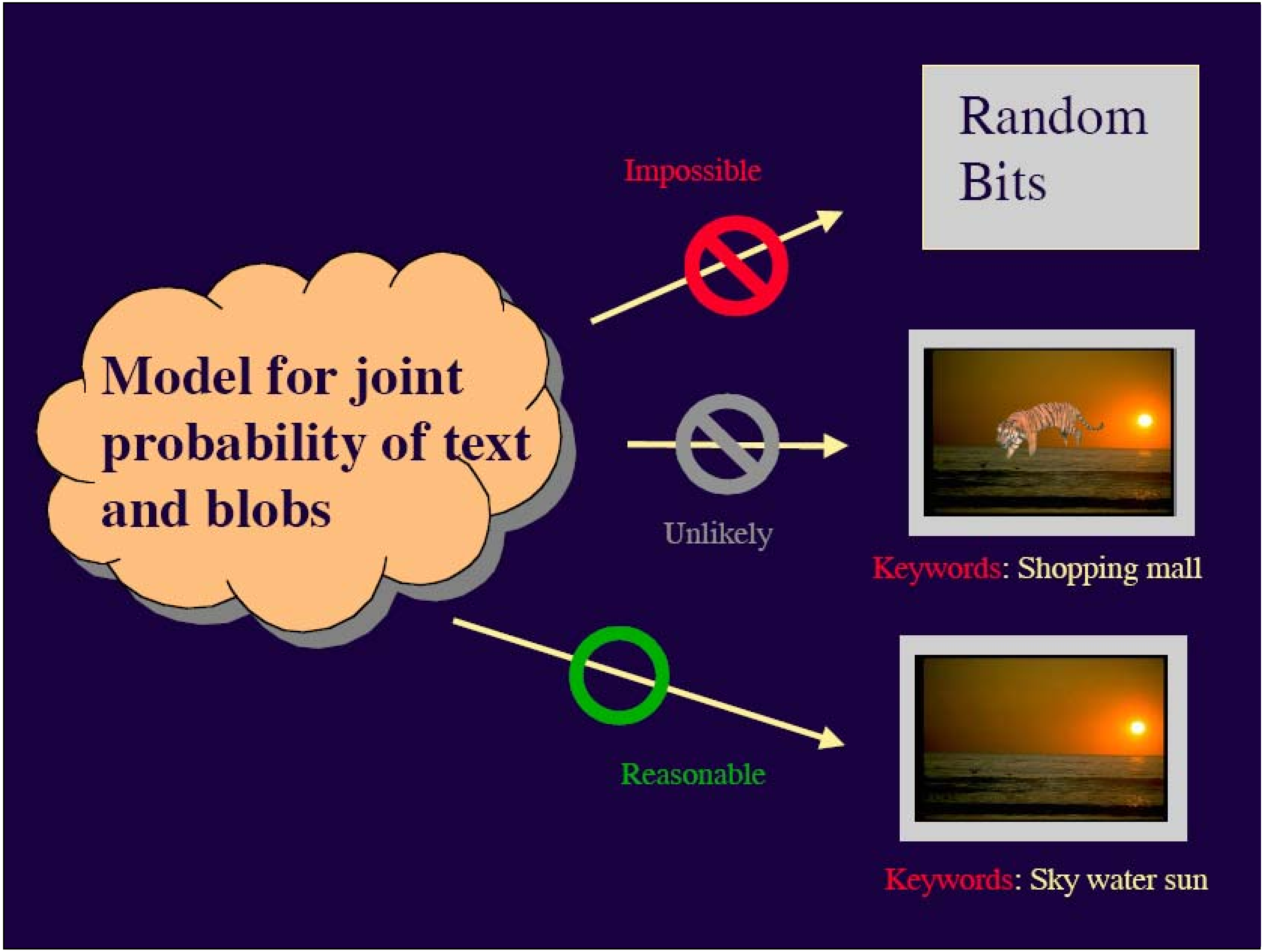
Each blob is a large vector of features

"This is a picture of the sun setting over the sea with waves in the foreground"

Language processing

sun sky waves sea

* Thanks to Blobworld team [Carson, Belongie, Greenspan, Malik], N-cuts team [Shi, Tal, Malik]

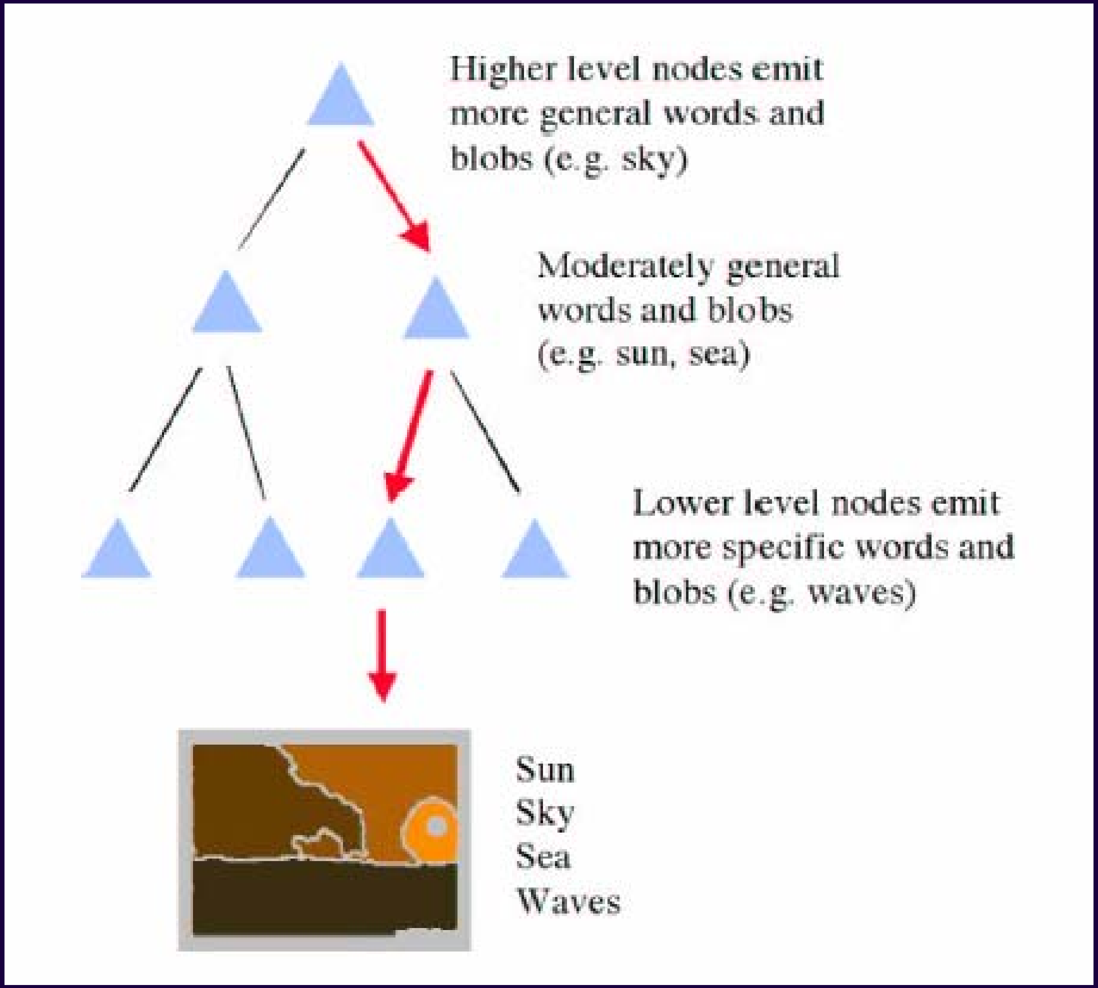

Random
Bits
Impossible
Model for joint probability of text and blobs
Unlikely
Keywords: Shopping mall
Reasonable
Keywords: Sky water sun

Hierarchical model

- Model for joint probability of text and blobs.

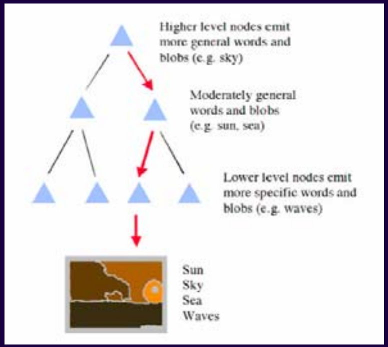
- Extension of Hofmann's model for text.
  - Hofmann, 1998; Hofmann and Puzicha, 1998
  - Each node generates a (region, word) pair.
  - Following a path from the root to a leaf generates full image and full text.

Hierarchical model



Probability Distribution

- Correlation considered model.



$$p(D|d)=\sum_{c} p(c)\prod_{(w,b)\in D}\left[\sum_{l} p((w,b)|l,c)\,p(l|d)\right]$$

- Conditionally independent model.

$$p(D|d)=\sum_{c} p(c)\prod_{w\in W}\left[\sum_{l} p(w|l,c)\,p(l|d)\right]^{\frac{N_W}{N_{w,d}}}\prod_{b\in B}\left[\sum_{l} p(b|l,c)\,p(l|d)\right]^{\frac{N_b}{N_{b,d}}}$$

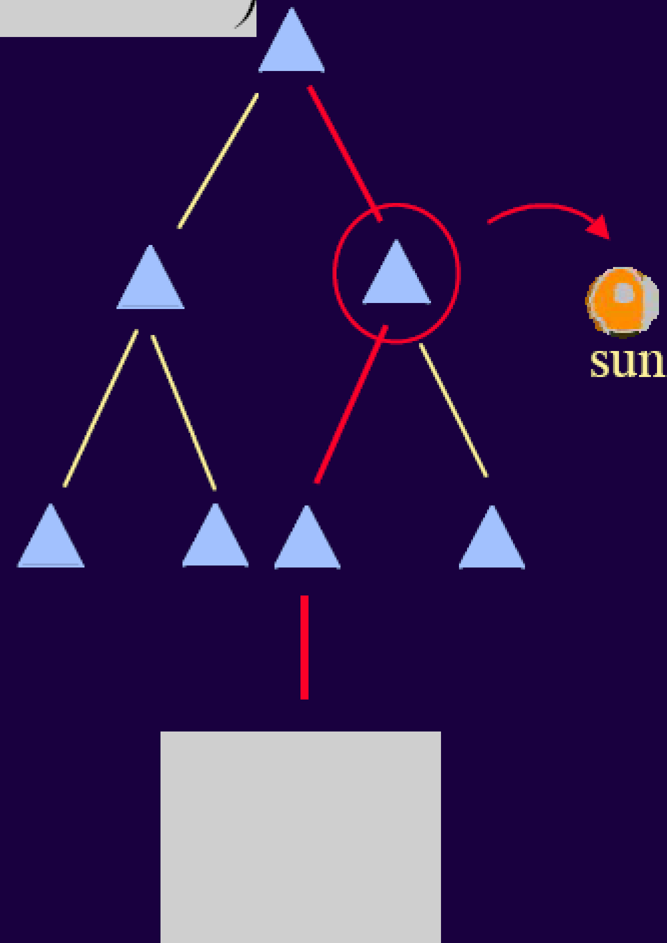
  – Observations D = (W + B), document d, cluster c, and level l.

$$P(D \mid d) = \sum_{c} P(c) \prod_{p \in D} \left(\sum_{l} P(p \mid l, c) P(l \mid d) \right)$$

cluster
c = 3

level
l = 2

pair
p =1

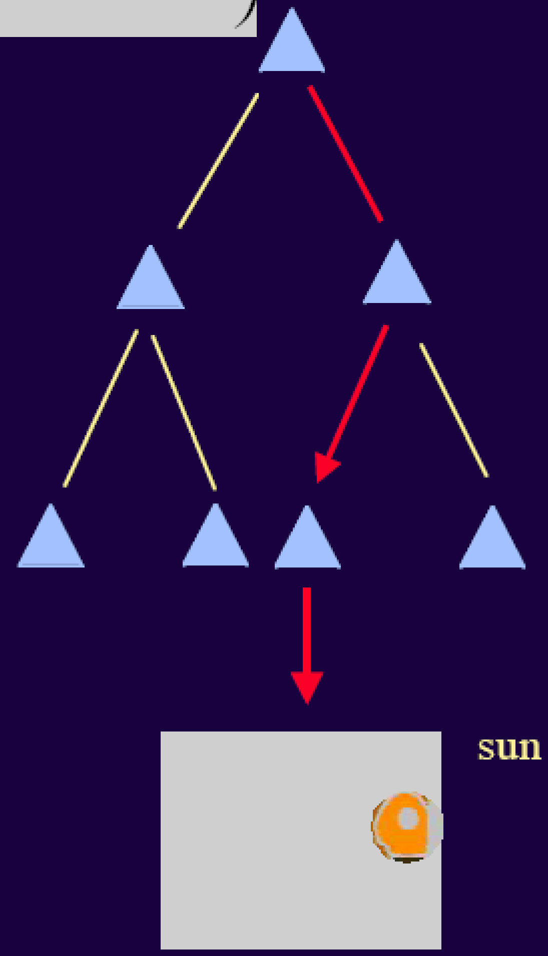


$$P(D \mid d) = \sum_{c} P(c) \prod_{p \in D} \left(\sum_{l} P(p \mid l, c) P(l \mid d) \right)$$

cluster
c = 3

level
l = 2

pair
p =1

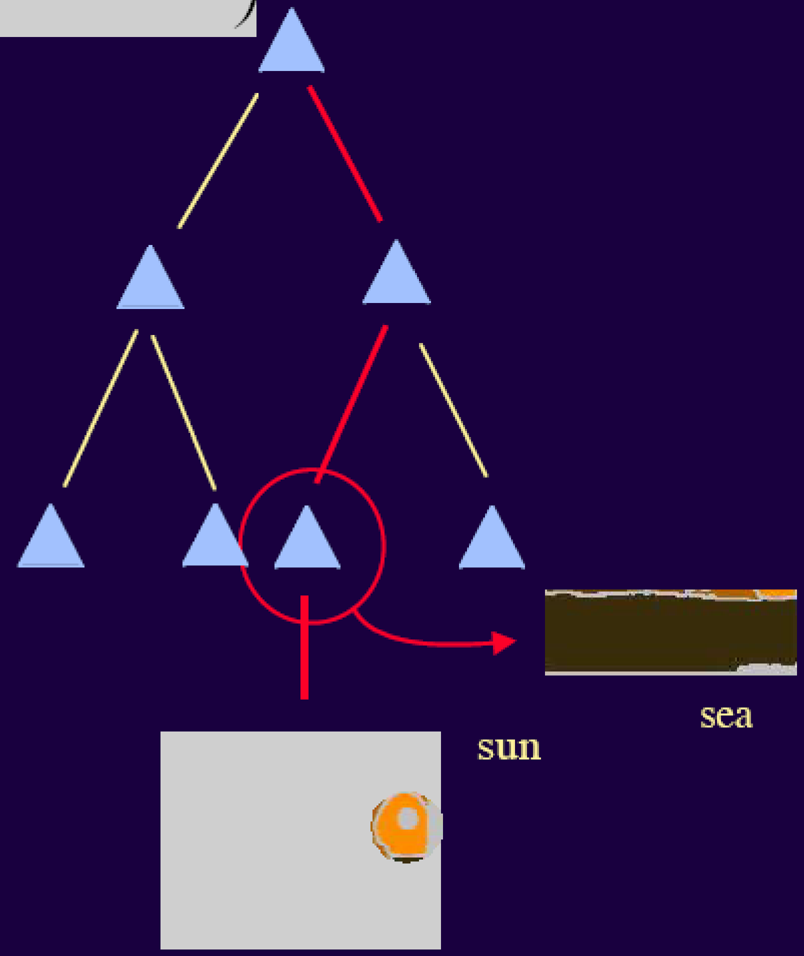


$$P(D \mid d) = \sum_{c} P(c) \prod_{p \in D} \left(\sum_{l} P(p \mid l, c) P(l \mid d) \right)$$

cluster
c = 3

level
l = 3

pair
p =2

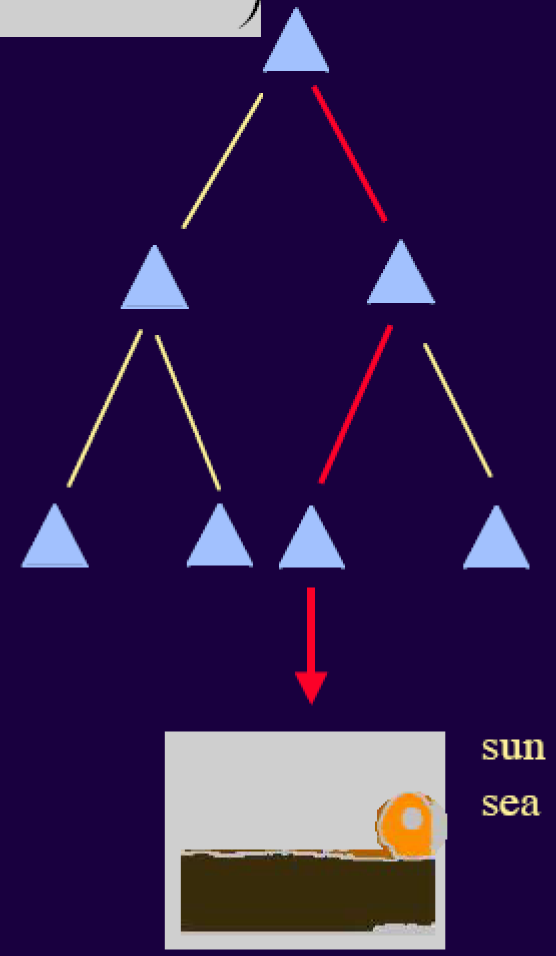


$$P(D \mid d) = \sum_{c} P(c) \prod_{p \in D} \left(\sum_{l} P(p \mid l, c) P(l \mid d) \right)$$

cluster
c = 3

level
l = 3

pair
p =2

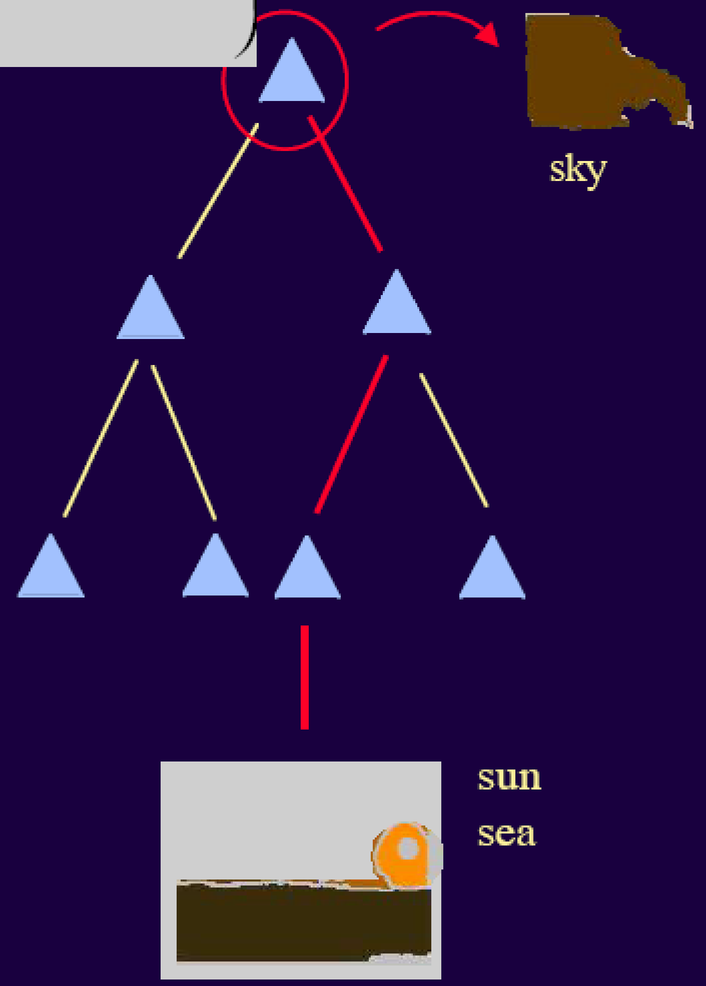


$$P(D \mid d) = \sum_{c} P(c) \prod_{p \in D} \left(\sum_{l} P(p \mid l, c) P(l \mid d) \right)$$

cluster
c = 3

level
l = 1

pair
p =3

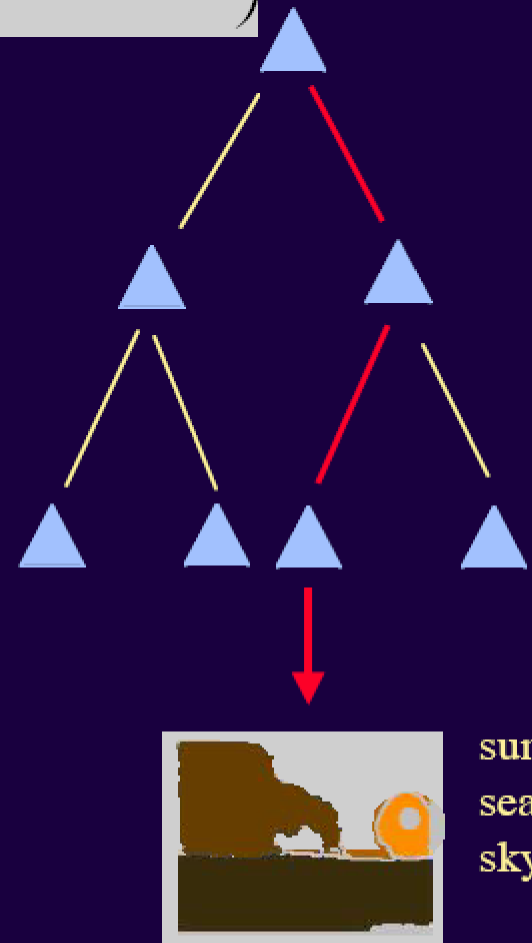


$$P(D \mid d) = \sum_{c} P(c) \prod_{p \in D} \left(\sum_{l} P(p \mid l, c) P(l \mid d) \right)$$

cluster
c = 3

level
l = 1

pair
p =3



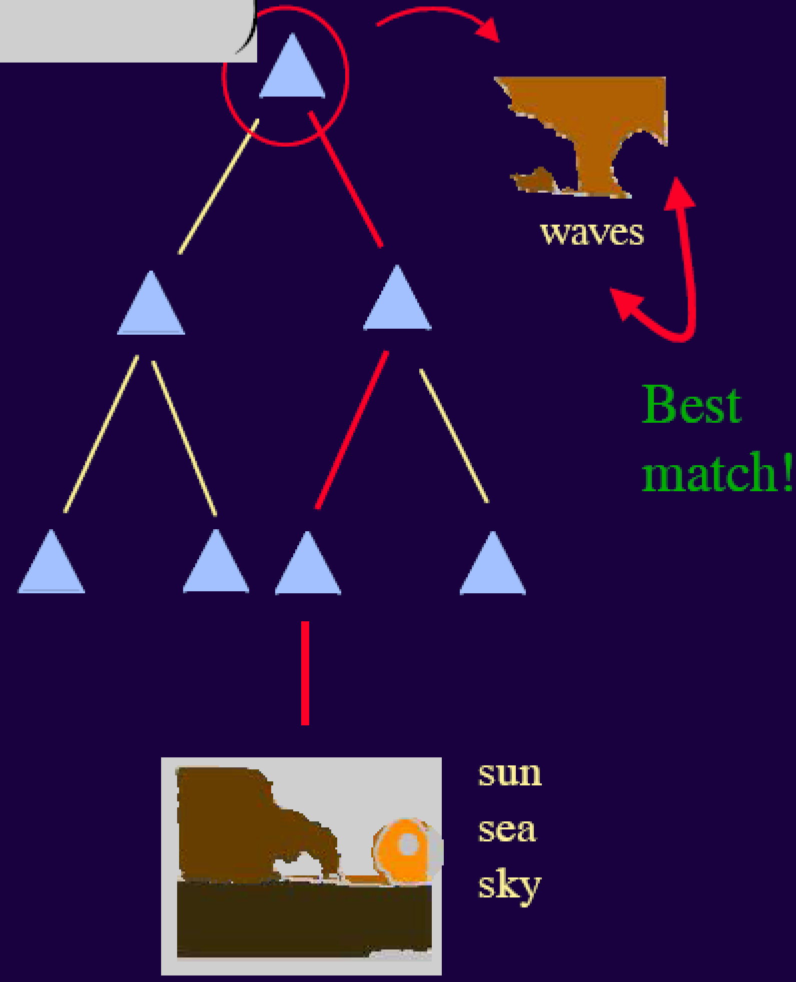
sun
sea
sky

$$P(D \mid d) = \sum_{c} P(c) \prod_{p \in D} \left(\sum_{l} P(p \mid l, c) P(l \mid d) \right)$$

cluster
c = 3

level
l = 1

pair
p =4

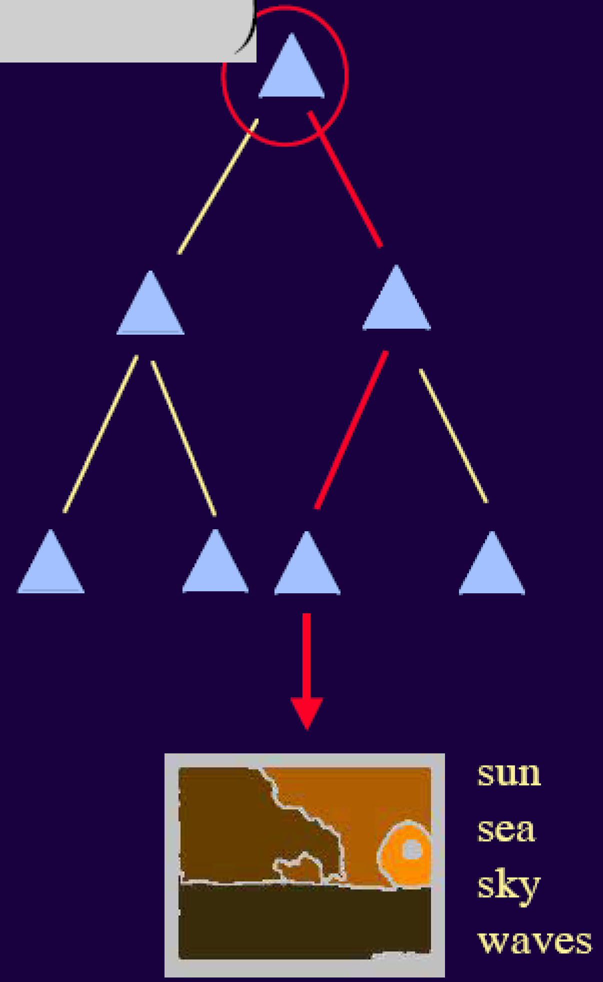


$$P(D \mid d) = \sum_{c} P(c) \prod_{p \in D} \left(\sum_{l} P(p \mid l, c) P(l \mid d) \right)$$

cluster
c = 3

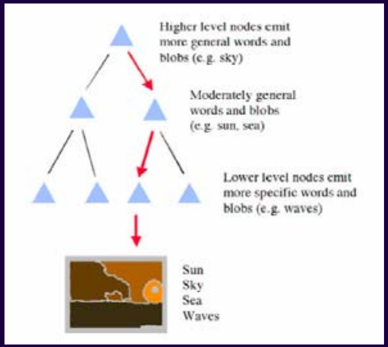
level
l = 1

pair
p =4



Probability Distribution

- Correlation considered model.



$$p(D|d) = \sum_{c} p(c) \prod_{(w,b)\in D} \left[\sum_{l} p((w,b)|l,c)\, p(l|d) \right]$$

- Conditionally independent model.

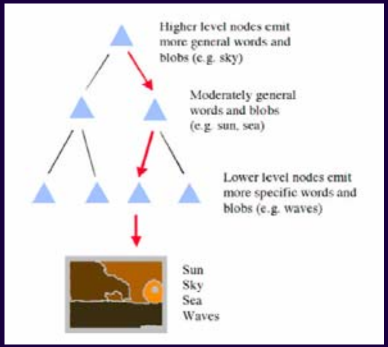
$$p(D|d) = \sum_{c} p(c) \prod_{w\in W} \left[\sum_{l} p(w|l,c)\, p(l|d) \right]^{\frac{N_W}{N_{w,d}}} \prod_{b\in B} \left[\sum_{l} p(b|l,c)\, p(l|d) \right]^{\frac{N_b}{N_{b,d}}}$$

  – Observations D = (W + B), document d, cluster c, and level l.

Predicting words using model

- Annotation, P(w | b) ≈ P(w, b)
- Correlation considered model.

$$p(w \Leftrightarrow b) \approx \sum_{c} p(c) \sum_{l} p((w,b)|l,c)p(l|d).$$



- Conditionally independent model.

$$p(w|b) \propto \sum_{c} p(c) \sum_{l} p(l)p(w|l,c)p(b|l,c).$$

Evaluation

- Compared annotation done by their model and 'empirical' word distribution model.
  - Annotate all region as common word (water?)
- Calculated Kullback-Leibler divergence.

$$E_{KL}^{(model)} = \sum_{w \in vocabulary} p(w) \log \frac{p(w)}{q(w|B)}.$$

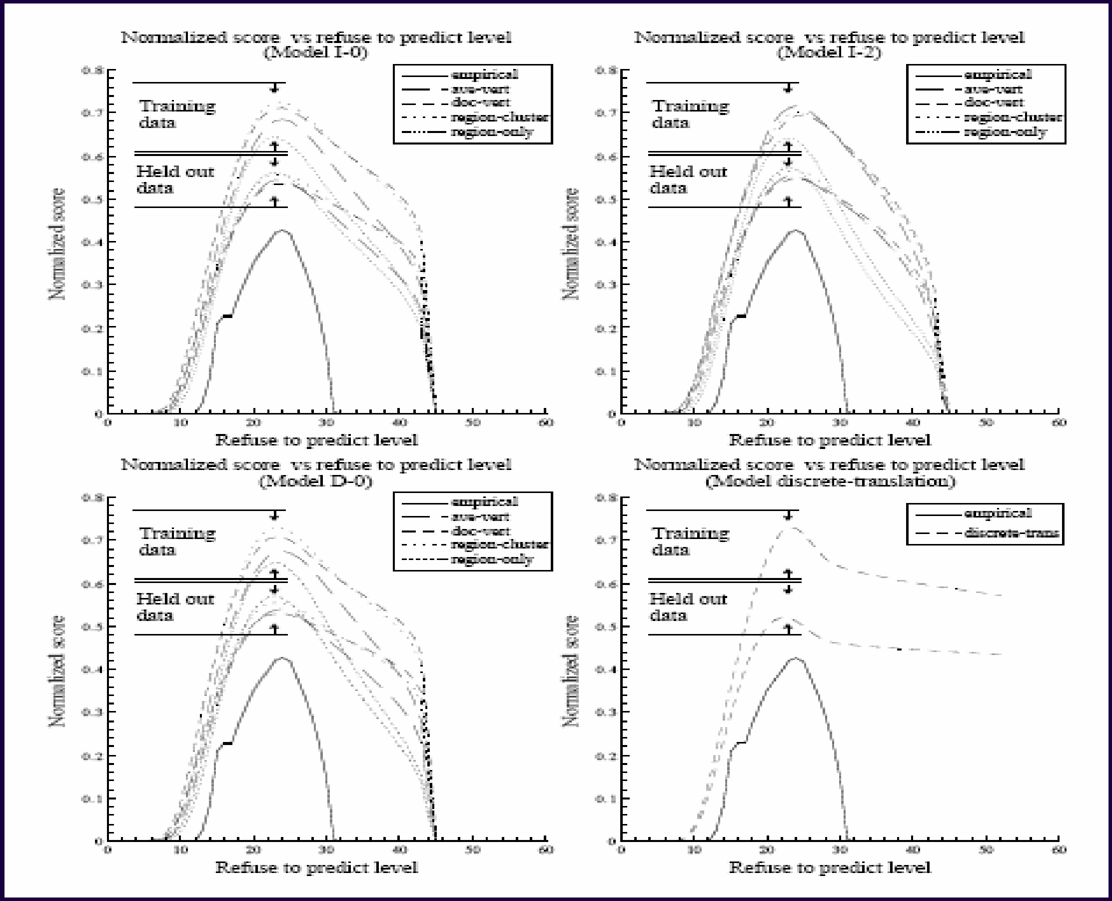
$$E_{KL}^{(model)} = \frac{1}{K} \sum_{w \in observed} \log \frac{p(w)}{q(w|B)}$$

- Also constructed their own function.
  - $E_{NS}^{(model)} = r/n - w/(N - n)$

Corel Database

392 CD's, each consisting of 100 annotated images.

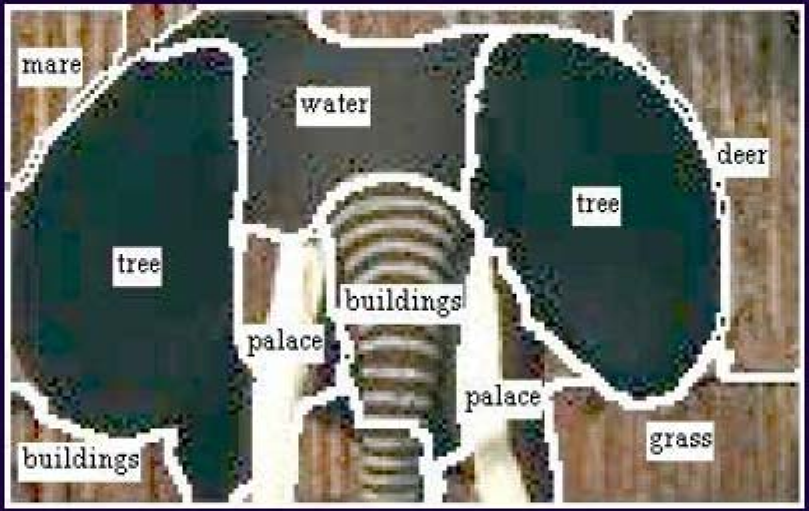
Experiments



Experiments

Method	Training data	Held out data	Novel data
linear-I-0-doc-vert	0.301 (0.005)	0.174 (0.007)	0.081 (0.007)
binary-I-0-ave-vert	0.294 (0.006)	0.154 (0.006)	0.064 (0.008)
binary-I-0-doc-vert	0.325 (0.006)	0.160 (0.007)	0.065 (0.008)
binary-I-0-region-cluster	0.332 (0.006)	0.168 (0.007)	0.068 (0.008)
binary-I-0-region-only	0.234 (0.006)	0.160 (0.006)	0.062 (0.008)
binary-I-2-ave-vert	0.331 (0.006)	0.164 (0.008)	0.068 (0.007)
binary-I-2-doc-vert	0.322 (0.006)	0.170 (0.008)	0.074 (0.008)
binary-I-2-region-cluster	0.324 (0.006)	0.179 (0.008)	0.076 (0.008)
binary-I-2-region-only	0.228 (0.006)	0.163 (0.006)	0.068 (0.007)
linear-D-0-doc-vert	0.321 (0.005)	0.167 (0.006)	0.076 (0.008)
binary-D-0-ave-vert	0.284 (0.007)	0.151 (0.007)	0.061 (0.008)
binary-D-0-doc-vert	0.321 (0.007)	0.157 (0.007)	0.064 (0.008)
binary-D-0-region-cluster	0.330 (0.006)	0.166 (0.008)	0.067 (0.008)
binary-D-0-region-only	0.239 (0.006)	0.162 (0.007)	0.064 (0.007)
binary-D-2-ave-vert	0.312 (0.005)	0.162 (0.003)	0.066 (0.005)
binary-D-2-doc-vert	0.358 (0.005)	0.172 (0.003)	0.069 (0.005)
binary-D-2-region-cluster	0.360 (0.005)	0.179 (0.003)	0.072 (0.005)
binary-D-2-region-only	0.248 (0.005)	0.167 (0.003)	0.066 (0.005)
linear-C-0-region-only	0.240 (0.005)	0.124 (0.007)	0.046 (0.006)
binary-C-0-ave-vert	0.252 (0.006)	0.143 (0.007)	0.060 (0.008)
binary-C-0-doc-vert	0.281 (0.006)	0.148 (0.006)	0.054 (0.007)
binary-C-0-region-cluster	0.290 (0.006)	0.157 (0.007)	0.064 (0.007)
binary-C-0-region-only	0.233 (0.006)	0.163 (0.006)	0.071 (0.006)
discrete-translation	0.318 (0.005)	0.111 (0.007)	0.016 (0.008)
MoM-LDA	0.125 (0.005)	0.107 (0.005)	0.041 (0.007)

grass
cat
horses
buildings
tiger
grass
mare
mare
aterfalls
sky
row
buildings
building
buildings
house
people
tree
tree
ground
flowers
ground
antlers
shops
town
crab
tiger
town
ground


mare
water
deer
tree
tree
buildings
palace
palace
grass
buildings

Searching Problem

- Given query image or query sentences, find a document that best matches.
  - Google Images
  - Content-based Image Retrieval

- This paper mentions that annotated words can be matched with queries…
  - $P(Q \mid d)$ vs $P(d \mid Q)$
  - Need to consider prior probabilities.

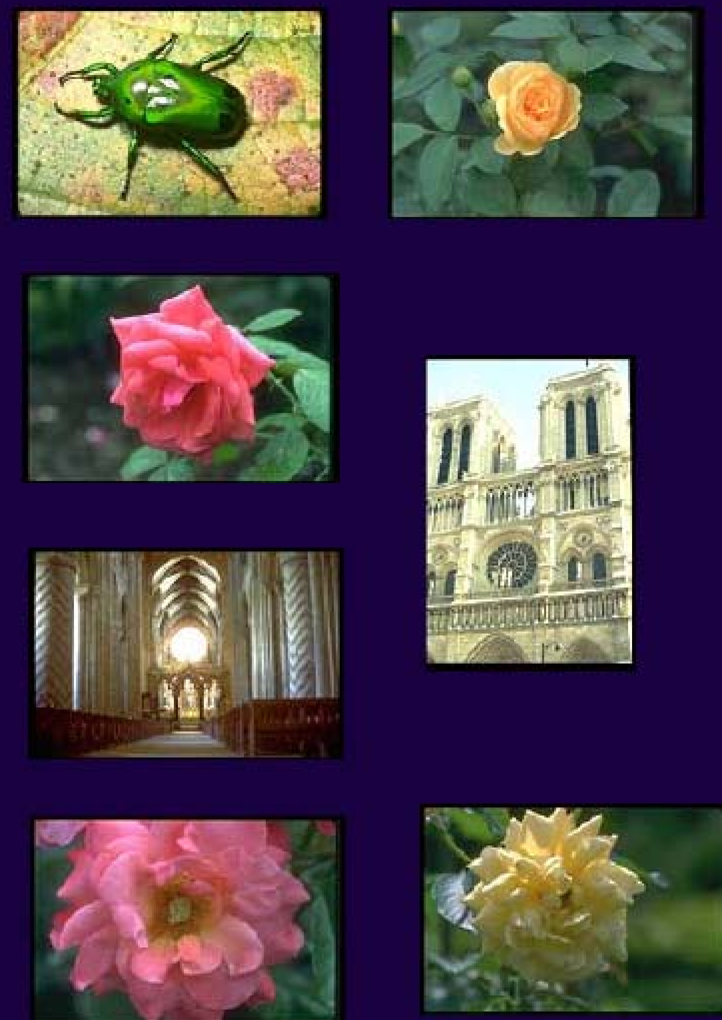
Conclusions

- Hierarchical model was proposed.
- Modeling image regions and words jointly.
- Annotation of image regions were done.

Keywords: rose flower plant leaves

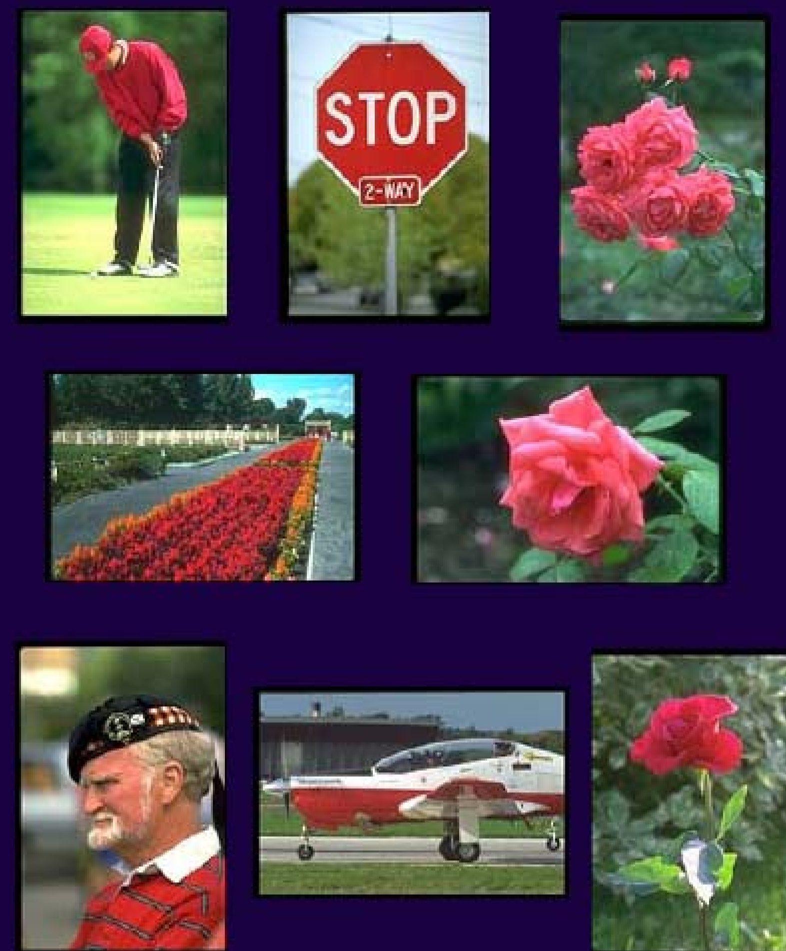
Query on

“Rose”



Example from Berkeley Blobworld system

Query on



Example from Berkeley
Blobworld system