Three attempts in PolEval 2017 Sentiment Analysis Task

Michal Ptaszynski, Fumito Masui, Arkadiusz Janz, Jan Kocoń, Maciej Piasecki
Kitami Institute of Technology Wroclaw University of Science and Technology

Monika Zaśko-Zielińska, Paweł Dybała
University of Wrocław Jagiellonian University
Outline

• PolEval 2017 Sentiment Analysis Task

• Three attempts
  – MuNNCo
  – NoCoLeS
  – JIBBoF

• Conclusions
PolEval 2017 Sentiment Analysis Task

• Polish: Underresourced language in NLP
• Multiple initiatives
  – NJKP: National Corpus of Polish
  – Słowosieć: Polish WordNet
  – PolNet: another Polish WordNet
  – Składnica: Treebank of Polish
  – PolEval: Series of NLP Tasks and Data
    http://poleval.pl/
PolEval 2017 Sentiment Analysis Task

• PolEval Tasks:
  – POS Tagging
  – Sentiment Analysis
PolEval 2017 Sentiment Analysis Task

Training Sets:
• 235 Sentences from the Skladnica manually annotated for sentiment and dependency
• 965 Reviews of perfumes and clothes manually annotated for sentiment and automatically annotated dependency
• Test Set:
  350 sentences (reviews)
PolEval 2017 Sentiment Analysis Task

<table>
<thead>
<tr>
<th>Tokens</th>
<th>Według mnie to zapach godny polecenia .</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS</td>
<td>adjunct comp pred subj adjunct adjunct punct</td>
</tr>
<tr>
<td>Depend.</td>
<td>3 1 0 3 6 4 3</td>
</tr>
<tr>
<td>Sentiment (English)</td>
<td>According_to me this_is_a smell worth recommending .</td>
</tr>
</tbody>
</table>
PolEval 2017 Sentiment Analysis Task

<table>
<thead>
<tr>
<th>Tokens</th>
<th>Według mnie to zapach godny polecenia .</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS</td>
<td>adjunct comp pred subj adjunct adjunct punct</td>
</tr>
<tr>
<td>Depend.</td>
<td>3 1 0 3 6 4 3</td>
</tr>
<tr>
<td>Sentiment</td>
<td>0 0 0 1 1 1 0</td>
</tr>
<tr>
<td>(English)</td>
<td>According to me this is a smell worth recommending</td>
</tr>
</tbody>
</table>

- Task: predict **sentiment** of each word/phrase using **other available information**
PolEval 2017 Sentiment Analysis Task
PolEval 2017 Sentiment Analysis Task

- test
- rev+sklad
- sklad
- rev

0% 20% 40% 60% 80% 100%

positive    neutral    negative
Three attempts

• Multiclass Nearest Neighbours Condominium
• Normalized Contextual Levenshtein Similarity
• Jaccard Index on Big Bag of Features
Multiclass Nearest Neighbours
Condominium
### Multiclass Nearest Neighbours

**Condominium**

<table>
<thead>
<tr>
<th>Tokens</th>
<th>Według mnie to zapach godny polecenia .</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS Depend.</td>
<td>adjunct comp pred subj adjunct adjunct punct</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Sentiment (English)</td>
<td>According to me this is a smell worth recommending .</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

8th Language & Technology Conference
### Multiclass Nearest Neighbours

**Condominium**

<table>
<thead>
<tr>
<th>Tokens</th>
<th>Według</th>
<th>mnie</th>
<th>to</th>
<th>zapach</th>
<th>godny</th>
<th>polecenia</th>
<th>.</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS</td>
<td>adjunct</td>
<td>comp</td>
<td>pred</td>
<td>subj</td>
<td>adjunct</td>
<td>adjunct</td>
<td>punct</td>
</tr>
<tr>
<td>Depend.</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>6</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Sentiment</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>(English)</td>
<td>According to me, this is a smell worth recommending.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Class 1

Class 2

Class 3

Class 4

...
Multiclass Nearest Neighbours
Condominium

<table>
<thead>
<tr>
<th>Tokens</th>
<th>POS</th>
<th>Według mnie to</th>
<th>zapach subj</th>
<th>godny polecenia</th>
<th>.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depend.</td>
<td>Sentiment (English)</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>
|        | According to me this is a smell worth recommending.

Class: **zapach_subj_3_1**

Samples:

- Według_adjunct_3 mnie_comp_1 to_pred_0 **zapach_subj_3** godny_adjunct_6 polecenia_adjunct_4 . punct_3
- Ten_adjunct_2 **zapach_subj_3** uwiódł_conjunct_5 mnie_obj_3 i_coord_0 uzależnił_conjunct_5 !_punct_5
- ...
Multiclass Nearest Neighbours
Condominium

<table>
<thead>
<tr>
<th>Tokens POS</th>
<th>według mniej to zapach godny polecenia .</th>
</tr>
</thead>
<tbody>
<tr>
<td>adjunct comp pred subj adjunct adjunct punct</td>
<td>3 6 4 3</td>
</tr>
<tr>
<td>3 1 0 3 6 4 1 1 0</td>
<td></td>
</tr>
</tbody>
</table>

- “block of flats” for 7729 unique classes
- k-Nearest Neighbour (k=1)
  - Computationally not demanding
- From predicted look only at sentiment
Multiclass Nearest Neighbours
Condominium

• Results:
  – All predictions: neutral (Accuracy = 72.64%)
  – Reason: most labels in training data were neutral
  – Not enough data for effective machine learning
Normalized Contextual Levenshtein Similarity
Normalized Contextual Levenshtein Similarity

<table>
<thead>
<tr>
<th>Tokens</th>
<th>Według_mnie to zapach_adjunct subj</th>
<th>godny_polecenia punt</th>
<th>POS</th>
<th>Według_mnie to zapach_adjunct subj godny_polecenia punct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjunct</td>
<td>comp pred</td>
<td>subj</td>
<td>Adjunct</td>
<td>Adjunct</td>
</tr>
<tr>
<td>Depend.</td>
<td>3 1 0</td>
<td>3 6 4</td>
<td>3</td>
<td>1 1 1</td>
</tr>
<tr>
<td>0 0 0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentiment</td>
<td>According_to_me this_is_a smell worth recommending</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Dictionary entry: zapach_subj_3_1

Sample sentences:
- Według_adjunct_3 mnie_comp_1 to_pred_0 zapach_subj_3 godny_adjunct_6 polecenia_adjunct_4 ._punct_3
- Ten_adjunct_2 zapach_subj_3 uwiodł_conjunct_5 mnie_obj_3 i_coord_0 uzależnił_conjunct_5 !_punct_5
- ...
Normalized Contextual Levenshtein Similarity

Sample sentence for Dictionary entry “zapach_subj_3_1”:
Według_adjunct_3 mnie_comp_1 to_pred_0 zapach_subj_3 godny_adjunct_6 polecenia_adjunct_4 _punct_3

Sentence-normalized Levenshtein Distance

New Input:
Ten_adjunct_2 zapach_subj_3 uwiódł_conjunct_5 mnie_obj_3 i_coord_0 uzależnił_conjunct_5 !_punct_5
Normalized Contextual Levenshtein Similarity

Sample sentence for Dictionary entry “zapach_subj_3_1”:
Według_adjunct_3  mnie_comp_1  to_pred_0  zapach_subj_3  godny_adjunct_6  polecenia_adjunct_4  ._punct_3

Sentence-normalized Levenshtein Distance

New Input:
Ten_adjunct_2  zapach_subj_3  uwiódł_conjunct_5  mnie_obj_3  i_coord_0  uzależnił_conjunct_5  !_punct_5

SnLD=
=1 – (number of operations)/(number of elements in sentence)
Normalized Contextual Levenshtein Similarity

• For each sentence element (word/phrase) from new input take Dictionary entry containing Sample sentence with highest similarity.

• Predicted Sentiment is sentiment of the Dictionary entry.
Normalized Contextual Levenshtein Similarity

• Results
• Accuracy=72.68%

<table>
<thead>
<tr>
<th>class-based result</th>
<th>positive</th>
<th>neutral</th>
<th>negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>7.09%</td>
<td>97.98%</td>
<td>1.10%</td>
</tr>
</tbody>
</table>
Normalized Contextual Levenshtein Similarity

• Results
• Accuracy=72.68%

Possible future improvements:
• Using less information (token, pos, depend.)
• Conceptualization of elements for similarity calculation
  – Letters $\rightarrow$ chunks of information (token, pos, dep.)
• Label weighting (e.g., token $>$ pos $>$ dep.)
Jaccard Index on Big Bag of Features
Jaccard Index on Big Bag of Features

Take sub-trees

Według_adjunct mnie_comp to_pred : polarity
to_pred zapach_subj : polarity
zapach_subj godny_adjunct polecenia_adjunct ._punct : polarity

Jaccard Index

New Input:
Ten_adjunct zapach_subj uwiódł_conjunct mnie_obj i_coord uzależnił_conjunct !_punct : predicted polarity

• If word not present in training data  : Use labels from WordNetPL 3.1 emo
Jaccard Index on a Bag of Features (2)

• Use a tagger for more accuracy:
  – Lemma with PoS as a base for features:
    • Polarity feature:
      – zapach_noun\textsuperscript{(root)}: neutral
        godny\_adj\textsuperscript{(child)}: positive
        polecenie\_noun\textsuperscript{(child)}: neutral
    • Dependency relations as a feature:
      – godny\_adj\textsuperscript{(child)}: adjunct; polecenie\_noun\textsuperscript{(child)}: adjunct
  • Label for a subtree with child as a root:
    – polecenie\_noun\textsuperscript{(child)}: neutral
plWordNet 3.1 emo – a large sense lexicon for sentiment analysis

• For even more accuracy, apply a WSD tool
  – text disambiguation: the way to access plWordNet
  – assign the polarity using plWordNet for disambiguated senses

„Uwielbiam ten piękny zapach”
  • zapach_noun -> zapach.1 : neutral
  • piękny_adj -> piękny.2: positive
  • uwielbiać_verb -> uwielbiać.1: positive
Training and testing

1. Generate bags for training / test set
   - apply a Tagger and WSD tool
   - index: link the bags created for subtrees with words
2. Process
   - test bags compared with training bags (bottom up approach)
   - assign a label of the most similar bag
     - zapach_noun(root): 
       godny_adj(child): positive, 
polecenie_noun(child): neutral
     }, label: positive

     - zapach_noun(root): 
       wart_adj(child): positive, 
polecenie_noun(child): neutral
     }, label: ???

* „godny_adj” maps to „godny.3” and „wart_adj” to „wart.1” - same synset!
Jaccard Index on Big Bag of Features

• Results
• Accuracy = 73.9%

<table>
<thead>
<tr>
<th>class-based result</th>
<th>positive</th>
<th>neutral</th>
<th>negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>34.12%</td>
<td>91.16%</td>
<td>10.11%</td>
</tr>
</tbody>
</table>

• Replacing all polarity labels (plWordNet) gives 73.6%
Conclusions

• PolEval 2017 Sentiment Analysis Task
• Three attempts
  – Multiclass Nearest Neighbours Condominium
    • Not enough data, most labels neutral
  – Normalized Contextual Levenshtein Similarity
    • Many unknown words in test data, could be improved
  – Jaccard Index on Big Bag of Features
    • WordNetPL emo annotations helped the most
Thank you for your attention!

Michal PTASZYNSKI
Department of Computer Science
Kitami Institute of Technology
165 Koen-cho, Kitami, 090-8507, Japan
TEL/FAX: +81-157-26-9327
E-mail: ptaszynski@ieee.org