AUTOMATIC GENERATION OF THE DOMAIN-SPECIFIC SENTIMENT RUSSIAN DICTIONARIES

Alina Dubatovka / SPbSU
Yuriii Kurochkin / Yandex
Elena Mikhailova / SPbSU

Dialogue 2016, Moscow, June 1-4, 2016
Goals

- Automatic extraction of sentiment words
- Automatic polarity detection
- Unsupervised
Methodology

- Hatzivassilogloum, McKeown 1997
  - "Tasty and healthy Breakfast"
  - "Cheap but nice hotel"

- The better the node is connected with other "positive" nodes and the worse with the "negative", the more positive it is
Graph builder

- (ADV |NEG) * ADJ(, ? (AND |BUT)? (ADV |NEG) * ADJ) +
- AND – conjunction "and"
- BUT – one of adversative conjunctions ("but", "instead", "however", “nevertheless ”)
- NEG – negation
- ADV – an adverb of measure and degree ("very", "quite", "too", "completely")
- ADJ – adjective
Example

• "Tasty, plentiful but not very varied and expensive breakfast"

• positive links: (tasty, plentiful), (tasty, varied), (plentiful and varied)

• negative links: (tasty, expensive), (plentiful, expensive), (varied, expensive).
Particle “not” and prefix “un-”

AUTOMATIC GENERATION OF THE DOMAIN-SPECIFIC SENTIMENT RUSSIAN DICTIONARIES
Graph Analyzer

• Initialization
• Weight of the graph edges
  – \( weight(word_1, word_2) = \#(word_1 \text{AND} word_2) - K \times \#(word_1 \text{BUT} \text{word}_2) \)
• Distance to the final set
  – The heaviest edge
  – The sum of the weights of edges
Description of experiments

• 259023 depersonalized unlabeled reviews
• Dataset size – 660 Mb
• Hotel domain
• Texts by real users
  – Misspellings
  – Grammatical errors
  – Informal words
  – unrelated information concerning flight, excursions, places of interest *etc*
“Large” dictionaries

<table>
<thead>
<tr>
<th>Algorithm without removing the &quot;un-&quot; prefix</th>
<th>Positive</th>
<th>Negative</th>
<th>Neutral</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm after removing the &quot;un-&quot; prefix</td>
<td>5252</td>
<td>2815</td>
<td>-</td>
<td>8067</td>
</tr>
<tr>
<td>“Large” dictionary</td>
<td>1948</td>
<td>1946</td>
<td>4951</td>
<td>8845</td>
</tr>
</tbody>
</table>
AUTOMATIC GENERATION OF THE DOMAIN-SPECIFIC SENTIMENT RUSSIAN DICTIONARIES

“Small” dictionaries

<table>
<thead>
<tr>
<th></th>
<th>Positive dictionary</th>
<th>Negative dictionary</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Manual” dictionary</td>
<td>173</td>
<td>127</td>
<td>300</td>
</tr>
<tr>
<td>Algorithm without “un-” prefix removing</td>
<td>164</td>
<td>74</td>
<td>238</td>
</tr>
<tr>
<td>Algorithm with “un-” prefix removing</td>
<td>163</td>
<td>83</td>
<td>246</td>
</tr>
</tbody>
</table>
Results without removing the "un-" prefix

<table>
<thead>
<tr>
<th>Metric</th>
<th>Positive dictionary</th>
<th>Negative dictionary</th>
<th>Total dictionary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>0.806</td>
<td>0.684</td>
<td>0.754</td>
</tr>
<tr>
<td>Precision</td>
<td>0.309</td>
<td>0.521</td>
<td>0.381</td>
</tr>
<tr>
<td>Precision without neutral words</td>
<td>0.77</td>
<td>0.827</td>
<td>0.796</td>
</tr>
<tr>
<td>$F_1$-measure</td>
<td>0.447</td>
<td>0.591</td>
<td>0.506</td>
</tr>
<tr>
<td>$F_1$-measure without neutral words</td>
<td>0.788</td>
<td>0.749</td>
<td>0.774</td>
</tr>
</tbody>
</table>
Results after removing the "un-" prefix

<table>
<thead>
<tr>
<th>Metric</th>
<th>Positive dictionary</th>
<th>Negative dictionary</th>
<th>Total dictionary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>0.793</td>
<td>0.683</td>
<td>0.746</td>
</tr>
<tr>
<td>Precision</td>
<td>0.314</td>
<td>0.502</td>
<td>0.38</td>
</tr>
<tr>
<td>Precision without neutral words</td>
<td>0.779</td>
<td>0.82</td>
<td>0.799</td>
</tr>
<tr>
<td>$F_1$-measure</td>
<td>0.45</td>
<td>0.579</td>
<td>0.504</td>
</tr>
<tr>
<td>$F_1$-measure without neutral words</td>
<td>0.786</td>
<td>0.745</td>
<td>0.772</td>
</tr>
</tbody>
</table>
Precision@n for positive dictionary
Precision@n for negative dictionary
Dependence on K

Without neutral words

With neutral words

Without neutral words

With neutral words