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

\[ p(f | e) = \sum_a p(f, a | e) \]

\[ p(c | v) = \sum_a p(c, a | v) \]

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

In Machine Translation $\rightarrow$ discrete tokens

In our task

sun sky waves sea concepts $\checkmark$

However, the features extracted from regions are continuous

Solution : Vector quantization $\rightarrow$ visterms $\checkmark$

$\Rightarrow \{f_{n1}, f_{n2}, \ldots, f_{nm}\} \rightarrow v_k$

$V_{10} V_{22} V_{35} V_{43}$
$c_5 c_1 c_{38} c_{71}$

$V_{20} V_{21} V_{50} V_{10}$
$c_{15} c_{21} c_{83}$

$V_{78} V_{78} V_1 V_1$
$c_{21} c_{19} c_1 c_{56} c_{38}$

water harbor sky clouds sea
Image annotation using translation probabilities

$p(c \mid v)$: Probabilities obtained from direct translation

$$P_0(c \mid d_v) = \frac{1}{|d_v|} \sum_{v \in d_v} P(c \mid v)$$
## Data Sets

<table>
<thead>
<tr>
<th>Data Set</th>
<th># Blocks</th>
<th># Concepts</th>
<th>Training Size</th>
<th>Test Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corel</td>
<td>24(6x4)</td>
<td>374</td>
<td>4500</td>
<td>500</td>
</tr>
<tr>
<td>TRECVID</td>
<td>35(7x5)</td>
<td>138 (75 used)*</td>
<td>34880</td>
<td>9220</td>
</tr>
</tbody>
</table>

* 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 visterns 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

<table>
<thead>
<tr>
<th>Corel</th>
<th>TREC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.125</td>
<td>0.124</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Corel</th>
<th>TREC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>0.125</td>
<td>0.124</td>
</tr>
<tr>
<td>Model 1 + Word-CO</td>
<td>0.145</td>
<td>0.124</td>
</tr>
</tbody>
</table>
Annotation Results (Corel set)

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

\[ P(f \mid e) = \sum_a P(f, a \mid e) = \sum_a P(m \mid e) P(a \mid m, e) P(f \mid a, m, e) \]

✓ “f” French Sentence of length m  (Concept language)
✓ “e” English Sentence of length 1  (Visterms language)
✓ “a” Alignment between the French sentence “f” and the English sentence “e”
✓ \( P(m \mid e) \) String length probability
✓ \( P(a \mid m, e) \) Alignment probability
✓ \( P(f \mid a, m, e) \) Word translation probabilities
Model 1-2-HMM

Model 1 assumptions:
\[ P(m | e) = \mathcal{E}(m | l) \]
\[ P(a | m, e) = (l + 1)^{-m} \quad \text{Each Alignment is equally probable} \]
\[ P(f | a, m, e) = \prod_{j=1}^{m} t(f_j | e_{a_j}) \]

Model 2 assumptions:
String length probabilities and Translation word probabilities as Model 1

\[ P(a | m, e) = \prod_{j=1}^{m} P(a_j | j, l, m) \quad \text{The alignment depends on the position of the concept} \]

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.

sun sky waves sea
IBM Model 1-2 & HMM (cont..)

Model 1 assumptions:

\[ P(m \mid e) = \epsilon(m \mid l) \]
\[ P(a \mid m, e) = (l + 1)^{-m} \quad \text{Each Alignment is equally probable} \]
\[ P(f \mid a, m, e) = \prod_{j=1}^{m} t(f_j \mid e_{a_j}) \]

HMM Model assumptions:

String length probabilities and Translation word probabilities as Model 1

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

- 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 \[ P(a_j \mid \text{Zebra} \rightarrow \text{B9}, l, m) \]
2. Grass Zebra Sky \[ P(a_j \mid \text{Zebra} \rightarrow \text{B9}, l, m) \]

* The model favors alignments close to each other.
IBM Model 3-4-5

\[ P(f | e) = \sum_a P(f, a | e) = \sum_{\tau, \pi \in \{f, a\}} P(\Phi | e)P(\tau | \Phi, e)P(\pi | \tau, \Phi, e) \]

- \( P(\Phi | e) \)  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.

- \( P(\pi | \tau, \Phi, e) \)  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)

\[
p(c \mid d_v) = \lambda \left( \sum_{v \in V} p(c \mid v)p(v \mid d_v) \right) + (1 - \lambda) p(c \mid G_C) \\
\]

\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)

\[
p(c \mid d_v) = \sum_{v \in d_v} p(v \mid d_v) p(c \mid v)
\]

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

\[
\text{score}(c \mid d_v) = \sum_{v \in d_v} DF_{\text{Train}}(v) p(c \mid v)
\]
Annotation Performance on TREC

<table>
<thead>
<tr>
<th>Model 1</th>
<th>0.124</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLIR using Model 1</td>
<td>0.126</td>
</tr>
</tbody>
</table>

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.