ENTITY BASED SENTIMENT ANALYSIS USING SYNTAX PATTERNS AND CONVOLUTIONAL NEURAL NETWORK

Karpov I. A.
Kozhevnikov M.V.
Kazorin V.I.
Nemov N.R.

Trained models and project code can be found at
http://github.com/lab533/RuSentiEval2016
Lexicon actualization
“выдавать” (“fib”)
представлять что-либо не тем, чем оно является на самом деле (to lie)
dелать донос, предавать (to betray)
передавать в чье-л. распоряжение (provide a loan)

Object matching
“Билайн, которым я пользовался два года, гораздо лучше MTC”
(“Beeline, that I’ve used for two years, is much better than MTS”)

Subjective fact interpretation
“Сбербанк подаст в суд иск по банкротству Мечела”
(“Sberbank will bring a bankruptcy case against Mechel to court”)

*Breaking Sticks and Ambiguities with Adaptive Skip-gram
Methods

Overall system architecture

Input text

Text preprocessing

External resources

Trained WV

Sentiment lexicon

Rule-based approach

Sentiment facts detection

Naive classification

Sentiment

CNN-based approach

Text vectorisation

CNN-classification

Sentiment
**Methods**  
**Text preprocessing**

**URLs cleaning**
> ВТБ, Россельхозбанк, Банк Москвы и Национальный Коммерческий Банк (РНКБ) http:…

**Nontextual data cleaning**
> #iPhone #android Сбербанк сообщил о проведении 11 августа технологических работ  
> #iPad #Samsung
> #США и их #санкции. #Ирония. #Сбербанк России приступил к выпуску банковских карт  
> на базе российской платежной…

**Tokenisation & morphology**
custom parser / mystem, smiles

**Named Entity (NE) recognition**
Wikipedia hyperlink structure
Methods
Text preprocessing

Syntax parsing

Со
сколько
лет
можно
взять
кредит
в
TARGET

PREP
NOUN. GENITIVE. INANIMATIVE. PLURAL. MALE

NOUN. ACCUSATIVE. INANIMATIVE. SINGULAR. MALE

ADV
INFN. PERFECTIVE

TARGET

PREP
NOUN. NOMINATIVE. INANIMATIVE. SINGULAR. MALE
WV_Banks_clear: 120,000 bank tweets
WV_TTK_clear: 120,000 telecom tweets
WV_Twitter: 1,500,000 gathered tweets
WV_news: 4,500,000 news texts
Methods Rule-based approach

Pre-trained dictionary
(2074 positive, 6136 negative)

top 2 most similar WV words from WV_twitter
(5,288 positive, 17,251 negative)

wordforms enrichment (60,288 positive, 189,953 negative)
Methods

Rule-based approach

уродливое

эдание

Сбербанкя

Я

ненавижу

Райффайзен

банк
<table>
<thead>
<tr>
<th>Pattern depth</th>
<th>pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>word → parent → child</td>
</tr>
<tr>
<td>3</td>
<td>word → parent → child → grand parent</td>
</tr>
<tr>
<td>3</td>
<td>word → child → * → child</td>
</tr>
<tr>
<td>4</td>
<td>word → parent → grand parent → great grand parent</td>
</tr>
<tr>
<td>4</td>
<td>word → parent → grand parent → child</td>
</tr>
</tbody>
</table>
CNN input:
substitute all "word + POS" pairs are by unique ids
align all sentences to length 50 (zero padding)
Input consists of 3 parts: linear order, parent patterns, sibling patterns

CNN architecture:
• embedding layer - to turn word ids to word vectors, we used only words, contained in training.
• convolution layer - layer with rectified linear unit (ReLU) activation where convolution patterns are applied as described in table 1;
• maxPooling layer - which is down-sampling convolution layer output;
• dropout layer - with dropout rate was set to 0.25;
• dense layer - with ReLU activation;
• dropout layer - with dropout rate was set to 0.5;
• softmax layer - to form classification output.
## Experiments

### Performance of rule- and CNN-based approaches in different configuration

<table>
<thead>
<tr>
<th>Domain</th>
<th>Approach</th>
<th>Training collection</th>
<th>WV</th>
<th>F₁ positive</th>
<th>F₁ negative</th>
<th>Macro-average F₁</th>
<th>Micro-average F₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banks</td>
<td>Rule-based</td>
<td>-</td>
<td>0.387</td>
<td>0.501</td>
<td>0.443</td>
<td>0.463</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rule-based with domain rules</td>
<td>-</td>
<td>0.394</td>
<td>0.524</td>
<td>0.459</td>
<td>0.482</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CNN</td>
<td>Banks Random</td>
<td>0.425</td>
<td>0.555</td>
<td>0.490</td>
<td>0.523</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Banks News</td>
<td>0.422</td>
<td>0.555</td>
<td>0.489</td>
<td>0.523</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Banks Twitter</td>
<td>0.429</td>
<td>0.552</td>
<td>0.490</td>
<td>0.522</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Banks &amp; TTK Random</td>
<td>0.446</td>
<td>0.618</td>
<td>0.532</td>
<td>0.574</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Banks &amp; TTK News</td>
<td>0.455</td>
<td>0.611</td>
<td>0.533</td>
<td>0.572</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Banks &amp; TTK Twitter</td>
<td>0.456</td>
<td>0.615</td>
<td>0.536</td>
<td>0.574</td>
<td></td>
</tr>
<tr>
<td>Telecom</td>
<td>Rule-based</td>
<td>-</td>
<td>0.280</td>
<td>0.682</td>
<td>0.481</td>
<td>0.569</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rule-based with domain rules</td>
<td>-</td>
<td>0.285</td>
<td>0.695</td>
<td>0.490</td>
<td>0.582</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CNN</td>
<td>TTK Random</td>
<td>0.097</td>
<td>0.556</td>
<td>0.326</td>
<td>0.497</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>TTK News</td>
<td>0.091</td>
<td>0.557</td>
<td>0.324</td>
<td>0.499</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>TTK Twitter</td>
<td>0.091</td>
<td>0.559</td>
<td>0.325</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Banks &amp; TTK Random</td>
<td>0.307</td>
<td>0.738</td>
<td>0.523</td>
<td>0.681</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Banks &amp; TTK News</td>
<td>0.298</td>
<td>0.740</td>
<td>0.519</td>
<td>0.682</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Banks &amp; TTK Twitter</td>
<td>0.313</td>
<td>0.739</td>
<td>0.526</td>
<td>0.682</td>
<td></td>
</tr>
</tbody>
</table>
### Experiments

Performance of rule- and CNN-based approaches in different configuration

<table>
<thead>
<tr>
<th>Domain</th>
<th>Approach</th>
<th>F&lt;sub&gt;1&lt;/sub&gt; positive</th>
<th>F&lt;sub&gt;1&lt;/sub&gt; negative</th>
<th>Macro-average F&lt;sub&gt;1&lt;/sub&gt;</th>
<th>Micro-average F&lt;sub&gt;1&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banks</td>
<td>Rule-based</td>
<td>0.394</td>
<td>0.524</td>
<td>0.459</td>
<td>0.482</td>
</tr>
<tr>
<td></td>
<td>CNN</td>
<td>0.456</td>
<td>0.615</td>
<td>0.536</td>
<td>0.574</td>
</tr>
<tr>
<td></td>
<td>Hybrid</td>
<td>0.457</td>
<td>0.619</td>
<td>0.538</td>
<td>0.577</td>
</tr>
<tr>
<td></td>
<td>SentiRuEval best</td>
<td></td>
<td></td>
<td>0.552</td>
<td></td>
</tr>
<tr>
<td>Telecom</td>
<td>Rule-based</td>
<td>0.285</td>
<td>0.695</td>
<td>0.490</td>
<td>0.582</td>
</tr>
<tr>
<td></td>
<td>CNN</td>
<td>0.313</td>
<td>0.739</td>
<td>0.526</td>
<td>0.682</td>
</tr>
<tr>
<td></td>
<td>Hybrid</td>
<td>0.313</td>
<td>0.74</td>
<td>0.527</td>
<td>0.684</td>
</tr>
<tr>
<td></td>
<td>SentiRuEval best</td>
<td></td>
<td></td>
<td>0.559</td>
<td></td>
</tr>
</tbody>
</table>
Conclusions

Rule-based linguistic method showed average performance result, which makes it useful when training collection is not available.

Few hand-written rules with well-filtered dictionaries can give a little boost to the CNN output, but the system degrades as rules count increases.

CNN show very high quality result that coincides with the best results of the competition, but this approach requires relatively large training collections.

Word2vec can extract deep semantic features between words if training corpora is large enough.