Systematic Evaluation of Machine Translation Methods for Image and Video Annotation

Paola Virga

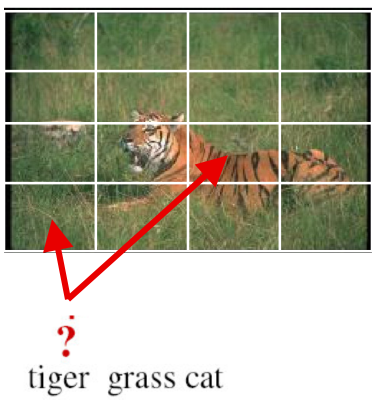
Johns Hopkins University, USA

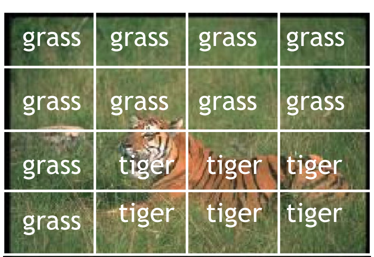
Pinar Duygulu

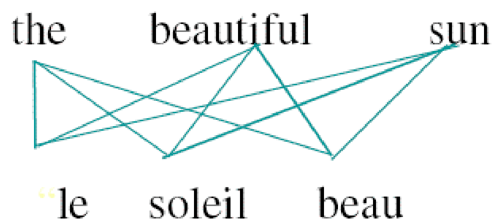
Bilkent University, Turkey

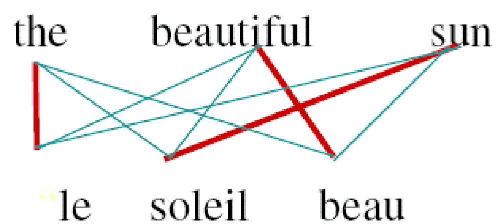
CIVR 2005

Inspiration from Machine Translation

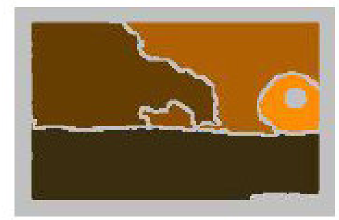




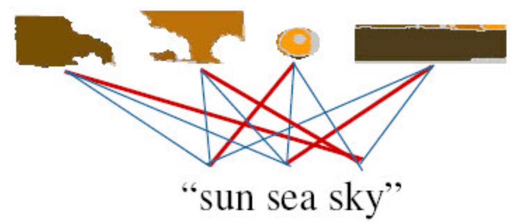




$$p(f|e) = \sum_{a} p(f,a|e)$$



"sun sea sky"



$$p(c|v) = \sum_{a} p(c,a|v)$$

Direct translation model

Discrete Representation of Image Regions (visterms) to create analogy to MT

In Machine Translation → discrete tokens

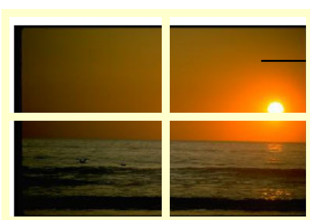
In our task

sun sky waves sea

concepts

However, the features extracted from regions are continuous

Solution : Vector quantization → visterms



$\{f_{n1}, f_{n2}, \ldots f_{nm}\} \text{ -> } v_k$

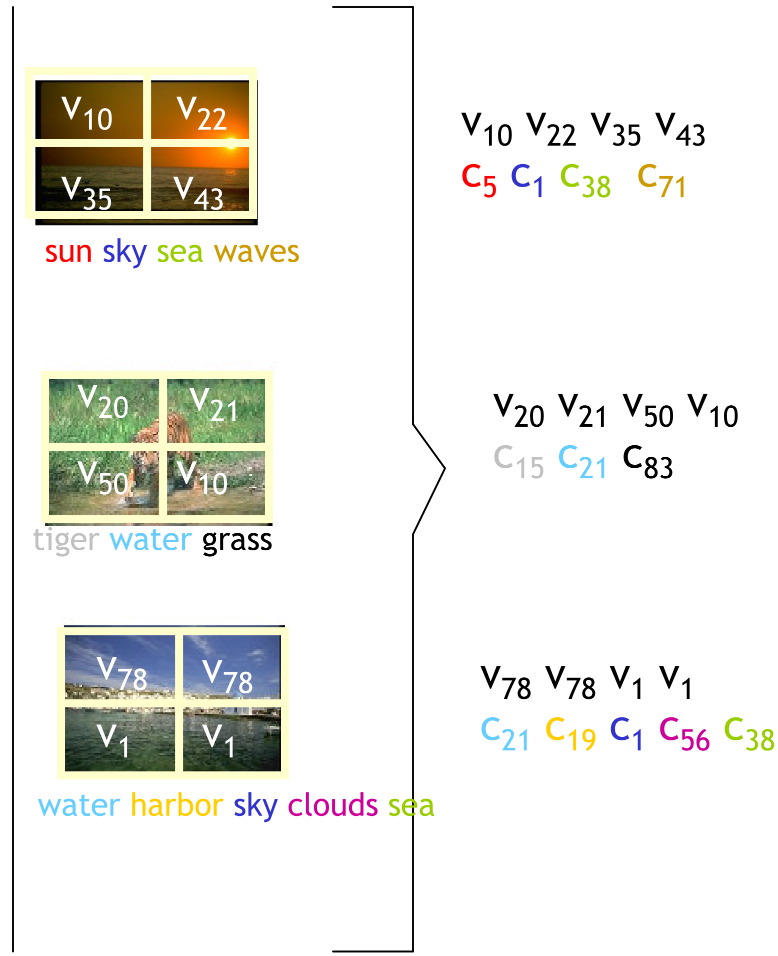
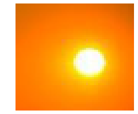
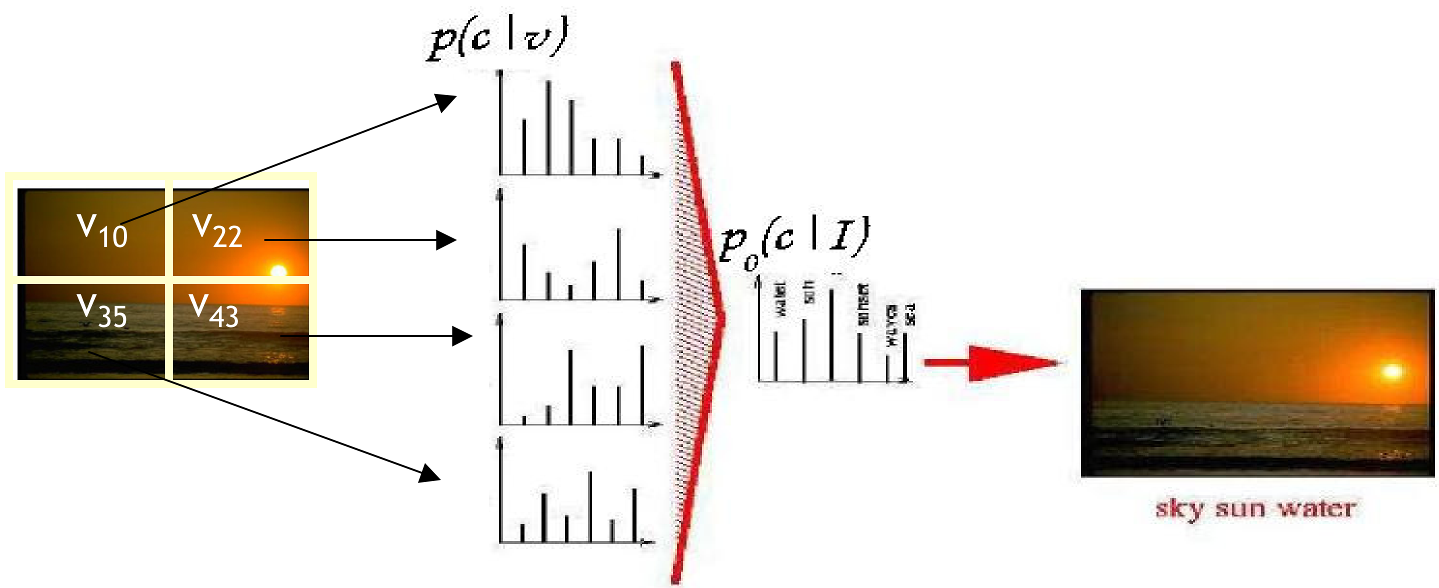


Image annotation using translation probabilities

p(c|v) : Probabilities obtained from direct translation

p (sun | )

$$P_0(c \mid d_V) = \frac{1}{|d_V|} \sum_{v \in d_V} P(c \mid v)$$



Data Sets

Data Set	# Blocks	# Concepts	Training Size	Test Size
Corel	24(6x4)	374	4500	500
TRECVID	35(7x5)	138 (75 used)*	34880	9220

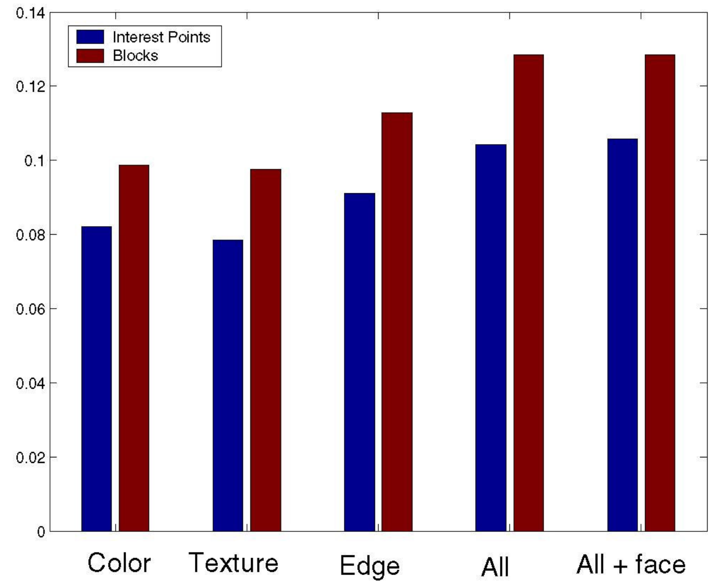
* Most frequent

Feature selection

Features : color, texture, edge
Extracted from blocks, or around interest points

Observations

- Features extracted from blocks give better performance than features extracted around interest points
- When the features are used individually
  Edge features give the best performance
- Training using all is the best
  - Using Information Gain to select visterms vocabulary didn't help
- Integrating number of faces, increases the performance slightly



mAP values for different features

Model and iteration selection

Strategies compared

- (a) IBM Model 1
- (b) HMM Model on top of (a)
- (c) IBM Model 4 on top of (b)

-> Observation : IBM Model 1 is the best

Corel	TREC
0.125	0.124

Number of iterations in Giza++ training affects the performance

-> Less iterations give better annotation performance
but cannot produce rare words

Integrating word co-occurrences

- Model 1 with word co-occurrence

$$P_1(c_i \mid d_V) = \sum_{j=1}^{|C|} P(c_i \mid c_j) P_0(c_j \mid d_V)$$

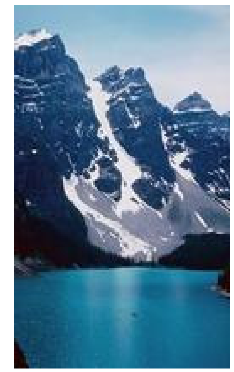
- Integrating word co-occurrences into the model helps for Corel but not for TREC

	Corel	TREC
Model 1	0.125	0.124
Model 1 + Word-CO	0.145	0.124

Annotation Results (Corel set)

field foals horses mare
tree horses foals mare field

flowers leaf petals stems
flowers leaf petals grass tulip



mountain sky snow water
sky mountain water clouds snow

people pool swimmers water
swimmers pool people water sky

jet plane sky
sky plane jet tree clouds

people sand sky water
sky water beach people hills

Top: manual annotations, bottom : predicted words (top 5 words with the highest probability)
Red : correct matches

MT Models Analysis

$$P(\mathbf{f} \mid \mathbf{e}) = \sum_{\mathbf{a}} P(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = \sum_{\mathbf{a}} P(m \mid \mathbf{e}) P(\mathbf{a} \mid m, \mathbf{e}) P(\mathbf{f} \mid \mathbf{a}, m, \mathbf{e})$$

- ✓ **"f"** French Sentence of length **m** **(Concept language)**
- ✓ **"e"** English Sentence of length **l** **(Visterms language)**
- ✓ **"a"** Alignment between the French sentence **"f"** and the English sentence **"e"**
- ✓ **P(m|e)** String length probability
- ✓ **P(a|m,e)** Alignment probability
- ✓ **P(f|a,m,e)** Word translation probabilities

Model 1-2-HMM

Model 1 assumptions:

$$P(m \mid \mathbf{e}) = \varepsilon(m \mid l)$$

$$P(\mathbf{a} \mid m, \mathbf{e}) = (l+1)^{-m}$$

Each Alignment is equally probable

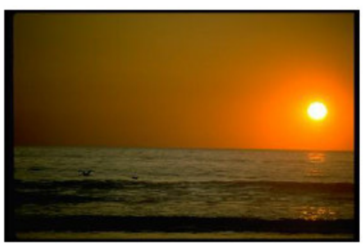
$$P(\mathbf{f} \mid \mathbf{a}, m, \mathbf{e}) = \prod_{j=1}^{m} t\left(f_j \mid e_{a_j}\right)$$

Model 2 assumptions:

String length probabilities and Translation word probabilities as Model 1

$$P(\mathbf{a} \mid m, \mathbf{e}) = \prod_{j=1}^{m} P\left(a_j \mid j, l, m\right)$$

The alignment depends on the position of the concept



sun sky waves sea

- The concept sentence associated to the image can be one of the following : {sun, sky, waves, sea} {sun, waves, sea, sky} {sky, sun, sea, waves} {sky, sea, waves, sun} …. The position of a concept in the annotation depends only on the annotator and not on the image itself.

IBM Model 1-2 &HMM (cont..)

Model 1 assumptions:

$$P(m \mid \mathbf{e}) = \varepsilon(m \mid l)$$

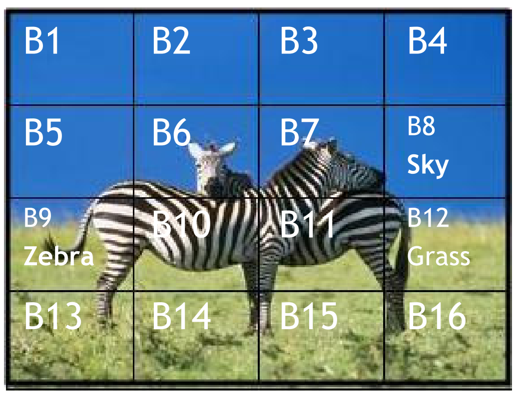
$$P(\mathbf{a} \mid m, \mathbf{e}) = (l+1)^{-m}$$ Each Alignment is equally probable

$$P(\mathbf{f} \mid \mathbf{a}, m, \mathbf{e}) = \prod_{j=1}^{m} t\left(f_j \mid e_{a_j}\right)$$

HMM Model assumptions:

String length probabilities and Translation word probabilities as Model 1

$$P(\mathbf{a} \mid m, \mathbf{e}) = \prod_{j=1}^{m} P(a_j \mid a_{j-1}, l, m)$$ The alignment depends on the previous alignment



Sky, Zebra, Grass

- Concepts as “sun” and “sky” are usually in adjacent blocks
- Given the lack of structure of the “concept” sentence, a possible scenario is:

1. Sky Zebra Grass $P\left(a_j \langle \text{Zebra} \rightarrow \text{B9} \rangle \mid a_{j-1} \langle \text{Sky} \rightarrow \text{B8} \rangle, l, m\right)$
2. Grass Zebra Sky $P\left(a_j \langle \text{Zebra} \rightarrow \text{B9} \rangle \mid a_{j-1} \langle \text{Grass} \rightarrow \text{B12} \rangle, l, m\right)$

* The model favorites alignments close to each other.

IBM Model 3-4-5

$$P(\mathbf{f}\,|\,\mathbf{e}) = \sum_{\mathbf{a}} P(\mathbf{f},\mathbf{a}\,|\,\mathbf{e}) = \sum_{\tau,\pi \in \langle \mathbf{f},\mathbf{a} \rangle} P(\mathbf{\Phi}\,|\,\mathbf{e}) P(\tau\,|\,\mathbf{\Phi},\mathbf{e}) P(\pi|\tau,\mathbf{\Phi},\mathbf{e})$$

- $P(\mathbf{\Phi}\,|\,\mathbf{e})$ Fertility probability

  Fertility is the number of concepts associated with a visterm. In our task such number does not depend on the concept but on the image itself. Depending on the resolution of the image a particular visterm can be associated with one or more concepts.

- $P(\pi|\tau,\mathbf{\Phi},\mathbf{e})$ Distortion probability

  The concept of distortion is used to deal with different language word orders: English is an SVO language while Arabic is a VSO language. It is not possible to apply it to our task since the "concept" language lacks of structure.

Inspiration from CLIR

- Treat Image Annotation as a Cross-lingual IR problem
  - Visual Document comprising visterms (target language) and a query comprising a concept (source language)

$$p(c \mid d_V) = \lambda \left(\sum_{v \in V} p(c \mid v) \, p(v \mid d_V) \right) + \underbrace{(1-\lambda) \, p(c \mid G_C)}_{\text{same } \forall d_V}$$

Inspiration from CLIR

- Treat Image Annotation as a Cross-lingual IR problem
  - Visual Document comprising visterms (target language) and a query comprising a concept (source language)

$$p(c \mid d_V) = \sum_{v \in d_V} p\left(v \mid d_V\right) p(c \mid v)$$

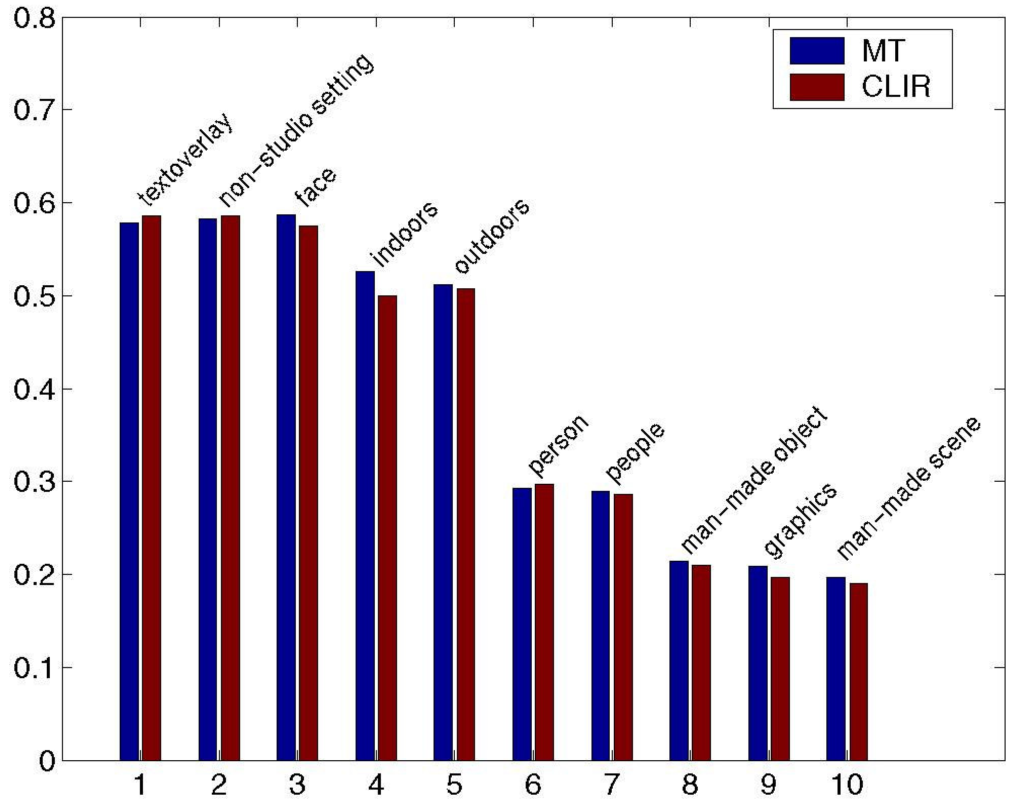
  - Image does not provide a good estimate of $p(v|d_v)$
  - Tried p(v) and DF(v), DF works best

$$score(c \mid d_V) = \sum_{v \in d_V} DF_{Train}(v)\, p(c \mid v)$$

Annotation Performance on TREC

Model 1	0.124
CLIR using Model 1	0.126

Significant at p=0.04



Average Precision values for the top 10 words
For some concepts we achieved up to 0.6

Conclusions

The simplest Translation Model (IBM Model 1) outperforms the more sophisticated ones.

Why:

1. IBM Models and HMM Model are suited for syntactically rich languages.

   Translations from Arabic to English are better than from Chinese to English.

2. The two-dimension image structure gets flattened, it is only partially preserved the horizontal order.
3. The length of the two parallel sentences (“concept” sentence and “visterms” sentence) are dramatically different, m<<l.