Entropy-Based Evaluation of Decoders

Stefan Steidl
stefan.steidl@informatik.uni-erlangen.de

Humaine WP5 Workshop, Belfast
2nd December 2004
Emotional Communication Skills

- Features of encoder
- Categories expressed

encoder

expressive cues

decoder

- Features of decoder
- Categories perceived

© Tanja Bänziger, Geneva Emotion Research Group
Emotional Communication Skills

- Features of encoder
- Categories expressed

- Features of decoder
- Categories perceived

- Psychologists: sensitivity of individuals to emotional expressions
- Computer scientists: performance of machine classifiers

© Tanja Bänziger, Geneva Emotion Research Group
Emotional Communication Skills

- Features of encoder
- Categories expressed

encoder

expressive cues

decoder

- Features of decoder
- Categories perceived

How good is he at this task?

- Psychologists: sensitivity of individuals to emotional expressions
- Computer scientists: performance of machine classifiers

© Tanja Bänziger, Geneva Emotion Research Group
Evaluation of the Decoder

Traditional:

- (Classwise averaged) recognition rates

Problems:

- (Hard) reference needed
- Dependent on the number and the similarity of the classes
- Confusions of similar emotions are as wrong as confusions of totally different emotions
- Independent of the “quality” of the reference
The Aibo-Emotion-Corpus

- Words labelled by 5 labellers as we do not know which emotion the child expressed
- 4 cover classes: **Anger**, **Motherese**, **Emphatic**, **Neutral**
- Majority voting as reference (3 or more labellers)

<table>
<thead>
<tr>
<th>emotion</th>
<th>frequency</th>
<th>5 of 5</th>
<th>4 of 5</th>
<th>3 of 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>anger</td>
<td>1645</td>
<td>191</td>
<td>473</td>
<td>981</td>
</tr>
<tr>
<td></td>
<td>3,3 %</td>
<td>11,6 %</td>
<td>28,7 %</td>
<td>59,5 %</td>
</tr>
<tr>
<td>emphatic</td>
<td>2528</td>
<td>67</td>
<td>550</td>
<td>1911</td>
</tr>
<tr>
<td></td>
<td>5,2 %</td>
<td>2,6 %</td>
<td>21,7 %</td>
<td>75,5 %</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>emotion</th>
<th>frequency</th>
<th>5 of 5</th>
<th>4 of 5</th>
<th>3 of 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>motherese</td>
<td>1261</td>
<td>16</td>
<td>510</td>
<td>735</td>
</tr>
<tr>
<td></td>
<td>2,6 %</td>
<td>1,2 %</td>
<td>40,4 %</td>
<td>58,2 %</td>
</tr>
<tr>
<td>neutral</td>
<td>39182</td>
<td>12654</td>
<td>16112</td>
<td>10416</td>
</tr>
<tr>
<td></td>
<td>80,9 %</td>
<td>32,2 %</td>
<td>41,1 %</td>
<td>26,5 %</td>
</tr>
</tbody>
</table>
Entropy-Based Evaluation of the Decoder

- Proposed in our ICASSP 2005 paper
- Creation of soft reference labels $l_{\text{ref}}$

\[
\begin{array}{cccccc}
L1 & L2 & L3 & L4 & \rightarrow & A & M & E & N \\
A & A & E & N & & 0,50 & 0,00 & 0,25 & 0,25 \\
\end{array}
\]

- Calculate the entropy

\[
0 \leq H(l_{\text{ref}}) = - \sum l_i \cdot \log_2 l_i \\
= -\left(\frac{1}{2} \cdot \log_2 \frac{1}{2} + \frac{1}{4} \cdot \log_2 \frac{1}{4} + \frac{1}{4} \cdot \log_2 \frac{1}{4}\right) = 1.5 \leq \log_2(4) = 2
\]

The entropy is a measure for the agreement of the labellers.
Entropy-Based Evaluation of the Decoder

- Decision of the decoder $l_{\text{dec}}$

\[
\begin{array}{c|c|c|c|c}
D & A & M & E & N \\
\hline
A & 1,00 & 0,00 & 0,00 & 0,00 \\
\end{array}
\]

- 1:1 weighting

\[ l = \frac{1}{2} \cdot l_{\text{ref}} + \frac{1}{2} \cdot l_{\text{dec}} \]

- Calculate the entropy $H(l)$

- Calculate the mean entropy of the whole data set or for a certain number of successive samples to plot histograms

- Compare the mean entropy or the histograms for different decoders
Entropy Histograms

The diagram shows the distribution of entropy values with relative frequency [%] on the y-axis and entropy on the x-axis. Two lines represent the human labeler and random choice, with the human labeler line being darker and more prominent. The graph indicates a higher frequency of entropy values around 0.8 and 1.0, with a peak in the random choice line at around 1.2.
Entropy Histograms

![Histogram of entropy with human labeler data]
Entropy Histograms

Stefan Steidl, LME
Entropy Histograms

Stefan Steidl, LME
Entropy Histograms

Stefan Steidl, LME

human labeler

rel. frequency [%]

entropy
Entropy Histograms
## Mean Entropy

<table>
<thead>
<tr>
<th>decoder</th>
<th>entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>human labeller</td>
<td>0.721</td>
</tr>
<tr>
<td>machine classifier</td>
<td>0.722</td>
</tr>
<tr>
<td>all ’neutral’</td>
<td>0.843</td>
</tr>
<tr>
<td>all ’emphatic’</td>
<td>1.049</td>
</tr>
<tr>
<td>random choice</td>
<td>1.050</td>
</tr>
<tr>
<td>all ’anger’</td>
<td>1.127</td>
</tr>
<tr>
<td>all ’motherese’</td>
<td>1.196</td>
</tr>
</tbody>
</table>
### Mean Entropy

<table>
<thead>
<tr>
<th>decoder</th>
<th>entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>human labeller</td>
<td>0.721</td>
</tr>
<tr>
<td>machine classifier</td>
<td>0.722</td>
</tr>
<tr>
<td>all 'neutral'</td>
<td>0.843</td>
</tr>
<tr>
<td>all 'emphatic'</td>
<td>1.049</td>
</tr>
<tr>
<td>random choice</td>
<td>1.050</td>
</tr>
<tr>
<td>all 'anger'</td>
<td>1.127</td>
</tr>
<tr>
<td>all 'motherese'</td>
<td>1.196</td>
</tr>
</tbody>
</table>

Is the recognition problem solved?
**Mean Entropy**

<table>
<thead>
<tr>
<th>decoder</th>
<th>entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>human labeller</td>
<td>0.721</td>
</tr>
<tr>
<td>machine classifier</td>
<td>0.722</td>
</tr>
<tr>
<td>all ’neutral’</td>
<td>0.843</td>
</tr>
<tr>
<td>all ’emphatic’</td>
<td>1.049</td>
</tr>
<tr>
<td>random choice</td>
<td>1.050</td>
</tr>
<tr>
<td>all ’anger’</td>
<td>1.127</td>
</tr>
<tr>
<td>all ’motherese’</td>
<td>1.196</td>
</tr>
</tbody>
</table>

Is the recognition problem solved? **Definitely not!**

Goal: classification as the majority does
Entropy Histogram

The figure shows the distribution of entropy values for human labelers and human majority voting. The x-axis represents the entropy values, while the y-axis shows the relative frequency in percentage. The histogram for human labelers is depicted with a solid line, and the histogram for human majority voting is shown with a dotted line.
Summary

- Evaluation of decoders with an entropy based measure which
  - uses a soft reference
  - requires $>2$ labellers
  - implicitly weights classification “errors” according to the likelihood that both emotions are confused by the labellers
End