A High Precision Method for Aspect Extraction in Russian

Speaker: Mayorov Vladimir, Andrianov I., Astrakhantsev N., Avanesov V., Kozlov I., Turdakov D.

Moscow, 2015
Aspect Extraction Tasks

Given a set of user review extract aspect terms

- **2 domains:** Automobile, Restaurants
- **3 aspect types:** Explicit, Implicit, Fact
- **5 aspect marks:** Relevant, Comparison, Previous, Irreal, Irony

- Task A: extract all **relevant explicit** aspect terms
- Task B: extract all **relevant** aspect terms
Method Overview

• Task A: Sequence-labeling task
  – BIO encoding
  – 2 aspect marks
    • Relevant
    • Other (Comparison, Previous, ...)
  – SVM classifier (5 classes: BRel, IRel, BOth, IOth, O)

• Task B: successive Task A solution for each aspect term type
Method Implementation

• Texterra Framework
  – Quick ML based NLP method implementation
  – Tools for text-based feature extraction
  – Useful quality testing tools

• Texterra NLP Tools\(^1\)
  – High-quality NLP methods for
    • Sentence detection
    • Tokenization
    • Morphological analysis
    • Syntactic parsing
    • NERC
    • ...

\(^1\) Texterra NLP Tools API is publicly available at [http://api.ispras.ru](http://api.ispras.ru)
Research & Development Process

• Implement system with basic features
• Implement additional features (iterative)
  – Errors analysis
  – Invent reasonable features
  – Implement it (using Texterra NLP Tools)
  – Run experiments (on Automobile domain)
  – Add to final solution (if positive F1-measure change)

• Process test corpora
Basic (NERC) Features

• Motivation:
  – The aspect extraction task is very similar to NERC
  – The most named entities in reviews are aspect terms

• Features:
  – Classification labels of 2 previous tokens
  – Word-form, part of speech tags in window of size 3
  – ...

• Evaluation results (F1):
  – Exact matching: (0.6303; 0.6443)
  – Partial matching: (0.7666; 0.7762)
Syntactic Features

• Motivation:
  – Aspect terms are syntactic subtrees
  – Aspect terms are connected with evaluative words

• Features:
  – Word-form, PoS tags of parent and children tokens
  – Syntactic tree distance
  – Classification labels of parent and children tokens

• Evaluation results (F1):
  – Exact matching: (0.6623; 0.6737) +0.0307
  – Partial matching: (0.7367; 0.7478) -0.0291
Automatic Term Recognition Features

• Motivation:
  – Aspect terms are domain specific terms

• Features:
  – Is word is part of Wikipedia page name
  – TF*IDF
  – ...

• Evaluation results (F1):
  – Exact matching: (0.6701; 0.6822) +0.0082
  – Partial matching: (0.7380; 0.7553) +0.0044
Topic Modeling Features

• Motivation:
  – Aspect terms words should be grouped in topics

• Features:
  – Distribution of a word

• Evaluation results (F1):
  – Exact matching: (0.6775; 0.6891) +0.0071
  – Partial matching: (0.7467; 0.7573) +0.0054
GloVe GMM Features

• Motivation:
  – There are a lot of synonyms of aspect term words

• Features:
  – Probability to be included in each cluster

• Evaluation results (F1)
  – Exact matching: (0.6773; 0.6885) \(-0.0003\)
  – Partial matching: (0.7493; 0.7604) \(+0.0029\)
## Features Impact

<table>
<thead>
<tr>
<th>Features Set</th>
<th>Exact Matching (F1)</th>
<th>Partial Matching (F1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>(0.6773; 0.6885)</td>
<td>(0.7493; 0.7604)</td>
</tr>
<tr>
<td>all - GloVe</td>
<td>(0.6775; 0.6891) +0.0003</td>
<td>(0.7467; 0.7573) -0.0029</td>
</tr>
<tr>
<td>all - TM</td>
<td>(0.6720; 0.6832) -0.0053</td>
<td>(0.7431; 0.7540) -0.0063</td>
</tr>
<tr>
<td>all - ATR</td>
<td>(0.6713; 0.6826) -0.0060</td>
<td>(0.7452; 0.7565) -0.0040</td>
</tr>
<tr>
<td>all - syntactic</td>
<td>(0.6850; 0.6968) +0.0080</td>
<td>(0.7579; 0.7685) +0.0084</td>
</tr>
<tr>
<td>all - NERC</td>
<td>(0.5682; 0.5810) -0.1083</td>
<td>(0.6611; 0.6747) -0.0870</td>
</tr>
</tbody>
</table>
## Task A Results

<table>
<thead>
<tr>
<th>run_id</th>
<th>Restaurants reviews</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exact matching</td>
<td>Partial matching</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td></td>
</tr>
<tr>
<td>baseline</td>
<td>0.55698</td>
<td>0.69033</td>
<td>0.60838</td>
<td>0.65803</td>
<td>0.69601</td>
<td>0.66511</td>
<td></td>
</tr>
<tr>
<td>2_1</td>
<td>0.72365</td>
<td>0.57380</td>
<td>0.63187</td>
<td>0.80775</td>
<td>0.61654</td>
<td>0.68909</td>
<td></td>
</tr>
<tr>
<td>4_1</td>
<td>0.55059</td>
<td>0.69010</td>
<td>0.60704</td>
<td>0.68857</td>
<td>0.79164</td>
<td>0.72840</td>
<td></td>
</tr>
<tr>
<td>11_1</td>
<td>0.51504</td>
<td>0.73502</td>
<td>0.59935</td>
<td>0.61472</td>
<td>0.74111</td>
<td>0.66270</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>run_id</th>
<th>Automobile reviews</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exact matching</td>
<td>Partial matching</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td></td>
</tr>
<tr>
<td>baseline</td>
<td>0.57467</td>
<td>0.62869</td>
<td>0.59413</td>
<td>0.74488</td>
<td>0.67238</td>
<td>0.69663</td>
<td></td>
</tr>
<tr>
<td>2_1</td>
<td>0.76004</td>
<td>0.62179</td>
<td>0.67611</td>
<td>0.85605</td>
<td>0.65509</td>
<td>0.73036</td>
<td></td>
</tr>
<tr>
<td>1_3</td>
<td>0.66187</td>
<td>0.65603</td>
<td>0.65131</td>
<td>0.79169</td>
<td>0.72715</td>
<td>0.74820</td>
<td></td>
</tr>
</tbody>
</table>
## Task B Results

<table>
<thead>
<tr>
<th>run_id</th>
<th>Restaurants reviews</th>
<th></th>
<th>Automobile reviews</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exact matching</td>
<td>Partial matching</td>
<td>Exact matching</td>
<td>Partial matching</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td>P</td>
</tr>
<tr>
<td>baseline</td>
<td>0.54647</td>
<td>0.64772</td>
<td>0.58720</td>
<td>0.67162</td>
</tr>
<tr>
<td>2_1</td>
<td>0.73359</td>
<td>0.51319</td>
<td>0.59617</td>
<td>0.81449</td>
</tr>
<tr>
<td>1_1</td>
<td>0.48403</td>
<td>0.57876</td>
<td>0.52242</td>
<td>0.66867</td>
</tr>
<tr>
<td>1_2</td>
<td>0.60943</td>
<td>0.60062</td>
<td>0.60012</td>
<td>0.75621</td>
</tr>
<tr>
<td>2_1</td>
<td>0.77010</td>
<td>0.55354</td>
<td>0.63662</td>
<td>0.86617</td>
</tr>
<tr>
<td>1_3</td>
<td>0.65632</td>
<td>0.61642</td>
<td>0.63014</td>
<td>0.81428</td>
</tr>
</tbody>
</table>
Questions