Method for Building a Multidimensional Affect Dictionary for a New Language Semi-automatically

Guillaume Pitel
Gregory Grefenstette
CEA LIST, France

LREC 2008
Marrakesh, Morocco

Contacts: guillaume.pitel@gmail.com, gregory.grenestette@cea.fr

Acknowledgments: Fondation Lagardère, ARC RAPSODIS (LORIA-INRIA Grand Est)
Maybe it was the unfriendly attitude of those hanging around the old complex.

Things started to make sense in November 2000, when authorities raided the site -- and said they found enough chemicals to make millions of doses of LSD.

"My husband and I started asking ourselves why they were working in the middle of the night. We thought it was pretty strange," said Lori Morrissey, who lives adjacent to the fenced, 26-acre site in a rural area slowly being overtaken by homes and families.
1. Lasswell Value Dictionary (1969)

✓ Eight dimensions:

- WEALTH, POWER, RECTITUDE, RESPECT, ENLIGHTENMENT, SKILL, AFFECTION, AND WELLBEING with positive or negative orientation
- e.g., admire: RESPECT (positive)

2. General Inquirer dictionary (Stone, et al. 1965) 9051 headwords

- 1,915 positive and 2,291 negative words (Pos/Neg)
- also labels: Active, Passive, ... , Pleasure, Pain, ... Human, Animate, ..., Region, Route,..., Fetch, Stay, ...

http://www.wjh.harvard.edu/~inquirer/inqdict.txt
<table>
<thead>
<tr>
<th>Lexical Entry</th>
<th>Part of Speech</th>
<th>Class</th>
<th>Centrality</th>
<th>Intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>arrogance</td>
<td>sn</td>
<td></td>
<td>0.7</td>
<td>0.9</td>
</tr>
<tr>
<td>gleeful</td>
<td>adj</td>
<td>“happiness”</td>
<td>0.7</td>
<td>0.6</td>
</tr>
<tr>
<td>gleeful</td>
<td>adj</td>
<td>“excitement”</td>
<td>0.3</td>
<td>0.6</td>
</tr>
</tbody>
</table>

42 pair affect classes (positive/negative)

http://www.infonortics.com/searchengines/sh01/slides-01/evans_files/v3_document.htm
Building an Affect Lexicon for a new Language

1. Define Affect dimensions (manual step)
   ✓ 3 hours

2. Choose a small set of Seed Words for each dimension endpoint (manual)
   ✓ One day
   ✓ We chose two sizes of « small »: 2-5 or 10

3. (For testing: create Gold Standard)
   ✓ ~5000 word-to-class mappings: 2 weeks
   ✓ Only 1 native speaker

4. Discover possible affect words (automatic)

5. Place candidates along axes (automatic)
Table 1: The 44 affect axes chosen for our classification experiment. Each axis has a positive and a negative pole. In parentheses is the number of words manually affected to each axis pole.
Choose Seed Words

1. Avantage (advantage)
   ✓ Avantage
   ✓ Avantageux
   ✓ Avantager

2. Désavantage (disadvantage)
   ✓ Désavantage
   ✓ Désavantager
   ✓ Désavantagée
   ✓ Défavoriser
   ✓ Défavorisée

- Find prototypical noun, adjective, verb
- Expanded using synonym dictionary and manual filtering
We tested 3 methods for placing candidates along their axes

1. **SL-PMI : Semantic Likeliness Pointwise Mutual Information from Information Retrieval**
   - Using the SemanticMap, a resource built from the Web.
   
   $$\text{SL-PMI}_C(w) = \frac{1}{|C|} \sum_{c \in C} \log_2 \frac{\epsilon + H_\delta(w, c)^2}{\epsilon + H_\delta(w, \ast)H_\delta(c, \ast)}$$
   
   Where $H_\delta(w_1, w_2)$ is the number of cooccurrences of words $w_1$ and $w_2$ in a $\delta$ words window.

2. **SL-LSA : Semantic Likeliness using LSA similarity measure**
   - Average cosine distance
   - With windows : [-2,+2], [-5,+5], [-10,+10], [-30,+30]
   - Using InfomapNLP + Europarl/French
We tested 3 methods for placing candidates along their axes

3. SL-dLSA+SVM : Semantic Likeliness from diversified Latent Semantic Analysis (LSA) and Support Vector Machines (SVM)
   ✓ Create forty-two 300-dimension LSA spaces
     ▪ Varying window size (14) × symmetry (3)
       → Window size: $\delta = [1…10, 15, 20, 25, 30]$
       → Windows : $[0, +\delta]$ $[-\delta, + \delta]$ $[-\delta, 0]$
   ✓ Concatenate spaces for each word (12600 dim)
   ✓ Train a 44-classes SVM classifier
SL-dLSA+SVM

This is my text and I love it because it is the best text ever...

This is my text and I love it because it is the best text ever...

This is my text and love it because it is the best text ever...

dLSA word signature: + + + + +...
Evaluation: Using five seed words per class

![Graph showing evaluation results with various methods.](image-url)
Evaluation: Using twenty seed words per class
Improvement (from 5 seeds to 20)
### Good Example of Classifying a New Emotive Word

<table>
<thead>
<tr>
<th>Class</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>27 Health/Sickness</td>
<td>.105</td>
</tr>
<tr>
<td>01 Advantage/Disadvantage</td>
<td>.065</td>
</tr>
<tr>
<td>09 Comfort/Irritation</td>
<td>.065</td>
</tr>
<tr>
<td>07 Clarity/Confusion</td>
<td>.062</td>
</tr>
<tr>
<td>22 Promise/Warning</td>
<td>.056</td>
</tr>
<tr>
<td>36 Sensitivity/Insensitivity</td>
<td>.042</td>
</tr>
<tr>
<td>03 Amity/Anger</td>
<td>.040</td>
</tr>
<tr>
<td><strong>16 Pleasure/Pain</strong></td>
<td>.037</td>
</tr>
<tr>
<td>21 Excitement/Boredom</td>
<td>.035</td>
</tr>
<tr>
<td>24 Public-spiritedness/Crime</td>
<td>.034</td>
</tr>
</tbody>
</table>

Table 4: Classification of the word “désagrément” using SL-dLSA+SVM with \( \#Classes = 10 \). Based on the gold standard L3, gold standard classes (in bold) for this word were 01 (Advantage/Disadvantage) and 16 (Pleasure/Pain). Other a posteriori acceptable classes numbers are in italic.
Table 5: Classification of the word “disgrâce” using SL-dLSA+SVM with L2, \#Classes = 10. Based on the gold standard L3, the manually assigned class for this word (01 Advantage/Disadvantage) is absent. A posteriori acceptable classes numbers are in italic.
Conclusions

1. An affect dictionary can be built rapidly for a new language using a little manual labor and semi-automatic techniques over a large corpus
   ✓ Best method: 10 times better than baseline
   ✓ Learning from 20 words per semantic axis is better than 5 (for all methods)

2. Semantic Likeliness (SL) from diversified Latent Semantic Analysis (dLSA) and Support Vector Machines (SVM) benefits more from more learning data than SL-PMI or SL-LSA
   ✓ Because of SVM vs. other methods?
   ✓ Because of the many concatenated LSA spaces?
The End
1. Though overall precision rates are comparable, different windows sizes for SL-LSA select different types of similarity, e.g.
   - Small windows: synonymous adverbs
   - Large windows: same domains

✓ Explains the results of SL-dLSA+SVM

2. Questions

✓ Can different window sizes be combined for other problems (disambiguation, alignment)
✓ Can we combine various SL-LSA?