Semantic Similarity for Aspect-Based Sentiment Analysis

P. D. Blinov, E. V. Kotelnikov
Vyatka State Humanities University, Kirov
Outline

1. Text data

2. The methods and results
   - Aspect term extraction
   - Aspect term polarity detection
   - Aspect term classification
   - Aspect category polarity detection
Data

Restaurants

\{Food, Interior, Service, Price, Whole\}

#terms = 2 822, #terms = 3 506

19 034 reviews

Automobiles

\{Comfort, Appearance, Reliability, Safety, Driveability, Costs, Whole\}

#terms = 3 152, #terms = 3 109

8 271 reviews

Explicit term sentiment distribution
Word vector space

\[
\text{xлеб} \rightarrow \begin{pmatrix} 0.1 \\ 0.4 \\ 0.3 \\ 0.1 \\ 0.2 \end{pmatrix}
\]

\[
\text{колесо} \rightarrow \begin{pmatrix} 0.7 \\ 0.2 \\ 0.3 \\ 0.8 \\ 0.9 \end{pmatrix}
\]

\[
\text{стильный} \rightarrow \begin{pmatrix} 0.6 \\ 2.3 \\ 0.5 \\ 1.8 \\ 3.9 \end{pmatrix}
\]

\[
sim(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \cdot \|\vec{b}\|}
\]

\[
sim(\text{хлеб, батон}) = 0.98 \quad sim(\text{хлеб, колесо}) = -0.85
\]
Aspect term extraction

- For each aspect $a$ get etalon seed words $B_a$ from the train collection
  - $B_{Food} = \{\text{меню, кухня, блюдо, еда, ...}\}$, $|B_{Food}|=136$
  - $B_{Whole} = \{\text{ресторан, заведение, место, ...}\}$, $|B_{Whole}|=38$
  - etc.

- For new term $t$ check sum similarity to $B_a$

$$sim(\vec{t}, B_a) = \sum_{i=1}^{k} \frac{\vec{t} \cdot \vec{b}_i}{\|\vec{t}\| \cdot \|\vec{b}_i\|}, \vec{b}_i \in B_a$$
Aspect term extraction

- Find threshold through CV procedure
- Apply a set of rules
  - Merge sequential terms (роллы на гриле → роллы на гриле)
  - Process quotes, parentheses (салат “Цезарь” → салат “Цезарь”)
  - etc.
## Aspect term extraction. Results

<table>
<thead>
<tr>
<th>run_id</th>
<th>Restaurants baseline</th>
<th>F1, exact matching (macro)</th>
<th>60.84</th>
<th>66.51</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Restaurants 2_1</td>
<td>F1, partial matching (macro)</td>
<td>63.19</td>
<td>68.91</td>
</tr>
<tr>
<td></td>
<td>Restaurants 4_1</td>
<td></td>
<td>60.70</td>
<td>72.84</td>
</tr>
<tr>
<td></td>
<td>Restaurants 4 of 14</td>
<td></td>
<td>4 of 14</td>
<td>1 of 14</td>
</tr>
<tr>
<td></td>
<td>Automobiles baseline</td>
<td>F1, exact matching (macro)</td>
<td>59.41</td>
<td>69.66</td>
</tr>
<tr>
<td></td>
<td>Automobiles 2_1</td>
<td>F1, partial matching (macro)</td>
<td>67.61</td>
<td>73.04</td>
</tr>
<tr>
<td></td>
<td>Automobiles 3_1</td>
<td></td>
<td>65.13</td>
<td>74.82</td>
</tr>
<tr>
<td></td>
<td>Automobiles 4_1</td>
<td></td>
<td>58.63</td>
<td>70.16</td>
</tr>
<tr>
<td></td>
<td>Automobiles 7 of 10</td>
<td></td>
<td>4 of 14</td>
<td>1 of 14</td>
</tr>
</tbody>
</table>
Aspect term extraction. Results

Poor on automobile domain

• Lack of data
  – Only 8 271 reviews
    (but 19 034 for the restaurants)

• Specific terms
  – двигатель 2.5 литра; Мотор 1700 DTI;
    m30b30 двигатель; Двигатель 406
  – ваз 2114; ВАЗ-2109; bmw 528i
Aspect term polarity detection

1. Sentiment lexicon creation
   - Sentiment candidates selection
     - All verbs, adjectives and not + <verb | adjective>
     - e.g. оживлять, прекрасный, не пафосный, etc.
   - Candidate weighting
     - Based on PMI
       \[
       \text{score}(t) = \log_2 \frac{\text{count}(t, \text{pos}) \cdot N}{\text{count}(t) \cdot \text{count}(\text{pos})} - \log_2 \frac{\text{count}(t, \text{neg}) \cdot N}{\text{count}(t) \cdot \text{count}(\text{neg})}
       \]
     - \(N\) – total number of tokens in corpus
     - Based on semantic similarity
       \[
       \text{sim}(\vec{t}, B_+) = \sum_{i=1}^{k} \frac{\vec{t} \cdot \vec{b}_i}{||\vec{t}|| \cdot ||\vec{b}_i||}
       \]
     - \(B_+ = \{\text{хороший, уютный, милый, ...}\}, |B_+| = 20\)
2. Aspect term classification

- 3 class: *Positive*, *Negative*, *Both*
- Terms from left and right contexts as features
- Gradient Boosting Classifier + Rule to handle “Both” polarity class
Aspect term polarity detection. Results

<table>
<thead>
<tr>
<th>run_id</th>
<th>Restaurants</th>
<th>F1, micro-averaging</th>
<th>F1, macro-averaging</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>Restaurants</td>
<td>71.04</td>
<td>26.71</td>
</tr>
<tr>
<td>4_1</td>
<td>Restaurants</td>
<td>82.49</td>
<td>55.45</td>
</tr>
<tr>
<td>3_1</td>
<td>Restaurants</td>
<td>66.96</td>
<td>26.96</td>
</tr>
<tr>
<td></td>
<td>Restaurants</td>
<td></td>
<td></td>
</tr>
<tr>
<td>baseline</td>
<td>Restaurants</td>
<td>61.92</td>
<td>26.48</td>
</tr>
<tr>
<td>4_1</td>
<td>Automobiles</td>
<td>74.28</td>
<td>56.84</td>
</tr>
<tr>
<td>1_2</td>
<td>Automobiles</td>
<td>65.31</td>
<td>34.22</td>
</tr>
<tr>
<td></td>
<td>Automobiles</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Aspect term classification

• Average vectors for multi-word terms
• Compute sum similarity for etalon sets of words $B_a$ for each aspect category $a$
• Assign term to category that maximize the similarity

\[
\text{молочный} \quad \text{коктейль} \quad \text{молочный_коктейль} \\
\begin{pmatrix}
0.2 \\
0.5 \\
0.3 \\
0.1 \\
0.2
\end{pmatrix}
\begin{pmatrix}
0.3 \\
0.4 \\
0.2 \\
0.7 \\
0.1
\end{pmatrix}
\begin{pmatrix}
0.2 \\
0.4 \\
0.2 \\
0.4 \\
0.1
\end{pmatrix}
\text{avg}
\]

\[
\text{sim}_{\text{food}} \quad B_{\text{food}} \\
\text{sim}_{\text{whole}} \quad B_{\text{whole}} \\
\arg \max_a (\text{sim}_a)
\]
# Aspect term classification. Results

<table>
<thead>
<tr>
<th></th>
<th>run_id</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Restaurants</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>baseline</td>
<td></td>
<td>79.96</td>
</tr>
<tr>
<td>8_1</td>
<td></td>
<td>86.53</td>
</tr>
<tr>
<td>4_1</td>
<td></td>
<td>81.10</td>
</tr>
<tr>
<td><strong>Automobiles</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>baseline</td>
<td></td>
<td>56.36</td>
</tr>
<tr>
<td>8_1</td>
<td></td>
<td>65.21</td>
</tr>
<tr>
<td>4_1</td>
<td></td>
<td>60.77</td>
</tr>
</tbody>
</table>

2 of 5
Aspect category polarity detection

• Use found terms and their polarity to calculate overall aspect sentiments
  – Positive $\rightarrow$ +1
  – Negative $\rightarrow$ –1

• Sum over terms
  – If sum $> 0$ then aspect is Positive
  – If sum $< 0$ then aspect is Negative
  – If there is at least one term with Both
  then aspect is marked as Both
## Aspect category polarity detection. Results

<table>
<thead>
<tr>
<th>run_id</th>
<th>Restaurants</th>
<th>F1</th>
<th>Automobiles</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Baseline</td>
<td>27.20</td>
<td>baseline</td>
<td>23.68</td>
</tr>
<tr>
<td></td>
<td>4_1</td>
<td>45.82</td>
<td>4_1</td>
<td>43.90</td>
</tr>
<tr>
<td></td>
<td>10_1</td>
<td>37.28</td>
<td>1 of 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1 of 5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Thank you!