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)

делать донос, предавать (to betray)

передавать в чье-л. распоряжение (provide a loan)

Object matching

“***Билайн**, которым я пользовался два года, гораздо **лучше МТС***”

(“*Beeline, that I’ve used for two years, is much better than MTS*”)

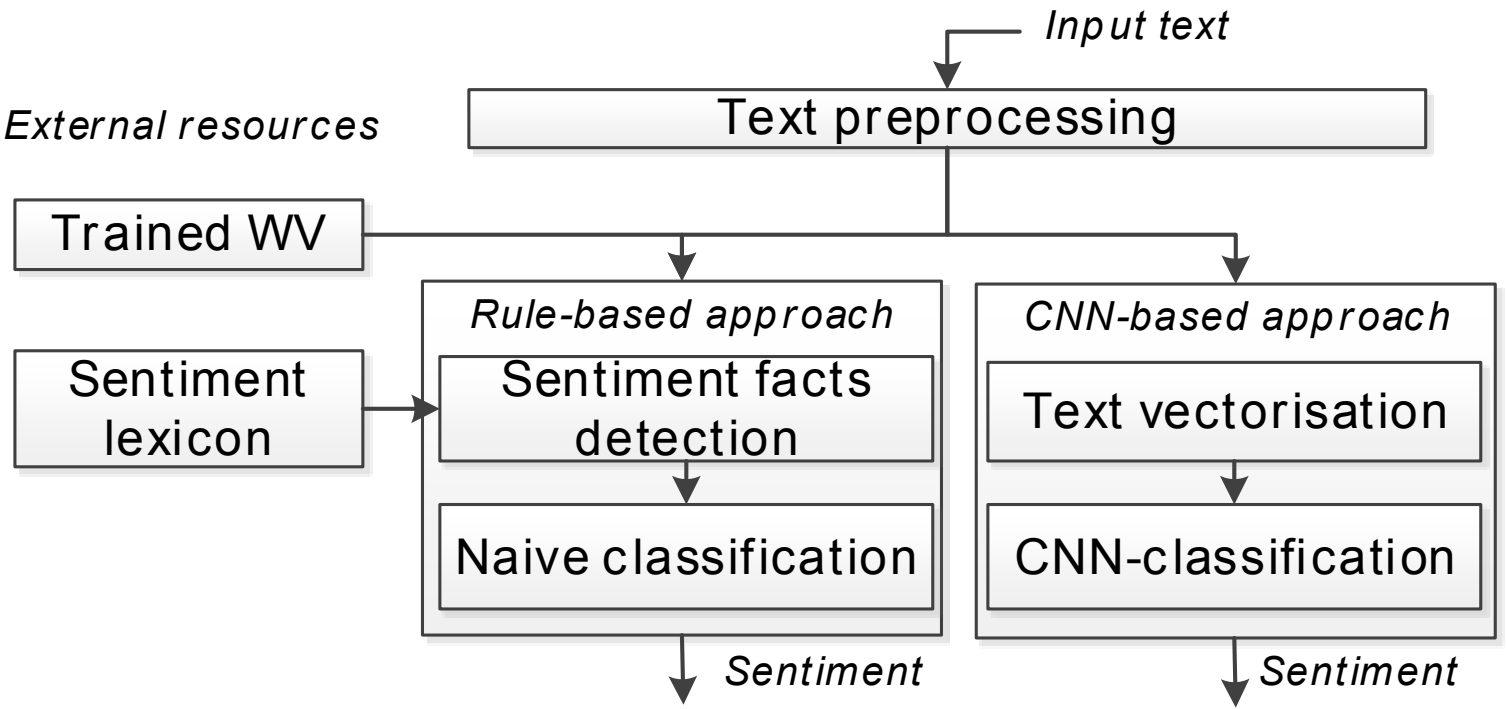
Subjective fact interpretation

“***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
http://jmlr.org/proceedings/papers/v51/bartunov16.pdf

Overall system architecture



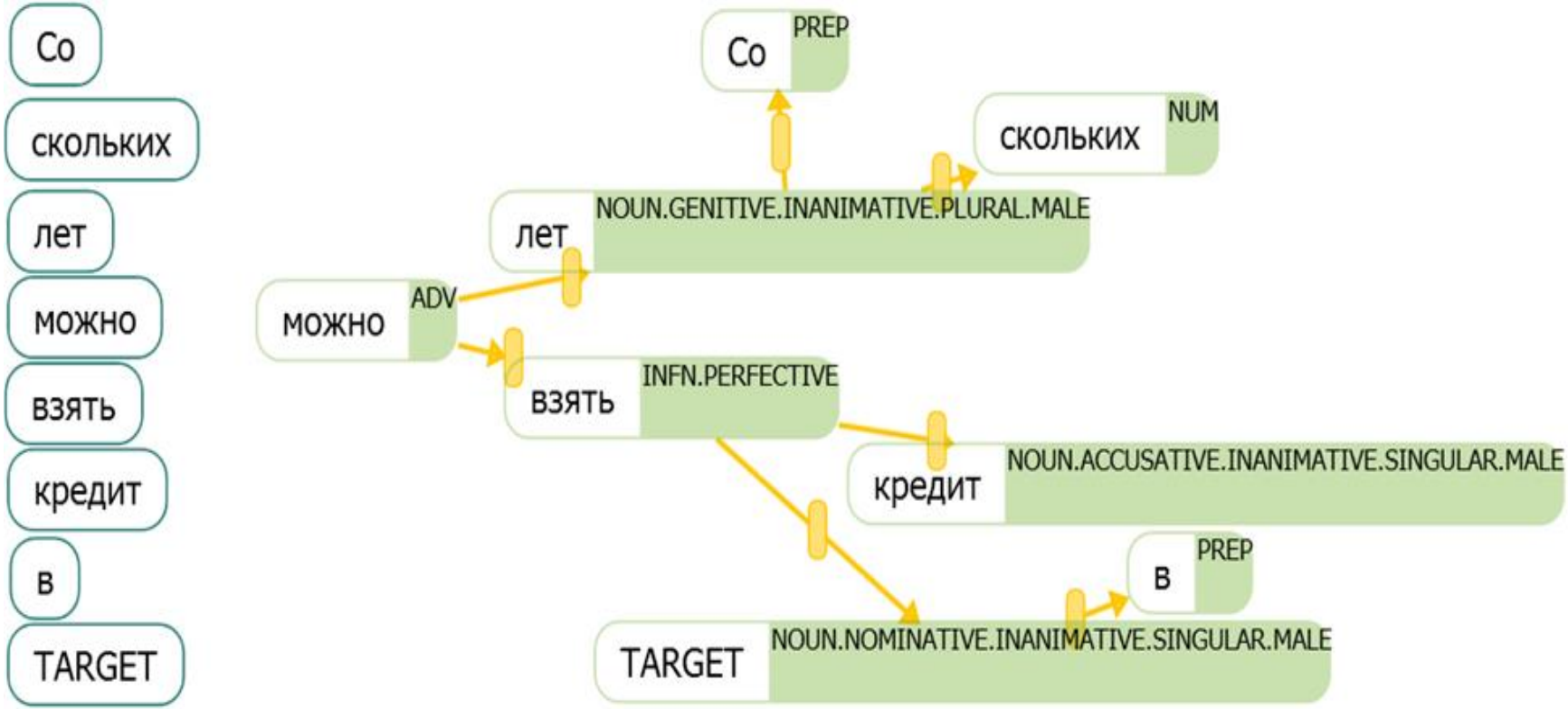
URLs cleaning
> ВТБ,Россельхозбанк,Банк Москвы и Национальный Коммерческий Банк (РНКБ) http:/…

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

Syntax parsing



WV_Banks_clear: 120,000 bank tweets

WV_TTK_ clear: 120,000 telecom tweets

WV_Twitter: 1,500,000 gathered tweets

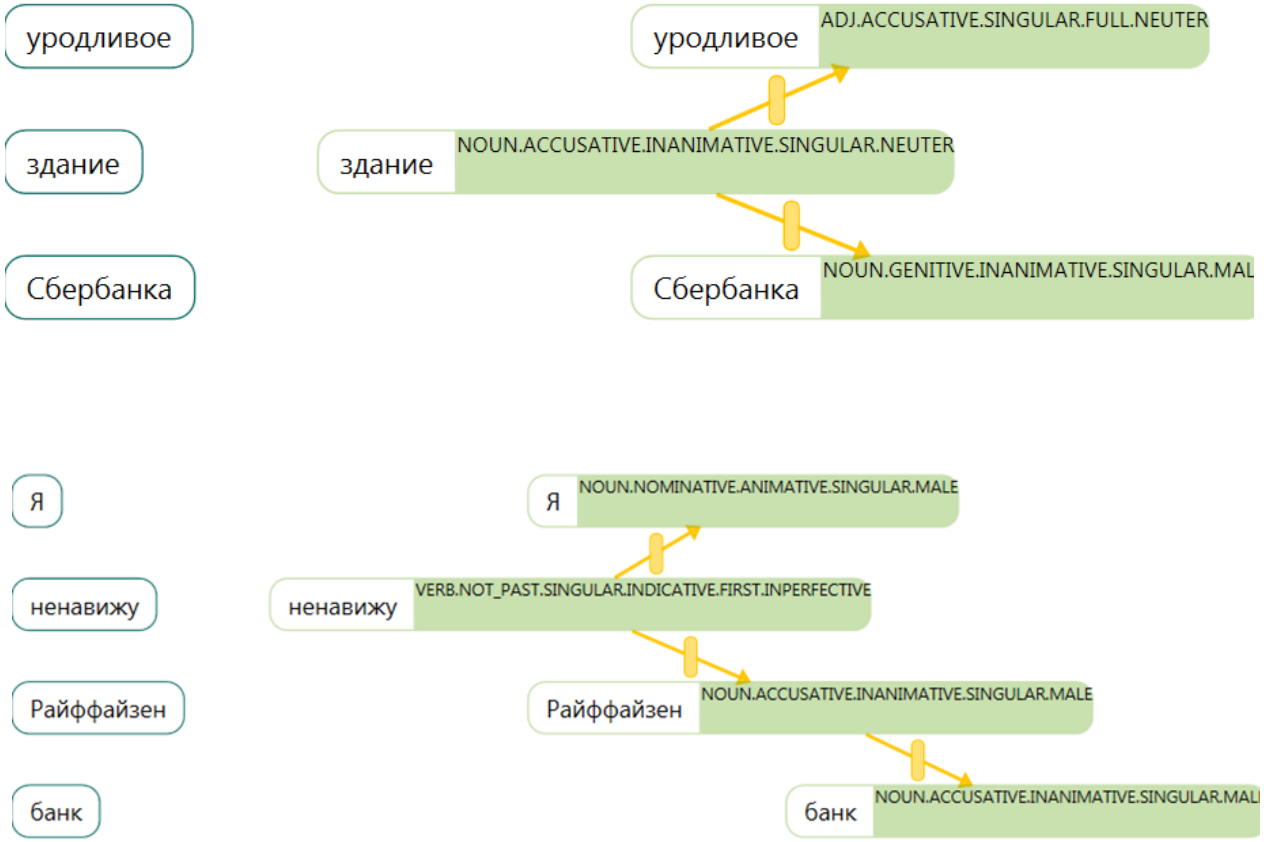
WV_news: 4,500,000 news texts

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



Convolutional neural network approach

Pattern depth	pattern
2	
2	
3	
3	
3	
4	
4	

Convolutional neural network approach

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

Domain	Approach	Training collection	WV	F_1 positive	F_1 negative	Macro-average F_1	Micro-average F_1
Banks	Rule-based	Banks	-	0.387	0.501	0.443	0.463
	Rule-based with domain rules	Banks	-	0.394	0.524	0.459	0.482
	CNN	Banks	Random	0.425	0.555	0.490	0.523
		Banks	News	0.422	0.555	0.489	0.523
		Banks	Twitter	0.429	0.552	0.490	0.522
		Banks & TTK	Random	0.446	0.618	0.532	0.574
		Banks & TTK	News	0.455	0.611	0.533	0.572
		Banks & TTK	Twitter	0.456	0.615	0.536	0.574
Telecom	Rule-based	TTK	-	0.280	0.682	0.481	0.569
	Rule-based with domain rules	TTK	-	0.285	0.695	0.490	0.582
	CNN	TTK	Random	0.097	0.556	0.326	0.497
		TTK	News	0.091	0.557	0.324	0.499
		TTK	Twitter	0.091	0.559	0.325	0.5
		Banks & TTK	Random	0.307	0.738	0.523	0.681
		Banks & TTK	News	0.298	0.740	0.519	0.682
		Banks & TTK	Twitter	0.313	0.739	0.526	0.682

Experiments

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

Domain	Approach	F_1 positive	F_1 negative	Macro-average F_1	Micro-average F_1
Banks	Rule-based	0.394	0.524	0.459	0.482
	CNN	0.456	0.615	0.536	0.574
	Hybrid	0.457	0.619	0.538	0,577
	SentiRuEval best			0.552	
Telecom	Rule-based	0.285	0.695	0.490	0.582
	CNN	0.313	0.739	0.526	0.682
	Hybrid	0.313	0.74	0.527	0.684
	SentiRuEval best			0.559	

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.