Automatically Annotating A Five-Billion-Word Corpus of Japanese Blogs for Affect and Sentiment Analysis

Michal Ptaszynski ¹, Rafal Rzepka ², Kenji Araki ², Yoshio Momouchi ³

¹) JSPS Research Fellow / Hokkai-Gakuen University, High-Tech Research Center
²) Hokkaido University, Graduate School of Information Science and Technology
³) Hokkai-Gakuen University, Department of Electronics and Information Engineering
Presentation Outline

• Introduction
• Emotion Blog Corpora
• YACIS Corpus Description
• Affective Information Annotation Tools
• Evaluation of Affective Annotations
• Statistics of Affective Annotations
• Conclusions and Future Work
Introduction

• Recently there has been much research done on Sentiment Analysis and Affect Analysis
  – Sentiment Analysis (SA): Positive/Negative
  – Affect Analysis (AA): Emotion types
Introduction

• Recently there has been much research done on Sentiment Analysis and Affect Analysis
  – Sentiment Analysis (SA): Positive/Negative
  – Affect Analysis (AA): Emotion types
  “John loves his red shiny Porsche boxster.”
Introduction

• Recently there has been much research done on Sentiment Analysis and Affect Analysis
  – Sentiment Analysis (SA): Positive/Negative
  – Affect Analysis (AA): Emotion types
    “John loves his red shiny Porshe boxster.”

SA: attitude=Positive, opinion object=Porshe boxster
AA: emotion=love, emotion object=Porshe boxster
Introduction

• Blogs have been recognized as a good source of information on attitudes and emotions.
Introduction

• Blogs have been recognized as a good source of information on attitudes and emotions.


Introduction

• It would be good to have a big blog corpus* annotated with emotions/attitudes for:
  – User modeling
  – Emotion-aware language/conversation model generation
  – Empathic agent development
  – Deeper affect analysis (e.g., appropriateness of emotions to context, etc.)

* This research is done on Japanese language
Emotion Blog Corpora

• Let’s compare Existing Emotion Corpora
Emotion Blog Corpora

• Let’s compare Existing Emotion Corpora


Emotion Blog Corpora

- Let’s compare Existing Emotion Corpora

<table>
<thead>
<tr>
<th>Number of sentences</th>
<th>number of documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>[10] 12,724</td>
<td>500</td>
</tr>
<tr>
<td>[12] 4,186</td>
<td>249</td>
</tr>
<tr>
<td>[13] 1,191</td>
<td>(separate sentences)</td>
</tr>
<tr>
<td>[14] 5,205</td>
<td>173</td>
</tr>
<tr>
<td>[15] 12,149</td>
<td>123</td>
</tr>
<tr>
<td>[16] 4,773</td>
<td>(separate sentences)</td>
</tr>
<tr>
<td>[17] ?</td>
<td>815,494</td>
</tr>
</tbody>
</table>
## Emotion Blog Corpora

- Let’s compare Existing Emotion Corpora

<table>
<thead>
<tr>
<th>Number of sentences</th>
<th>number of documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>[10]</td>
<td>12,724</td>
</tr>
<tr>
<td>[12]</td>
<td>4,186</td>
</tr>
<tr>
<td>[13]</td>
<td>1,191</td>
</tr>
<tr>
<td>[14]</td>
<td>5,205</td>
</tr>
<tr>
<td>[15]</td>
<td>12,149</td>
</tr>
<tr>
<td>[16]</td>
<td>4,773</td>
</tr>
<tr>
<td>[17]</td>
<td>?</td>
</tr>
</tbody>
</table>

6 out of 8 are based on blogs
Emotion Blog Corpora

- Let’s compare Existing Emotion Corpora

<table>
<thead>
<tr>
<th>Number of sentences</th>
<th>number of documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>12,724</td>
<td>500</td>
</tr>
<tr>
<td>10,657</td>
<td></td>
</tr>
<tr>
<td>4,186</td>
<td></td>
</tr>
<tr>
<td>1,191</td>
<td></td>
</tr>
<tr>
<td>5,205</td>
<td></td>
</tr>
<tr>
<td>12,149</td>
<td></td>
</tr>
<tr>
<td>4,773</td>
<td></td>
</tr>
<tr>
<td>?</td>
<td>815,494</td>
</tr>
</tbody>
</table>

6 out of 8 are based on blogs
3 out of 8 are Japanese
Emotion Blog Corpora

• Let’s compare Existing Emotion Corpora

- Emotion blog corpora are rather small
- Japanese emotion blog corpora are the smallest
Emotion Blog Corpora

• Let’s compare Existing Emotion Corpora
Amount of annotations
Let’s compare Existing Emotion Corpora

<table>
<thead>
<tr>
<th>emotion classes</th>
<th>annotated affective information</th>
<th>syntactic annotations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>emotive expressions</td>
<td>emotive/ non-emot.</td>
</tr>
<tr>
<td>[10] 8</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>[11] 0</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>[12] 0</td>
<td>○</td>
<td>×</td>
</tr>
<tr>
<td>[13] 8</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>[14] 6</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>[16] 9</td>
<td>○</td>
<td>×</td>
</tr>
<tr>
<td>[17] 123</td>
<td>×</td>
<td>×</td>
</tr>
</tbody>
</table>
Emotion Blog Corpora

- Let’s compare Existing Emotion Corpora

<table>
<thead>
<tr>
<th>emotion classes</th>
<th>annotated affective information</th>
</tr>
</thead>
<tbody>
<tr>
<td>emotive/ emotions</td>
<td>emotive/ non-emot.</td>
</tr>
<tr>
<td>[10] 8</td>
<td>○</td>
</tr>
<tr>
<td>[11] 0</td>
<td>○</td>
</tr>
<tr>
<td>[12] 0</td>
<td>○</td>
</tr>
<tr>
<td>[13] 8</td>
<td>○</td>
</tr>
<tr>
<td>[14] 6</td>
<td>○</td>
</tr>
<tr>
<td>[16] 9</td>
<td>○</td>
</tr>
<tr>
<td>[17] 123</td>
<td>○</td>
</tr>
</tbody>
</table>

- Some are focused on syntax/morphology annotations
- Some are focused on emotion annotations
Emotion Blog Corpora

- We need a corpus that is:
  1. Large
  2. Annotated with syntax/morphology
  3. Annotated with emotions
Emotion Blog Corpora

How large are Japanese web-based corpora?

• JpWaC [1]
• jBlogs [2]
• KNP [3]
• Kawahara&Kurohashi [4]
• Yahoo! Blog corpus [5]
• YACIS [6]

Emotion Blog Corpora

How large are Japanese web-based corpora?
• JpWaC [1]
• jBlogs [2]
• KNP [3]
• Kawahara&Kurohashi [4]
• Yahoo! Blog corpus [5]
• YACIS [6]

Emotion Blog Corpora

<table>
<thead>
<tr>
<th>corpus name</th>
<th>scale (in words)</th>
<th>number of documents (Web pages)</th>
<th>number of sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>YACIS</td>
<td>5,600,597,095</td>
<td>12,938,606</td>
<td>354,288,529</td>
</tr>
<tr>
<td>JpWaC</td>
<td>409,384,411</td>
<td>49,544</td>
<td>12,759,201</td>
</tr>
<tr>
<td>jBlogs</td>
<td>61,885,180</td>
<td>28,530</td>
<td>[not revealed]</td>
</tr>
<tr>
<td>KNB</td>
<td>66,952</td>
<td>249</td>
<td>4,186</td>
</tr>
</tbody>
</table>

- All have: tokenization, POS, lemma,
- KNB and YACIS have: DP, NER, other.
Emotion Blog Corpora

• YACIS corpus
2. Tokenized, lemmatized, annotated with POS, dependency structure, named entities.

Emotion Blog Corpora

- YACIS is:
  1. Large ☑
  2. Annotated with syntax/morphology ☑
  3. Annotated with emotions
Affect Annotation Tools

- **Tools**
  - ML-Ask: Affect Analysis system for text-based utterances
  - CAO: Emoticon Analysis system
Affect Annotation Tools

• ML-Ask
  – Distinguish emotive sentences from non-emotive
  – Annotate emotive expressions in emotive sentences
  – Generalize emotive expressions into:
    • Emotion classes
    • Valence/activation

Affect Annotation Tools

- **ML-Ask**
  - Distinguish emotive sentences from non-emotive
  - Annotate emotive expressions in emotive sentences
  - Generalize emotive expressions into:
    - Emotion classes
    - Valence/activation


Affect Annotation Tools

• CAO
  – Detect emoticons in sentence
  – Extract/analyze emoticons
  – Annotate sentences containing emoticons with
    • Emotion classes
    • Valence/activation


Affect Annotation Tools

- CAO
  - Detect emoticons in sentence
  - Extract/analyze emoticons
  - Annotate sentences containing emoticons with
    - Emotion classes
    - Valence/activation

*) condition: sentence must contain emoticon

Affect Annotation Tools

• System output

Sentence: なぜかレディー・ガガを観ると恐怖を感じる(;´艸`)  
Spaced: なぜか レディー・ガガ を 観 と 恐怖 感じる(;´艸`)  
Transliteration: Nazeka Lady Gaga wo miru to kyoufu kanjiru (;´艸`)  
Translation: Somehow Lady Gaga frightens me (;´艸`)  

<table>
<thead>
<tr>
<th>CAO output:</th>
<th>Emotion score</th>
<th>Anger (0.00703125)</th>
<th>Sorrow (0.004665203)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extracted emoticon: (;´艸`)</td>
<td>Fear (0.02708333)</td>
<td>Shame (0.004424779)</td>
<td></td>
</tr>
<tr>
<td>Emoticon segmentation:</td>
<td>Surprize (0.01973684)</td>
<td>Joy (0.002962932)</td>
<td></td>
</tr>
<tr>
<td>S₁</td>
<td>BŁ</td>
<td>S₂</td>
<td>ELME祇</td>
</tr>
<tr>
<td>N/A</td>
<td>(</td>
<td>;</td>
<td>´艸`</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ML-Ask output:</th>
<th>なぜかレディー・ガガを観ると恐怖を感じる(;´艸`)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sentence: emotive</td>
<td>emotions: (1), FEAR:恐怖</td>
</tr>
<tr>
<td>emotemes: EMOTICON:(;´艸`)</td>
<td>2D: NEGATIVE, ACTIVE</td>
</tr>
</tbody>
</table>
Affect Annotation Tools

• System output

Sentence: なぜかレディー・ガガを見ると恐怖を感じる(;'犇'；)
Spaced: なぜか レディー・ガガ を 見る と 恐怖 感じる (;'犇'；)
Transliteration: Nazeka Lady Gaga wo miru to kyoufu kanjiru (;'犇'；)
Translation: Somehow Lady Gaga frightens me (;'犇'；)

Both systems have been evaluated separately, but not together (CAO supports ML-Ask)
Evaluation of Affective Annotations

• Data set
  – Applied earlier in evaluation of CAO in [1]
  – 1000 random mid-sized sentences from YACIS
  – 418 emotive, 582 non-emotive
  – 42 laypeople annotated emotion classes on sentences

Evaluation of Affective Annotations

• Results

<table>
<thead>
<tr>
<th></th>
<th>emotive/ non-emotive</th>
<th>emotion classes</th>
<th>2D (valence and activation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML-Ask</td>
<td>98.8%</td>
<td>73.4%</td>
<td>88.6%</td>
</tr>
<tr>
<td>CAO</td>
<td>97.6%</td>
<td>80.2%</td>
<td>94.6%</td>
</tr>
<tr>
<td>ML-Ask+CAO</td>
<td>100.0%</td>
<td>89.9%</td>
<td>97.5%</td>
</tr>
</tbody>
</table>
## Evaluation of Affective Annotations

### Results

- In dataset many sentences contained emoticons
- ML-Ask was better on blogs than on original dataset *

<table>
<thead>
<tr>
<th></th>
<th>emotive/non-emotive</th>
<th>emotion classes</th>
<th>2D (valence and activation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML-Ask</td>
<td>98.8%</td>
<td>73.4%</td>
<td>88.6%</td>
</tr>
<tr>
<td>CAO</td>
<td>97.6%</td>
<td>80.2%</td>
<td>94.6%</td>
</tr>
<tr>
<td>ML-Ask+CAO</td>
<td>100.0%</td>
<td>89.9%</td>
<td>97.5%</td>
</tr>
</tbody>
</table>

- Both systems combined were always better
- On YACIS blog corpus the two systems together have presumably good performance.

* Proof that dataset influences results.
Statistics of Affective Annotations
Statistics of Affective Annotations

- Emotive / Non emotive

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td># of emotive sentences</td>
<td>233,591,502</td>
</tr>
<tr>
<td># of non-emotive sentence</td>
<td>120,408,023</td>
</tr>
<tr>
<td>ratio (emotive/non-emotive)</td>
<td>1.94</td>
</tr>
</tbody>
</table>

- 2 times more emotive sentences than non-emotive
Statistics of Affective Annotations

• Emotion classes
Statistics of Affective Annotations

- Emotion classes
- Three emotion classes were dominant
  - Joy 喜 (+), dislike 厭 (-), fondness 好 (+)
  - Japanese generally express more positive emotions on blogs

<table>
<thead>
<tr>
<th>emotion class</th>
<th># of sentences</th>
<th>%</th>
<th>emotion class</th>
<th># of sentences</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>joy</td>
<td>16,728,452</td>
<td>31%</td>
<td>excitement</td>
<td>2,833,388</td>
<td>5%</td>
</tr>
<tr>
<td>dislike</td>
<td>10,806,765</td>
<td>20%</td>
<td>surprize</td>
<td>2,398,535</td>
<td>5%</td>
</tr>
<tr>
<td>fondness</td>
<td>9,861,466</td>
<td>19%</td>
<td>gloom</td>
<td>2,144,492</td>
<td>4%</td>
</tr>
<tr>
<td>fear</td>
<td>3,308,288</td>
<td>6%</td>
<td>anger</td>
<td>1,140,865</td>
<td>2%</td>
</tr>
<tr>
<td>relief</td>
<td>3,104,774</td>
<td>6%</td>
<td>shame</td>
<td>952,188</td>
<td>2%</td>
</tr>
</tbody>
</table>
Statistics of Affective Annotations

• Emotion classes
• Three emotion classes were dominant
  – Joy 喜 (+), dislike 厭 (-), fondness 好 (+)
  – Japanese generally express more positive emotions on blogs

* In previous research, on 2channel* two dominating emotions were: dislike (-) and excitement (+/-)

*) forum famous for rude and vulgar language

Statistics of Affective Annotations

- Emotion classes

In lexicon:

<table>
<thead>
<tr>
<th>emotion class</th>
<th>number of expressions</th>
<th>emotion class</th>
<th>number of expressions</th>
</tr>
</thead>
<tbody>
<tr>
<td>dislike</td>
<td>532</td>
<td>fondness</td>
<td>197</td>
</tr>
<tr>
<td>excitement</td>
<td>269</td>
<td>fear</td>
<td>147</td>
</tr>
<tr>
<td>sadness</td>
<td>232</td>
<td>surprise</td>
<td>129</td>
</tr>
<tr>
<td>joy</td>
<td>224</td>
<td>relief</td>
<td>106</td>
</tr>
<tr>
<td>anger</td>
<td>199</td>
<td>shame</td>
<td>65</td>
</tr>
</tbody>
</table>

But what if number of expressions in lexicon influences results? (more expr. = ? more annot.)

- In YACIS corpus

<table>
<thead>
<tr>
<th>emotion class</th>
<th># of sentences</th>
<th>%</th>
<th>emotion class</th>
<th># of sentences</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>joy</td>
<td>16,728,452</td>
<td>31%</td>
<td>excitement</td>
<td>2,833,388</td>
<td>5%</td>
</tr>
<tr>
<td>dislike</td>
<td>10,806,765</td>
<td>20%</td>
<td>surprise</td>
<td>2,398,535</td>
<td>5%</td>
</tr>
<tr>
<td>fondness</td>
<td>9,861,466</td>
<td>19%</td>
<td>gloom</td>
<td>2,144,492</td>
<td>4%</td>
</tr>
<tr>
<td>fear</td>
<td>3,308,288</td>
<td>6%</td>
<td>anger</td>
<td>1,140,865</td>
<td>2%</td>
</tr>
<tr>
<td>relief</td>
<td>3,104,774</td>
<td>6%</td>
<td>shame</td>
<td>952,188</td>
<td>2%</td>
</tr>
</tbody>
</table>
Statistics of Affective Annotations

- Emotion classes

In lexicon:

<table>
<thead>
<tr>
<th>emotion class</th>
<th>number of expressions</th>
<th>emotion class</th>
<th>number of expressions</th>
</tr>
</thead>
<tbody>
<tr>
<td>dislike</td>
<td>532</td>
<td>fondness</td>
<td>197</td>
</tr>
<tr>
<td>excitement</td>
<td>269</td>
<td>fear</td>
<td>147</td>
</tr>
<tr>
<td>sadness</td>
<td>232</td>
<td>surprise</td>
<td>129</td>
</tr>
<tr>
<td>joy</td>
<td>224</td>
<td>relief</td>
<td>106</td>
</tr>
<tr>
<td>anger</td>
<td>199</td>
<td>shame</td>
<td>65</td>
</tr>
</tbody>
</table>

* No significant correlation between number of words in lexicon and frequency of emotion class! *
* $\rho=0.38$

Its good because it means lexicon does not influence the results.
Statistics of Affective Annotations

• Comparison with other corpora
Statistics of Affective Annotations

• YACIS and KNB
  – KNB:
    – 249 pages, 67,000 words
    – No emotion types, but valence/attitude and opinion-related annotations
  – Compared:
    POSITIVE SENTENCES vs. NEGATIVE SENTENCES in YACIS and KNB

Statistics of Affective Annotations

- YACIS and KNB

- Similar ratio was observed for the two blog corpora: large (YACIS) and small (KNB).

<table>
<thead>
<tr>
<th></th>
<th>positive</th>
<th>negative</th>
<th>ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>KNB</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>emotional attitude</td>
<td>317</td>
<td>208</td>
<td>1.52</td>
</tr>
<tr>
<td>opinion</td>
<td>489</td>
<td>289</td>
<td>1.69</td>
</tr>
<tr>
<td>merit</td>
<td>449</td>
<td>264</td>
<td>1.70</td>
</tr>
<tr>
<td>acceptance or rejection</td>
<td>125</td>
<td>41</td>
<td>3.05</td>
</tr>
<tr>
<td>event</td>
<td>43</td>
<td>63</td>
<td>0.68</td>
</tr>
<tr>
<td>sum</td>
<td>1,423</td>
<td>865</td>
<td>1.65</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>YACIS</strong></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>only</td>
<td>22,381,992</td>
<td>12,837,728</td>
<td>1.74</td>
</tr>
<tr>
<td>only+ mostly</td>
<td>23,753,762</td>
<td>13,605,514</td>
<td>1.75</td>
</tr>
<tr>
<td>only+ mostly</td>
<td>31,071,945</td>
<td>17,496,901</td>
<td>1.78</td>
</tr>
<tr>
<td>only+ mostly</td>
<td>32,752,589</td>
<td>18,442,602</td>
<td>1.78</td>
</tr>
</tbody>
</table>

Statistics of Affective Annotations

- YACIS and KNB

Japanese generally express more positive emotions on blogs. (confirmation)

Statistics of Affective Annotations

• YACIS and Minato et al.
  – Corpus by Minato et al. (2006) [1]
    • Examples (sentences) gathered from Emotion Dictionary [2] and analyzed.
    • Annotations (in general):
      – Emotion classes on sentences
      – POS

Statistics of Affective Annotations

• YACIS and Minato et al.
  – Compared: DISTRIBUTION OF EMOTION CLASSES
  – Between: YACIS, Minato et al. and Nakamura’s Emotion Dictionary[3] *

*) dictionary used as affect lexicon in ML-Ask

Statistics of Affective Annotations

- **YACIS vs. Minato et al. vs. Nakamura**

<table>
<thead>
<tr>
<th></th>
<th>Minato et al.</th>
<th>YACIS</th>
<th>Nakamura</th>
</tr>
</thead>
<tbody>
<tr>
<td>dislike</td>
<td>355</td>
<td>14,184,697</td>
<td>532</td>
</tr>
<tr>
<td>joy</td>
<td>295</td>
<td>22,100,500</td>
<td>224</td>
</tr>
<tr>
<td>fondness</td>
<td>205</td>
<td>13,817,116</td>
<td>197</td>
</tr>
<tr>
<td>sorrow</td>
<td>205</td>
<td>2,881,166</td>
<td>232</td>
</tr>
<tr>
<td>anger</td>
<td>160</td>
<td>1,564,059</td>
<td>199</td>
</tr>
<tr>
<td>fear</td>
<td>145</td>
<td>4,496,250</td>
<td>147</td>
</tr>
<tr>
<td>surprise</td>
<td>25</td>
<td>3,108,017</td>
<td>129</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Minato et al. and Nakamura</th>
<th>Minato et al. and YACIS</th>
<th>YACIS and Nakamura</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spearman’s ρ</td>
<td>0.88</td>
<td>0.63</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Statistics of Affective Annotations

- **YACIS vs. Minato et al. vs. Nakamura**

1. **No significant correlation between:**
   - YACIS and Nakamura *

   * confirmation of previous calculation
good because lexicon does not influence the results

<table>
<thead>
<tr>
<th></th>
<th>Minato et al.</th>
<th>YACIS</th>
<th>Nakamura</th>
<th>Spearman’s $\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>dislike</td>
<td>355</td>
<td>14,184,697</td>
<td>532</td>
<td>0.88</td>
</tr>
<tr>
<td>joy</td>
<td>295</td>
<td>22,100,500</td>
<td>224</td>
<td></td>
</tr>
<tr>
<td>fondness</td>
<td>205</td>
<td>13,817,116</td>
<td>197</td>
<td></td>
</tr>
<tr>
<td>sorrow</td>
<td>205</td>
<td>2,881,166</td>
<td>232</td>
<td></td>
</tr>
<tr>
<td>anger</td>
<td>160</td>
<td>1,564,059</td>
<td>199</td>
<td></td>
</tr>
<tr>
<td>fear</td>
<td>145</td>
<td>4,496,250</td>
<td>147</td>
<td></td>
</tr>
<tr>
<td>surprise</td>
<td>25</td>
<td>3,108,017</td>
<td>129</td>
<td></td>
</tr>
</tbody>
</table>

Statistics of Affective Annotations

- YACIS vs. Minato et al. vs. Nakamura

2. Medium correlation between YACIS and Minato (could be some similarities, but nothing sure)

<table>
<thead>
<tr>
<th></th>
<th>Minato et al.</th>
<th>YACIS</th>
<th>Nakamura</th>
</tr>
</thead>
<tbody>
<tr>
<td>dislike</td>
<td>355</td>
<td>14,184,697</td>
<td>532</td>
</tr>
<tr>
<td>joy</td>
<td>295</td>
<td>22,100,500</td>
<td>224</td>
</tr>
<tr>
<td>fondness</td>
<td>205</td>
<td>13,817,116</td>
<td>197</td>
</tr>
<tr>
<td>sorrow</td>
<td>205</td>
<td>2,881,166</td>
<td>232</td>
</tr>
<tr>
<td>anger</td>
<td>160</td>
<td>1,564,056</td>
<td>199</td>
</tr>
<tr>
<td>fear</td>
<td>145</td>
<td>4,496,250</td>
<td>147</td>
</tr>
<tr>
<td>surprise</td>
<td>25</td>
<td>3,108,017</td>
<td>129</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Minato et al. and Nakamura</th>
<th>Minato et al. and YACIS</th>
<th>YACIS and Nakamura</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spearman’s $\rho$</td>
<td>0.88</td>
<td>0.63</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Statistics of Affective Annotations

YACIS vs. Minato et al. vs. Nakamura

3. STRONG correlation between Minato and Nakamura
(both are dictionaries, but differ in: time, media, collecting person’s background, approach, assumptions......)

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Minato et al.</th>
<th>YACIS</th>
<th>Nakamura</th>
</tr>
</thead>
<tbody>
<tr>
<td>dislike</td>
<td>355</td>
<td>14,184,697</td>
<td>532</td>
</tr>
<tr>
<td>joy</td>
<td>295</td>
<td>22,100,500</td>
<td>224</td>
</tr>
<tr>
<td>fondness</td>
<td>205</td>
<td>13,817,116</td>
<td>197</td>
</tr>
<tr>
<td>sorrow</td>
<td>205</td>
<td>2,881,166</td>
<td>232</td>
</tr>
<tr>
<td>anger</td>
<td>160</td>
<td>1,564,059</td>
<td>199</td>
</tr>
<tr>
<td>fear</td>
<td>145</td>
<td>4,496,250</td>
<td>147</td>
</tr>
<tr>
<td>surprise</td>
<td>25</td>
<td>3,108,017</td>
<td>129</td>
</tr>
</tbody>
</table>

Spearman’s $\rho$

<table>
<thead>
<tr>
<th>Minato et al. and Nakamura</th>
<th>Minato et al. and YACIS</th>
<th>YACIS and Nakamura</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.88</td>
<td>0.63</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Statistics of Affective Annotations

• What does it mean?
  – There could be a tendency in Japanese language in general to produce more expressions of some emotions (dislike 厳・嫌, excitement 昂, sadness 哀・悲, joy 喜,) than the other (surprise 驚, fear 怖, anger 怒)
  – The number of expressions in language is not equivalent to the frequency of usage
    • Some expressions are used more frequently (relief, fear) than the others (anger sadness)
Conclusions

• Presented a study on Emotion Corpora with a focus on Emotion Blog Corpora
• Used 2 tools for Affective Information Annotation (ML-Ask and CAO) to Annotate YACIS
  – The tools showed good performance on YACIS
• Presented Statistics of Affective Annotations
Conclusions

• Statistics of Affective Annotations
  – Japanese express on blogs more positive emotions than negative
  – Japanese language seems to have a tendency to generate more expressions for some emotion types than the others
  – However, number of expressions and usage frequency are not related
    • Emotion class with many expressions is sometimes expressed rarely (sadness, anger),
    • Emotion class with few expressions is sometimes expressed more often (relief, fear)
Future Work

• Online interface!
• More detailed evaluation (e.g. of dependency)
• Lexicon generation (Sketch Engine?)
• N-gram version for download without limitations (like Google n-gram)
• Applications
  – Affect lexicon enlargement
  – User modeling
  – Emotion-aware language/conversation model generation
  – Empathic dialog agent development (no need to Google)
  – Deeper affect analysis (e.g., appropriateness of emotions to context, etc.)
  – ...

http://www.sketchengine.co.uk/
Thank you for your attention!

Michal Ptaszynski
ptaszynski@ieee.org
Discussion

• Copyrights
  – YACIS will not be put on sale
  – Only for scientific purposes
  – Usage of corpus will need a two-side agreement

• Gathering of the corpus is similar to search engines
  – If YACIS was illegal, Google, Yahoo,... would be even more illegal.
### Web-crawled corpora

9 / 11 of >1 bil. corpora are Web-crawled

<table>
<thead>
<tr>
<th>corpus name</th>
<th>scale (in words)</th>
<th>language</th>
<th>domain</th>
<th>annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liu&amp;Curran [23]</td>
<td>10 billion</td>
<td>English</td>
<td>whole Web</td>
<td>tokenization;</td>
</tr>
<tr>
<td>YACIS</td>
<td>5.6 billion</td>
<td>Japanese</td>
<td>Blogs (Ameba)</td>
<td>parsing, NER, affect (emotive expressions, Russell-2D, emotion objects);</td>
</tr>
<tr>
<td>BiWeC [21]</td>
<td>5.5 billion</td>
<td>English</td>
<td>whole Web (.uk and .au domains)</td>
<td>POS, lemma;</td>
</tr>
<tr>
<td>ukWaC</td>
<td>2 billion</td>
<td>English</td>
<td>whole Web (.uk domain)</td>
<td>POS, lemma;</td>
</tr>
<tr>
<td>PukWaC (Parsed-ukWaC) [27]</td>
<td>2 billion</td>
<td>English</td>
<td>whole Web (.uk domain)</td>
<td>POS, lemma, dependency parsing;</td>
</tr>
<tr>
<td>itWaC [20], [27]</td>
<td>2 billion</td>
<td>Italian</td>
<td>whole Web (.it domain)</td>
<td>POS, lemma;</td>
</tr>
<tr>
<td>Gigaword [32]</td>
<td>2 billion</td>
<td>Hungarian</td>
<td>whole Web (.hu domain)</td>
<td>POS, lemma;</td>
</tr>
<tr>
<td>deWaC [27]</td>
<td>1.7 billion</td>
<td>German</td>
<td>whole Web (.de domain)</td>
<td>POS, lemma;</td>
</tr>
<tr>
<td>frWaC [27]</td>
<td>1.6 billion</td>
<td>French</td>
<td>whole Web (.fr domain)</td>
<td>POS, lemma;</td>
</tr>
<tr>
<td>Corpus</td>
<td>1 billion</td>
<td>Brazilian</td>
<td>multi-domain (newspapers, Web,</td>
<td>tokenization, sentence segmentation;</td>
</tr>
<tr>
<td>Brasiliero [40]</td>
<td></td>
<td></td>
<td>talk transcriptions)</td>
<td>POS, lemma;</td>
</tr>
<tr>
<td>National Corpus of Polish [41]</td>
<td></td>
<td></td>
<td>multi-domain (newspapers,</td>
<td>POS, lemma, dependency parsing, named entities, word senses;</td>
</tr>
<tr>
<td>JpWaC [31]</td>
<td>400 million</td>
<td>Japanese</td>
<td>literature, Web, etc.)</td>
<td>tokenization, POS, lemma;</td>
</tr>
<tr>
<td>jBlogs [31]</td>
<td>62 million</td>
<td>Japanese</td>
<td>Blogs (Ameba, Goo, Livedoor,</td>
<td>tokenization, POS, lemma;</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Yahoo!)</td>
<td></td>
</tr>
</tbody>
</table>
### Web-crawled corpora

9 / 11 of >1 bil. corpora are Web-crawled

<table>
<thead>
<tr>
<th>corpus name</th>
<th>scale (in words)</th>
<th>language</th>
<th>domain</th>
<th>annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liu&amp;Curran [23]</td>
<td>10 billion</td>
<td>English</td>
<td>whole Web</td>
<td>tokenization; POS, lemma, dependency parsing, NER, affects (emotive expressions, Russell-2D, emotion objects);</td>
</tr>
<tr>
<td>YACIS</td>
<td>5.6 billion</td>
<td>Japanese</td>
<td>Blogs (Ameba)</td>
<td>POS, lemma;</td>
</tr>
<tr>
<td>BiWeC [21]</td>
<td>5.5 billion</td>
<td>English</td>
<td>whole Web (.uk and .au domains)</td>
<td>POS, lemma;</td>
</tr>
<tr>
<td>ukWaC</td>
<td>2 billion</td>
<td>English</td>
<td>whole Web (.uk domain)</td>
<td>POS, lemma;</td>
</tr>
<tr>
<td>PukWaC (Parsed-ukWaC) [27]</td>
<td>2 billion</td>
<td>English</td>
<td>whole Web (.uk domain)</td>
<td>POS, lemma;</td>
</tr>
<tr>
<td>itWaC [20, 27]</td>
<td>2 billion</td>
<td>Italian</td>
<td>whole Web (.it domain)</td>
<td>POS, lemma;</td>
</tr>
<tr>
<td>Gigaword [32]</td>
<td>2 billion</td>
<td>Hungarian</td>
<td>whole Web (.hu domain)</td>
<td>POS, lemma;</td>
</tr>
<tr>
<td>deWaC [27]</td>
<td>1.7 billion</td>
<td>German</td>
<td>whole Web (.de domain)</td>
<td>POS, lemma;</td>
</tr>
<tr>
<td>frWaC [27]</td>
<td>1.6 billion</td>
<td>French</td>
<td>whole Web (.fr domain)</td>
<td>POS, lemma;</td>
</tr>
<tr>
<td>Corpus</td>
<td>1 billion</td>
<td>Brazilian</td>
<td>multi-domain (newspapers, Web, talk transcriptions)</td>
<td>POS, lemma;</td>
</tr>
<tr>
<td>Brasiliero [40]</td>
<td>1 billion</td>
<td>Portuguese</td>
<td>multi-domain (newspapers, literature, Web, etc.)</td>
<td>POS, lemma;</td>
</tr>
<tr>
<td>National Corpus of Polish [41]</td>
<td>1 billion</td>
<td>Polish</td>
<td>multi-domain (newspapers, literature, Web, etc.)</td>
<td>POS, lemma;</td>
</tr>
<tr>
<td>JpWaC [31]</td>
<td>400 million</td>
<td>Japanese</td>
<td>whole Web (.jp domain)</td>
<td>POS, lemma;</td>
</tr>
<tr>
<td>jBlogs [31]</td>
<td>62 million</td>
<td>Japanese</td>
<td>Blogs (Ameba, Goo, Livedoor, Yahoo!)</td>
<td>POS, lemma;</td>
</tr>
</tbody>
</table>
Web-crawled corpora

9 / 11 of >1 bil. corpora are Web-crawled

Web presumably contains more text than written data

Most of our culture exists on the Web(?)
Japanese Web/blog-based corpora

<table>
<thead>
<tr>
<th>corpus name</th>
<th>scale (in words)</th>
<th>number of documents (Web pages)</th>
<th>number of sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>YACIS</td>
<td>5,600,597,095</td>
<td>12,938,606</td>
<td>354,288,529</td>
</tr>
<tr>
<td>JpWaC</td>
<td>409,384,411</td>
<td>49,544</td>
<td>12,759,201</td>
</tr>
<tr>
<td>jBlogs</td>
<td>61,885,180</td>
<td>28,530</td>
<td>[not revealed]</td>
</tr>
<tr>
<td>KNB</td>
<td>66,952</td>
<td>249</td>
<td>4,186</td>
</tr>
</tbody>
</table>
### Japanese Web/blog-based corpora

<table>
<thead>
<tr>
<th>corpus name</th>
<th>size (uncompressed in GB, text only, no annotation)</th>
<th>domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>YACIS</td>
<td>26.6</td>
<td>Blogs (Ameba); Whole Web (11 domains within .jp); Blogs (Ameba, Goo, Livedoor, Yahoo!); Blogs (written by students);</td>
</tr>
<tr>
<td>JpWaC</td>
<td>7.3</td>
<td></td>
</tr>
<tr>
<td>jBlogs</td>
<td>.25 (compressed)</td>
<td></td>
</tr>
<tr>
<td>KNB</td>
<td>450 kB</td>
<td></td>
</tr>
</tbody>
</table>
How large are large corpora?

- The notion of a “large scale corpus” has appeared in linguistic and computational linguistic literature for many years.
  - (perhaps) first use of phrase “large scale corpus” 1987
How large are large corpora?

• The notion of a “large scale corpus” has appeared in linguistic and computational linguistic literature for many years.
The notion of a “large scale corpus” has appeared in linguistic and computational linguistic literature for many years. The number of papers on large scale corpora has increased significantly over time, with a peak in the late 1990s and early 2000s.
How large are large corpora?

- The notion of a “large scale corpus” has appeared in linguistic and computational linguistic literature for many years.

- **1992:** 1 million word corpus
- **2001:** 70 million word corpus
- **2006:** 2 billion word corpus
- **2009:** 5 billion word corpus

- **March 2012:** 29 papers
- **July 2012:** 97 papers
ML-Ask – Automatic affect annotation tool

- Our approach to affect analysis:

In language there are:

1. emotive/evaluative expressions*
2. emotiveness indicators. “Emotemes” – Japanese emotive morphemes**

“Oh, but John is such a nice person!”

“Oh, but John is such a rude person!”

**) M. Ptaszyński, Moeru gengo - Intânetto kei-jiban no ue no nihongo kaiwa ni okeru kanjōhyōgen no kōzō to kigőrontekikinō no bunseki – “2channeru, denshikeijiban o rei toshite” –(Boisterous language. Analysis of structures and semiotic functions of emotive expressions in conversation on Japanese Internet bulletin board forum - 2channel -), UAM, Poznań (2006)


Michał Ptaszyński, Paweł Dybala, Rafał Rzepka and Kenji Araki. Affecting Corpora:Experiments with Automatic Affect Annotation System - A Case Study of the 2channel Forum -, The Conference of the Pacific Association for Computational Linguistics (PACLING-09), September 1-4, 2009, Hokkaido University, Sapporo, Japan
ML-Ask – Automatic affect annotation tool

Gathered manually (907 items)

emotems DB

exclamatives
すげえ sugee (great!)
うおお wooo (whoa!)

mimetics (gitaigo)
ワクワク wakuwaku (heart pounding)
ドキドキ dokidoki (go pit-a-pat)

vulgarities
-やがる -yagaru (fu**ing do sth)
くそ kuso (shit)
馬鹿 baka (stupid)

hypocorystics
-ちゃん -chan (name suffix)

textual representations of voice modulation and body language (emoticons)
“!” “??”, “…” , (T_T), (-д-), __[ ]__〇

Nakamura’s dictionary (2100 items)

emotive expressions DB

nouns
愛情 aijou (love)
安心 anshin (relief)
恐怖 kyofu (fear)

verbs
喜ぶ yorokobu (be glad)
悲しむ kanashimu (feel sad)
むかつく mukatsuku (get angry)

phrases / idioms
虫酸が走る mushizu ga hashiru (give one the creeps)
心が解ける kokoro ga tokeru (one’s heart is melting in relief)
歓天喜地 kantenkichi (delight larger than Haven and Earth)

adjectives
嬉しい ureshii (happy)
悔しい kuyashii (mortifying)
怖い kowai (scary)

10-type emotion classification:
1. Joy, delight
2. Anger
3. Sorrow, sadness, gloom
4. Fear
5. Shame, shyness, bashfulness
6. Liking, fondness
7. Dislike, detestation
8. Excitement
9. Relief
10. Surprise, amazement

input

emotive expressions DB

output

emotems DB

Found emotemes: ne, !
(for English: oh, so-)
Utterance is: emotive
Found emotive expressions: omoshiroi
(interesting)
Conveyed emotion types: joy (positive)
Problematic inputs:

あんまり面白くなかったな…
Anmari *omoshiroku nakatta na…*
Oh, it wasn’t that interesting...

+ CVS
+ Russell’s 2D affect model

activated

kou / takaburi (excitement)
kyou / odoroki (surprise, amazement)
chi / haji (shame, shyness, bashfulness)

negative

do / ikari (anger)
fu / kowagari (fear)
en / iya (dislike, detestation)

ai / aware (sorrow, sadness, gloom)

positive

ki / yorokobi (joy, delight)
kou / suki (liking, fondness)
an / yasuragi (relief)

deactivated
CAO

Emoticons:

• Only for 1-line Eastern emoticons
Divide emoticons into semantic areas (mouths, eyes, etc. and calculate their emotion score.)

Deeper Affect Analysis

• Contextual Affect Analysis


Potential Critical Errors in Real World Tasks

• Facial expressions

• Expression:
  User is crying
    – (presence of tears and facial expression);

• Assumption:
  User is sad;
Potential Critical Errors in Real World Tasks

• Facial expressions

• **Context**: The user is cutting an onion in the kitchen;

• Assumption: User is sad;
Potential Critical Errors in Real World Tasks

• **Language**

• **Expression**: User has used the word “happy”; (嬉しい、すっきり)

• **Assumption**: User is happy;
Potential Critical Errors in Real World Tasks

• Language

Context:

- Expression: User has used the word "happy";
- Assumption: User is happy;

1. "I’m not happy"
   あまり嬉しくないな。
2. "I'm so happy that bastard was hit by a car!"
   あいつが車に引かれたと聞いてすっきりした。
Potential Critical Errors in Real World Tasks

- Language

**Context:**

1. "I’m not happy"
   あまり嬉しくないな。
2. "I'm so happy that bastard was hit by a car!"
   あいつが車に引かれたと聞いてすっきりした。

**Assumption:**

User is happy;

**Grammatical context**

**Social/conceptual context**
Contextual Appropriateness of Emotions

– 試験に合格してうれしい！ [joy, happiness]
  “Oh, I’m so happy (because) I passed the exam!”

– 彼女に振られて悲しい... [depression]
  “Oh, I’m is so depressed (because) my girlfriend left...”
Contextual Appropriateness of Emotions

- 試験に合格してうれしい！  [joy, happiness]
  “Oh, I’m so happy (because) I passed the exam!”
- あの野郎が車に引かれたと聞いてすっきり！  [joy, happiness]
  “Oh, I’m so happy (because) that bastard was hit by a car!”
- 彼女に振られて悲しい…  [depression]
  “Oh, I’m is so depressed (because) my girlfriend left...”
- バレンタイン・デーが来るから悲しいね…  [depression]
  “Oh, I’m so depressed (because) the Valentine’s Day is coming...”
Contextual Appropriateness of Emotions

– 試験に合格してうれしい！
  “Oh, I’m so happy (because) I passed the exam!”  [joy, happiness]

– あの野郎が車に引かれたと聞いてすっきり!
  “Oh, I’m so happy (because) that bastard was hit by a car!”  [joy, happiness]

– 彼女に振られて悲しい...
  “Oh, I’m is so depressed (because) my girlfriend left...”  [depression]

– バレンタイン・デーが来るから悲しいね...
  “Oh, I’m so depressed (because) the Valentine’s Day is coming...”  [depression]
Contextual Appropriateness of Emotions

cause of emotion = context of expression of emotion
makes the expression either Appropriate or Inappropriate
Contextual Appropriateness of Emotions

“Oh, I’m so happy (because) I passed the exam!” [joy, happiness], 試験に合格してうれしい！

“Oh, I’m so happy (because) that bastard was hit by a car!” [joy, happiness]

“Oh, I’m so depressed (because) my girlfriend left…” 彼女に振られて悲しい…

“Oh, I’m so depressed (because) the Valentine’s Day is coming…” バレンタイン・デーが来るから悲しいね…
Contextual Appropriateness of Emotions

“...(because) I passed the exam!”

“I’m so happy…”
“...glad…”
“He looked so happy…”
“My mom was so happy…”

context of expression of emotion

“試験に合格して” / “I passed the exam”
Contextual Appropriateness of Emotions

"...(because) I passed the exam!"

"I’m so happy…"
"I’m so glad…"
"He looked so happy…"
"My mom was so happy…"

WEB MINING

emotive expression DB

cause of emotion = context of expression of emotion

List of emotions appearing most often = appropriate /natural/ for this context
Contextual Appropriateness of Emotions

List of emotions appearing most often = appropriate /natural/ for this context

Confront expression from the sentence with the list

“I’m so happy (because) I passed the exam!”
## Contextual Appropriateness of Emotions

<table>
<thead>
<tr>
<th>Negative</th>
<th>Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Inappropriate</strong></td>
<td><strong>Appropriate</strong></td>
</tr>
<tr>
<td><strong>Appropriate</strong></td>
<td></td>
</tr>
<tr>
<td><strong>ML-Ask:</strong> dislike;</td>
<td>User: I’m so happy I passed the exam!</td>
</tr>
<tr>
<td><strong>Web mining:</strong> anger, dislike</td>
<td>ML-Ask: joy;</td>
</tr>
<tr>
<td><strong>Agent:</strong> Yeah, you have a reason to</td>
<td>Web mining: joy, excitement</td>
</tr>
<tr>
<td>be angry.</td>
<td>Agent: Yeah! That’s great! [sympathy]</td>
</tr>
<tr>
<td><strong>[empathy]</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Inappropriate</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Appropriate</strong></td>
<td></td>
</tr>
<tr>
<td><strong>ML-Ask:</strong> dislike, depression;</td>
<td>User: I’d be happy if that bastard was hit by a car!</td>
</tr>
<tr>
<td><strong>Web mining:</strong> excitement, joy</td>
<td>ML-Ask: joy;</td>
</tr>
<tr>
<td><strong>Agent:</strong> You should be happy!</td>
<td>Web mining: fear, sadness</td>
</tr>
<tr>
<td><strong>[consolation]</strong></td>
<td>Agent: Are you sure this is what you really feel? [counsel]</td>
</tr>
</tbody>
</table>

Every emotion can be either appropriate or not.
Results

1. Contextual Appropriateness Verification of emotion types
   • 70%

2. Contextual Appropriateness Verification of emotion valence
   • 80%

Michal Ptaszynski, Michal Mazur, Pawel Dybala, Rafal Rzepka, Kenji Araki and Yoshio Momouchi: “Towards Computational Fronesis: Verifying Contextual Appropriateness of Emotions” International Journal of Distance Education Technologies, Special Issue on Emotional Intelligence for Online Learning, 2012 (to appear)