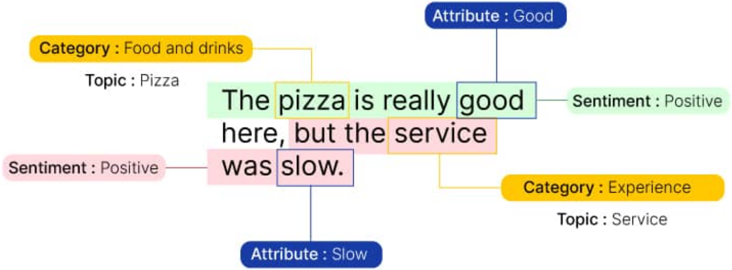
Advances in Sentiment Attitude Extraction from Mass-Media Analytical Texts in Russian

Nicolay Rusnachenko

nicolay-r.github.io

Newcastle University
United Kingdom

Sentiment Analysis



Text classification

The first attempt to propose the task[1]:

$$\langle d \rangle \rightarrow c$$

d – document
c – related class positive, negative

"The picture quality of this camera at night time is amazing"

$$\langle d \rangle \rightarrow \textit{positive}$$

[1] Peter Turney. «Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews». B: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. 2002, c. 417—424.

Targeted sentiment analysis

Considering entity as an input parameter[2]:

$$\langle d, e_j \rangle \rightarrow c$$

e_j – object, or entity

"The picture quality of this camera$_e$
at night time is amazing, especially with tripod$_e$"

$$\langle d, camera \rangle \rightarrow positive \quad \langle d, tripod \rangle \rightarrow ?$$

[2] Long Jiang и др. «Target-dependent twitter sentiment classification». В: *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*. 2011, с. 151—160.

Aspect Based Sentiment Analysis

Focusing on two core tasks[3]:

1. Aspect extraction;
2. Aspect sentiment analysis:

$$\langle d, e_j, a_k \rangle \to c$$

a_k – aspect, object characteristics

"The picture quality of this camera$_e$ is amazing ..."[3]

$$\langle d, camera, picture\ quality \rangle \to positive$$

[3] Bing Liu и Lei Zhang. «A survey of opinion mining and sentiment analysis». B: *Mining text data*. Springer, 2012, c. 415—463.

Attitude Definition

Opinions between mentioned named entities (e_j, e_m):

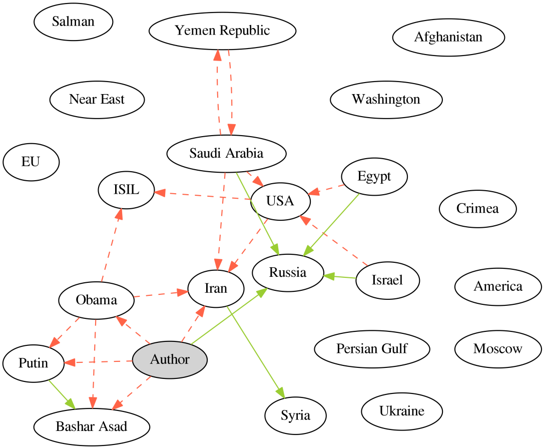
$$\langle d, e_j, e_m, a_k, h_t, t_l \rangle \rightarrow c$$

a_k – aspect
e_m – subject
e_j – object
h_t – author
t_l – time
c – sentiment class ($\textsc{pos}$, $\textsc{neg}$)

" ... Moscow$_e$ dissatisfied with the Warsaw's$_e$ decision ... "

$$\langle e_m, e_j \rangle \rightarrow \textsc{neg}$$

Document-Level Attitude Representation



Sentiment Attitude Extraction Task

Input:

1. Collection of analytical articles $\langle D_i, E_i \rangle$ (in Russian)
   - Each article includes: document D_i, list of mentioned named entites E_i
2. For synonymous mentions: given a collection of synonyms:

   Russia_e , RF_e , $\text{Russian Federation}_e$

Task: For each D_i complete the list of sentiment attitudes (pairs $\langle e_i, e_j, l_{i,j} \rangle$)[4], with label $l_{i,j} \in \{\text{POS}, \text{NEG}\}$

[4] Natalia Loukachevitch и Nicolay Rusnachenko. «Extracting sentiment attitudes from analytical texts». В: *Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialogue-2018 (arXiv:1808.08932)* (2018), с. 459—468.

Task aspects and problems

1. Large amount of named entities (*NE*);
2. Text structure complexities:

« Trump_e accused China_e and Russia_e of "playing devaluation of currencies"»

$(\text{Trump}_{\text{subj}}, \text{China}_{\text{obj}}) \rightarrow$ negative

$(\text{Trump}_{\text{subj}}, \text{Russia}_{\text{obj}}) \rightarrow$ negative

Approach

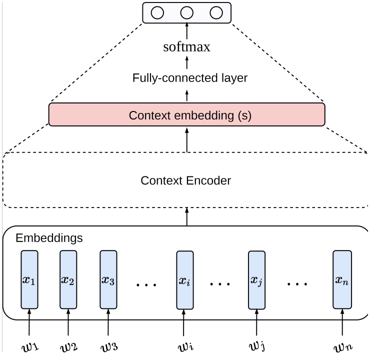
- Contexts as the main idea
- The criterion for the presence of a relationship: a relatively short distance between entities in the text, i.e. in **context**
- **Annotated context** – context with labeled subject-object pair in it $\langle e_i, e_j \rangle$
- Retrieval of attitudes – $\textsc{pos}$ and $\textsc{neg}$ labeling among a set *neutrally labeled* contexts

Automatic Annotation Approaches

1. CNN (including the one with attention mechanism):
   - CNN, PCNN
   - AttPCNN$_e$,
2. BERT-based language models[5]:
   - SentRuBERT

[5] Nicolay Rusnachenko. «Language Models Application in Sentiment Attitude Extraction Task». В: *Proceedings of the Institute for System Programming of the RAS (Proceedings of ISP RAS), vol.33*. 3. 2021, с. 199—222.

Neural Networks

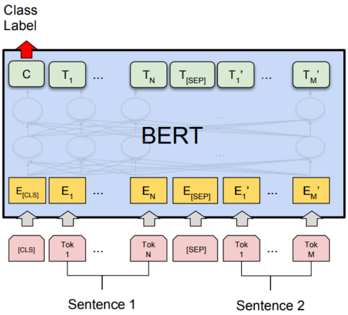


Embedding features for words in context:

1. Term vector from the pretrained model Word2Vec;
2. **Distance** vector from a given term to every pair participant $\langle subj, obj \rangle$ in separate;
3. **Part-of-speech** vectorised representation;
4. Term presence in the side lexicon.

Context labeling: application of FC-layer (fully connected layer) towards the *vectorised context* s: $o = W \cdot s + b$

Language Models



Input sequences:

- TextA: Input context terms
- TextB (Optional), **as prompt**:

  $\underline{E}_{subj}$ towards $\underline{E}_{obj}$ in « $\underline{E}_{subj}$ … $\underline{E}_{obj}$ » is NEG

Context labeling: FC-layer application towards the averaged embedded vectors

Input Formatting Details

Input data
There is no need to talk about Caucasus region$_e$ separation due to the confrontation between Russia$_{subj}$ and Turkey$_{obj}$, although there is a danger.

↓

For Conventional Neural Networks
There is ***not-need***$_{neg}$ to talk about E separation due to the ***confrontation***$_{neg}$ $\underline{E}_{subj}$ and $\underline{E}_{obj}$ `COMMA` although there is a danger `DOT`

↓

For Language Models
TEXTA: there is no need to talk about E separation due to the confrontation $\underline{E}_{subj}$ and $\underline{E}_{obj}$, although there is a danger.
TEXTB: $\underline{E}_{subj}$ towards $\underline{E}_{obj}$ in « $\underline{E}_{subj}$ and $\underline{E}_{obj}$ » is NEG

Datasets

RuSentRel[1]: articles about Russia's international relations

Documents	**73**
Sentences per document	105.8
Entities per document	247
POS and NEG pairs per document	11.47

1 `https://github.com/nicolay-r/RuSentRel/tree/v1.1`

RuSentRel[5] supervised learning results, 3-FOLD CV

Model	$F_1(P, N)$
SentRuBERT	**33.4**
$\text{AttPCNN}_{\text{ends}}$	29.9
PCNN	29.6
Experts agreement	55.0

For `MPQA-3.0`, $F_1 = 36.0$[6]

[6] Eunsol Choi и др. «Document-level sentiment inference with social, faction, and discourse context». В: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2016, с. 333—343.

Distant Supervision Application I

1. **News collecion**: Russian articles from mass-media sources (**8.8M**);
2. Knowledge Base **RuSentiFrames**[2]: describes sentiment association, conveyed by *predicate* in a form of a verb on noun (311 frames)
   - **roles**: A0 (agent), A1 (theme);
   - **dimensions**: authors attitude towards the participants mentioned in text; **polarity** – score between participants;

Frame «хвалиться» (bragging)	Description
ENTRIES	bragging, boasting
ROLES	`A0`: those who bragging
	`A1`: the object of bragging
POLARITY	`A0`→`A1`, POS
	`author`→`A0`, NEG

2 `https://github.com/nicolay-r/RuSentiFrames`

Distant Supervision Application II

Main assumption: news title has a simple structure.

... **Subject**$_e$ ... {*frame*$_{A0 \rightarrow A1}$}$_k$ ... **Object**$_e$ ...

Distant supervision performed in two steps[7]:

1. Collect the list A of the *most-sentiment attitudes* (*subject* $\rightarrow$ *object*) from news titles using frame A0→A1 polarity across all news titles
2. Filter news titles and sentences, which contains at least one pair with A0→A1 score as in A

[7] Nicolay Rusnachenko, Natalia Loukachevitch и Elena Tutubalina. «Distant supervision for sentiment attitude extraction». В: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*. 2019, с. 1022—1030.

Frame Title

Title
Tillerson$_e$: USA$_e$ won't remove ***sanctions***$_{neg}$ from Russia$_e$ before the return of ... Crimea$_e$

↓ USA→Russia$_{neg}$, USA→Crimea$_{neg}$

Most sentiment attitudes	
Query	Search results
USA→Russia$_{neg}$	pair found, scores match; POS: 32%, NEG: 68%
USA→Crimea$_{neg}$	pair not found

↓ USA→Russia$_{neg}$

Sentence
Secretary of State **USA**$_e$ Rex Tillerson$_e$, speaking in Brussels$_e$ at a meeting Foreign$_e$ heads of NATO$_e$ affiliates stated that the sanctions from **Russians**$_e$ will only be removed after the return of Crimea$_e$, according to CNN$_e$.

Datasets

RuAttitudes – automatically marked up collection of texts using the Distant Supervision approach over a large amount of mass-media short news per 2017 year.

Documents	**134442**
Attitudes per document	2.26

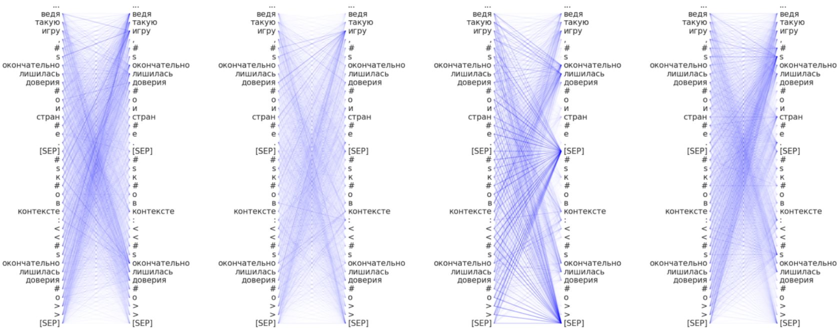
RuSentRel[5] distant-supervision results, 3-FOLD CV

Model	$F_1(P, N)$
SentRuBERT (pretrain + ft)	**39.0**
$\text{AttPCNN}_{\text{ends}}$	32.2
SentRuBERT	**33.4**
$\text{AttPCNN}_{\text{ends}}$	29.9
PCNN	29.6
Experts agreement	55.0

Official RuSentRel leaderboard

SentRuBERT Attention weights analysis

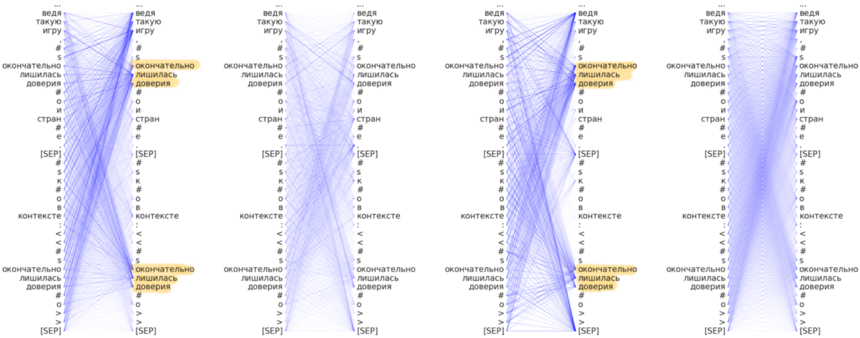
SentRuBERT (head 2, layers from left-to-right: 2, 4, 8, 11)[3]



3 . . . playing such a-game, #S finally lost the-trust of #O and countries #E. [SEP] #S to #O in context: “#S has finally lost the-trust of #O” [SEP]

SentRuBERT Attention weights analysis (II)

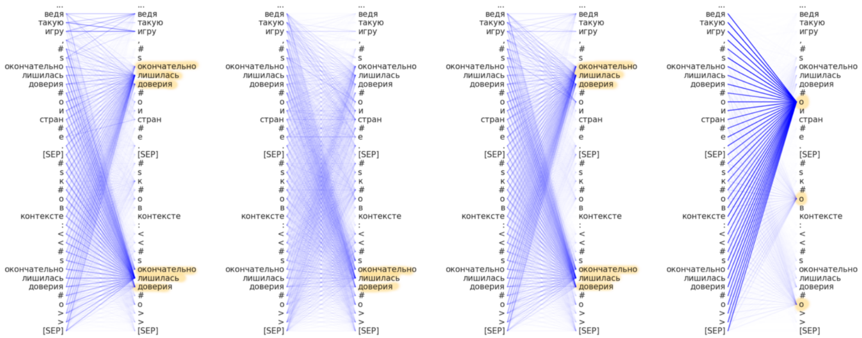
+ 4 epochs on RuAttitudes[4]



4 ... playing such a-game, #S finally lost the-trust of #O and countries #E. [SEP] #S to #O in context: "#S has finally$_{frame}$ lost$_{frame}$ the-trust$_{frame}$ of #O" [SEP]

SentRuBERT Attention weights analysis (III)

[5] + 4 epochs for finetunning on RuSentRel

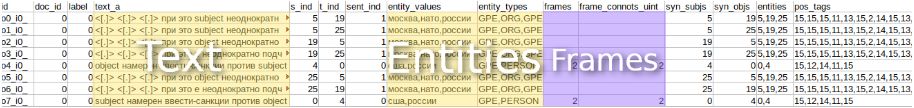


5 ... playing such a-game, #S finally lost the-trust of #O and countries #E. [SEP] #S to #O in context: “#S has finally$_{frame}$ lost$_{frame}$ the-trust$_{frame}$ of #O” [SEP]

Projects

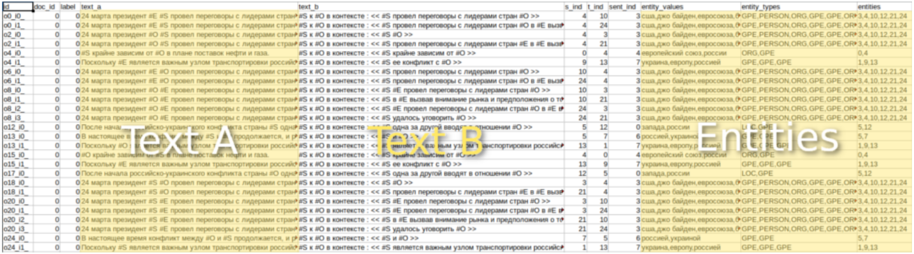
Sampling

In the case of conventional neural networks (frames, features):



id	doc_id	label	text_a	s_ind	t_ind	sent_ind	entity_values	entity_types	frames	frame_connots_uint	syn_subjs	syn_objs	entities	pos_tags
o0_i0_	0	0	<[.]> <[.]> <[.]> при это subject неоднократн	5	19	1	москва,нато,россии	GPE,ORG,GPE			5	19	5,19,25	15,15,15,11,13,15,2,14,15,13,
o1_i0_	0	0	<[.]> <[.]> <[.]> при это subject неоднократн	5	25	1	москва,нато,россии	GPE,ORG,GPE			5	25	5,19,25	15,15,15,11,13,15,2,14,15,13,
o2_i0_	0	0	<[.]> <[.]> [illegible] при это object неоднократно	19	5	1	[illegible]	GPE,ORG,GPE			19	5	5,19,25	15,15,15,11,13,15,2,14,15,13,
o3_i0_	0	0	<[.]> <[.]> [illegible] неоднократно подч	19	25	1	[illegible]	[illegible]			19	25	5,19,25	15,15,15,11,13,15,2,14,15,13,
o4_i0_	0	0	object намер[illegible] против subject	4	0	0	[illegible]	[illegible]	2	2	4	0	0,4	15,12,14,11,15
o5_i0_	0	0	<[.]> <[.]> <[.]> при это object неоднократно	25	5	1	москва,нато,россии	GPE,ORG,GPE			25	5	5,19,25	15,15,15,11,13,15,2,14,15,13,
o6_i0_	0	0	<[.]> <[.]> <[.]> при это е неоднократно подч	25	19	1	москва,нато,россии	GPE,ORG,GPE			25	19	5,19,25	15,15,15,11,13,15,2,14,15,13,
o7_i0_	0	0	subject намерен ввести-санкции против object	0	4	0	сша,россии	GPE,PERSON	2	2	0	4	0,4	15,12,14,11,15

In case of BERT-based language models (TextA, TextB):



id	doc_id	label	text_a	text_b	s_ind	t_ind	sent_ind	entity_values	entity_types	entities
o0_i0_	0	0	24 марта президент #E #S провел переговоры с лидерами стран	#S к #O в контексте : << #S провел переговоры с лидерами стран #O >>	4	10	3	сша,джо байден,евросоюза,(	GPE,PERSON,ORG,GPE,GPE,OR	3,4,10,12,21,24
o0_i1_	0	0	24 марта президент #E #S провел переговоры с лидерами стран	#S к #O в контексте : << #S провел переговоры с лидерами стран #O в #E выз	4	24	3	сша,джо байден,евросоюза,(	GPE,PERSON,ORG,GPE,GPE,OR	3,4,10,12,21,24
o2_i0_	0	0	24 марта президент #O #S провел переговоры с лидерами стран	#S к #O в контексте : << #S #O >>	4	3	3	сша,джо байден,евросоюза,(	GPE,PERSON,ORG,GPE,GPE,OR	3,4,10,12,21,24
o2_i1_	0	0	24 марта президент #O #S провел переговоры с лидерами стран	#S к #O в контексте : << #S провел переговоры с лидерами стран #E в #E выз	4	21	3	сша,джо байден,евросоюза,(	GPE,PERSON,ORG,GPE,GPE,OR	3,4,10,12,21,24
o4_i0_	0	0	#S крайне зависим от #O в плане поставок нефти и газа.	#S к #O в контексте : << #S крайне зависим от #O >>	0	4	4	европейский союз,россии	ORG,GPE	0,4
o4_i1_	0	0	Поскольку #E является важным узлом транспортировки российс	#S к #O в контексте : << #S ее конфликт с #O >>	9	13	7	украина,европу,россией	GPE,GPE,GPE	1,9,13
o6_i0_	0	0	24 марта президент #E #O провел переговоры с лидерами стран	#S к #O в контексте : << #S провел переговоры с лидерами стран #O >>	10	4	3	сша,джо байден,евросоюза,(	GPE,PERSON,ORG,GPE,GPE,OR	3,4,10,12,21,24
o6_i1_	0	0	24 марта президент #E #O провел переговоры с лидерами стран	#S к #O в контексте : << #S провел переговоры с лидерами стран #O в #E выз	24	4	3	сша,джо байден,евросоюза,(	GPE,PERSON,ORG,GPE,GPE,OR	3,4,10,12,21,24
o8_i0_	0	0	24 марта президент #O #E провел переговоры с лидерами стран	#S к #O в контексте : << #S #E провел переговоры с лидерами стран #O >>	10	3	3	сша,джо байден,евросоюза,(	GPE,PERSON,ORG,GPE,GPE,OR	3,4,10,12,21,24
o8_i1_	0	0	24 марта президент #O #E провел переговоры с лидерами стран	#S к #O в контексте : << #S в #E вызвав внимание рынка и предположения о т	10	21	3	сша,джо байден,евросоюза,(	GPE,PERSON,ORG,GPE,GPE,OR	3,4,10,12,21,24
o8_i2_	0	0	24 марта президент #O #E провел переговоры с лидерами стран	#S к #O в контексте : << #S #E провел переговоры с лидерами стран #O в #E в	24	3	3	сша,джо байден,евросоюза,(	GPE,PERSON,ORG,GPE,GPE,OR	3,4,10,12,21,24
o8_i3_	0	0	24 марта президент #O #E провел переговоры с лидерами стран	#S к #O в контексте : << #S удалось уговорить #O >>	24	21	3	сша,джо байден,евросоюза,(	GPE,PERSON,ORG,GPE,GPE,OR	3,4,10,12,21,24
o12_i0_	0	0	После нача[illegible]ийско-украин[illegible]кого конф[illegible]кта страны #S одна	#S к #O в контексте [illegible] одна за другой вводят [illegible]отношении #O >>	5	12	0	запада,россии	[illegible]	5,12
o13_i0_	0	0	В настоящее в[illegible] #S [illegible] продолжается, и р	#S к #O в контексте : <<[illegible] >>	5	7	6	россией,украино[illegible]	[illegible]	5,7
o13_i1_	0	0	Поскольку #O [illegible]ортировки российс	#S к #O в контексте : <<[illegible] важным узлом [illegible]анспортировки российс	13	1	7	украина,европу,р[illegible]	[illegible]	1,9,13
o15_i0_	0	0	#O крайне зав[illegible] #S в плане поставок нефти и газа.	#S к #O в контексте : << #S [illegible] зависим от #O >>	4	0	4	европейский союз,россии	ORG,GPE	0,4
o15_i1_	0	0	Поскольку #E является важным узлом транспортировки российс	#S к #O в контексте : << #S ее конфликт с #O >>	13	9	7	украина,европу,россией	GPE,GPE,GPE	1,9,13
o17_i0_	0	0	После начала российско-украинского конфликта страны #O одна	#S к #O в контексте : << #S одна за другой вводят в отношении #O >>	12	5	0	запада,россии	LOC,GPE	5,12
o18_i0_	0	0	24 марта президент #S #O провел переговоры с лидерами стран	#S к #O в контексте : << #S #O >>	3	4	3	сша,джо байден,евросоюза,(	GPE,PERSON,ORG,GPE,GPE,OR	3,4,10,12,21,24
o18_i1_	0	0	24 марта президент #S #O провел переговоры с лидерами стран	#S к #O в контексте : << #S провел переговоры с лидерами стран #E в #E выз	21	4	3	сша,джо байден,евросоюза,(	GPE,PERSON,ORG,GPE,GPE,OR	3,4,10,12,21,24
o20_i0_	0	0	24 марта президент #S #E провел переговоры с лидерами стран	#S к #O в контексте : << #S #E провел переговоры с лидерами стран #O >>	3	10	3	сша,джо байден,евросоюза,(	GPE,PERSON,ORG,GPE,GPE,OR	3,4,10,12,21,24
o20_i1_	0	0	24 марта президент #S #E провел переговоры с лидерами стран	#S к #O в контексте : << #S #E провел переговоры с лидерами стран #O в #E в	3	24	3	сша,джо байден,евросоюза,(	GPE,PERSON,ORG,GPE,GPE,OR	3,4,10,12,21,24
o20_i2_	0	0	24 марта президент #S #E провел переговоры с лидерами стран	#S к #O в контексте : << #S в #E вызвав внимание рынка и предположения о т	21	10	3	сша,джо байден,евросоюза,(	GPE,PERSON,ORG,GPE,GPE,OR	3,4,10,12,21,24
o20_i3_	0	0	24 марта президент #S #E провел переговоры с лидерами стран	#S к #O в контексте : << #S удалось уговорить #O >>	21	24	3	сша,джо байден,евросоюза,(	GPE,PERSON,ORG,GPE,GPE,OR	3,4,10,12,21,24
o24_i0_	0	0	В настоящее время конфликт между #O и #S продолжается, и р	#S к #O в контексте : << #S и #O >>	7	5	6	россией,украиной	GPE,GPE	5,7
o24_i1_	0	0	Поскольку #S является важным узлом транспортировки российс	#S к #O в контексте : << #S является важным узлом транспортировки российс	1	13	7	украина,европу,россией	GPE,GPE,GPE	1,9,13

AREkit – Text Opinion Sampler

AREkit – Document level **A**ttitude and **R**elation **E**xtraction tool**kit** for sampling mass-media news into datasets for your ML-model training and evaluation

github.com/AREkit

Conclusion

1. Sentiment Attitude Extraction task[6] advances:
   - Supervised Learning
   - Distant Supervision
2. Manually annotated data
   - Relatively high quality
   - Low volume, since consuming a lot of time
3. Distant supervision
   - Solution for an automatic training-data annotation;
   - Expected to be noisy, in terms of correctness; the latter relies on the quality of the knowledge base and its usage.
4. Attention mechanism allows to visually see the most important information considered in the result class decision;

6 http://nlpprogress.com/russian/sentiment-analysis.html

Thank you for attention!

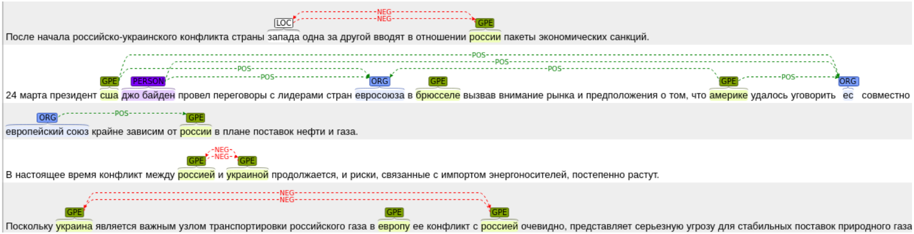
`https://nicolay-r.github.io`

ARElight

ARElight project page

ARElight – Inference Attitudes

The complete example is available here[7].
For SentRuBERT[8]:



7 https://raw.githubusercontent.com/nicolay-r/ARElight/main/data/texts-inosmi-rus/e1.txt
8 http://172.17.0.2/examples/demo/wui_bert.py