

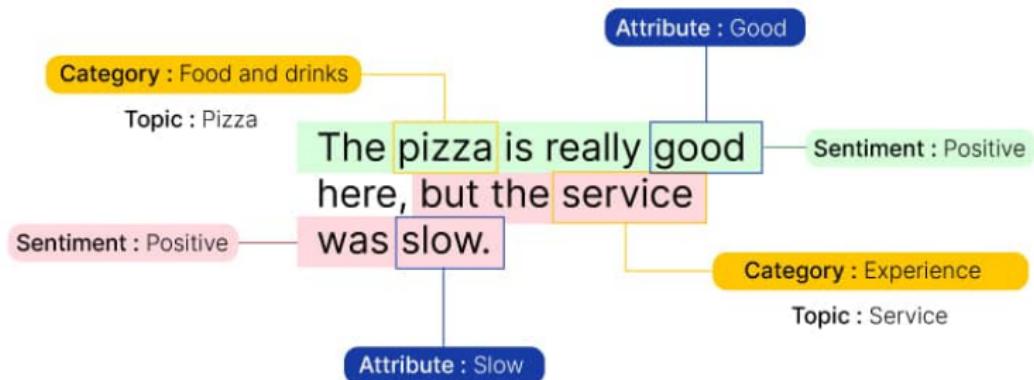
**ARElight** – приложение для обработки и извлечения  
оценочных отношений из больших новостных документов

Nicolay Rusnachenko

[rusnicolay@gmail.com](mailto:rusnicolay@gmail.com)

[nicolay-r.github.io](https://nicolay-r.github.io)

# Sentiment Analysis



## Text classification

Первая попытка предложения постановки задачи<sup>[1]</sup>:

“Качество картинки этой камеры в ночное время – потрясающее”

$$d \rightarrow \text{positive}$$

---

[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, с. 417–424.

## Targeted sentiment analysis

Предполагает указание **сущности<sub>e</sub>** в качестве параметра<sup>[2]</sup>:

“Качество снимков такой **камеры<sub>e</sub>** в  
ночное время просто потрясающее,  
особенно если пользоваться **штативом<sub>e</sub>**”

$\langle d, \text{камера} \rangle \rightarrow \text{positive}$      $\langle d, \text{штатив} \rangle \rightarrow ?$

---

[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

Два основных направления<sup>[3]</sup>:

- ① Извлечение аспекта
- ② Аспектно-ориентированный анализ тональности

добавляется **аспект** (характеристика объекта)

“Качество картинки этой **камеры<sub>e</sub>** – потрясающее . . .”<sup>[3]</sup>

$\langle d, \text{камера}, \text{качество картинки} \rangle \rightarrow \text{positive}$

---

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

## Attitude Definition

Отношения размеченные между именованными сущностями ( $e_j, e_m$ ):

$$\langle d, e_j, \textcolor{red}{e_m}, a_k, h_t, t_l \rangle \rightarrow c$$

$a_k$  – аспект

$e_m$  – субъект

$e_j$  – объект

$h_t$  – автор

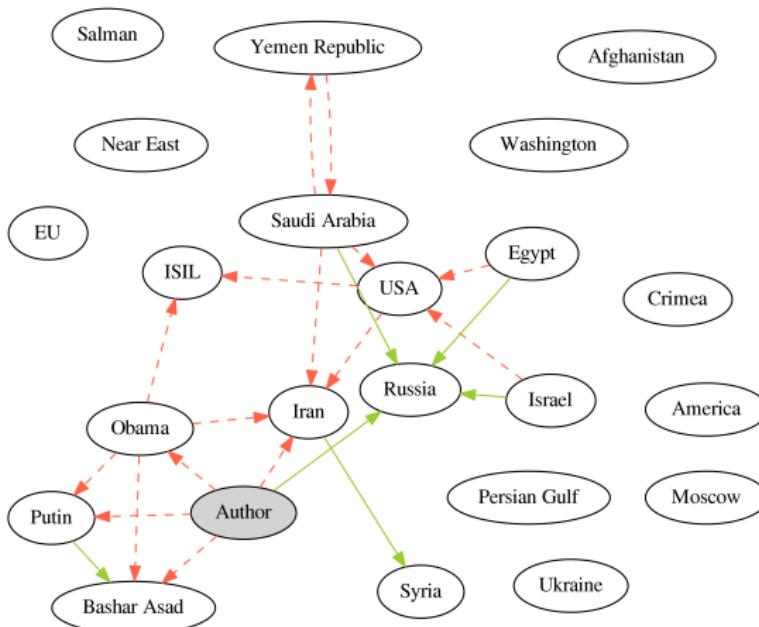
$t_l$  – время

$c$  – класс тональности (POS, NEG)

“ ... Москва<sub>e</sub> недовольна решением Варшавы<sub>e</sub> ... ”

$$\langle e_m, e_j \rangle \rightarrow \text{NEG}$$

# Document-Level Attitude Representation



## Задача извлечения оценочных отношений

- Даны коллекция аналитических статей
- Каждая статья включает: документ  $D_i$ , список упомянутых именованных сущностей  $E_i$
- Для синонимичных упоминаний: вводится коллекция синонимов ( Россия<sub>e</sub>, РФ<sub>e</sub>, Российская Федерация<sub>e</sub> )
- Необходимо для каждого  $D_i$  составить список оценочных отношений (пар  $\langle e_i, e_j \rangle$ )<sup>[4]</sup>, где оценка пары может быть из множества {POS, 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](https://arxiv.org/abs/1808.08932)) (2018), с. 459—468.

## Примеры извлечения оценочных отношений

Пример<sup>1</sup>:



Пример<sup>2</sup>:



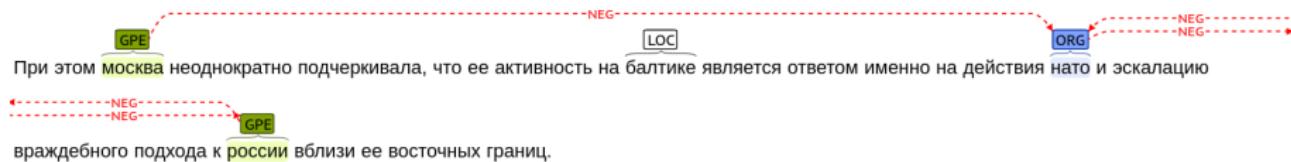
1 Ввиду введения санкций

2 Россия<sub>e</sub> и Китай<sub>e</sub> нейтральны друг к другу

## Примеры извлечения оценочных отношений (II)

Пример:

При этом **москва** неоднократно подчеркивала, что ее активность на балтике является ответом именно на действия **нато** и эскалацию  
враждебного подхода к **россии** вблизи ее восточных границ.



Пример<sup>3</sup>:

Говорить о разделении кавказского региона из-за конфронтации **россии** и **турции** пока не приходится, хотя опасность есть.



3 Ввиду *разделения*, а также *опасности* контекст слева и справа относительно оценочной пары

## Основная идея

- Критерий наличия отношения: относительно короткое расстояние между сущностями в тексте, т.е. в **контексте**.
- **Размеченный контекст** – контекст с выделенной парой  $\langle e_i, e_j \rangle$
- Извлечение отношений – разметка POS и NEG среди множества *нейтрально отмеченных* контекстов.

# Подходы автоматической разметки

- ① Сверточные и рекуррентные нейронные сети с механизмом внимания:
  - CNN, PCNN, LSTM, BiLSTM;
  - ATT $CNN_e$ , ATT $PCNN_e$ , IAN $_{ends}$ , BiLSTM, ATT-BLSTM;
- ② Языковые модели<sup>[5]</sup>:
  - mBERT, RUBERT, SENTRUBERT.

---

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

## Представление входной информации

### Входные данные

Говорить о разделении **кавказского региона<sub>e</sub>** из-за конфронтации **России<sub>subj</sub>** и  
**Турции<sub>obj</sub>** пока не приходится, хотя опасность есть.



Для сверточных и рекуррентных нейронных сетей  
говорить о разделение  $E$  из-за **конфронтация<sub>neg</sub>**  $E_{subj}$  и  $E_{obj}$  **не-приходиться<sub>neg</sub>**  
**СОММА** хотя опасность есть **DOT**



### Для языковых моделей

ТЕХТА: Говорить о разделении  $E$  из-за конфронтации  $E_{subj}$  и  $E_{obj}$  пока не-приходится , хотя опасность есть .

TEXTB<sub>NLI</sub>:  $E_{subj}$  к  $E_{obj}$  в контексте « $E_{subj}$  и  $E_{obj}$ »

3 Фреймы размечены коллекцией RuSentiFrames:

<https://github.com/nicolay-r/RuSentiFrames/tree/v2.0>

## Коллекции данных

- ① **RuSentRel**<sup>4</sup>: статьи про международные отношения России

Параметр	Значение
Число документов	73
Предложений на документ	105.8
Сущностей на документ	247
POS и NEG пар сущностей на документ	11.47

- ② **RuAttitudes**<sup>5</sup>: автоматически размеченная коллекция текстов с использованием подхода Distant Supervision (RuSentiFrames коллекция).

Версия	2.0-LARGE
Документов	<b>134442</b>
Отношений на документ	2.26

4 <https://github.com/nicolay-r/RuSentRel/tree/v1.1>

5 <https://github.com/nicolay-r/RuAttitudes/tree/v2.0>

## Результаты на коллекции RuSentRel<sup>[5]</sup>, 3-fold CV

<https://github.com/nicolay-r/RuSentRel-Leaderboard>

Модель	$F_1(P, N)$
SentenceRuBERT ( $NLI_{\text{pretrain}} + NLI_{\text{ft}}$ )*	<b>39.0</b>
PCNNends*	32.2
SentenceRuBERT (NLI)	<b>33.4</b>
AttPCNN <sub>ends</sub>	29.9
IAN <sub>ends</sub>	30.8
PCNN	29.6
Согласие экспертов	55.0

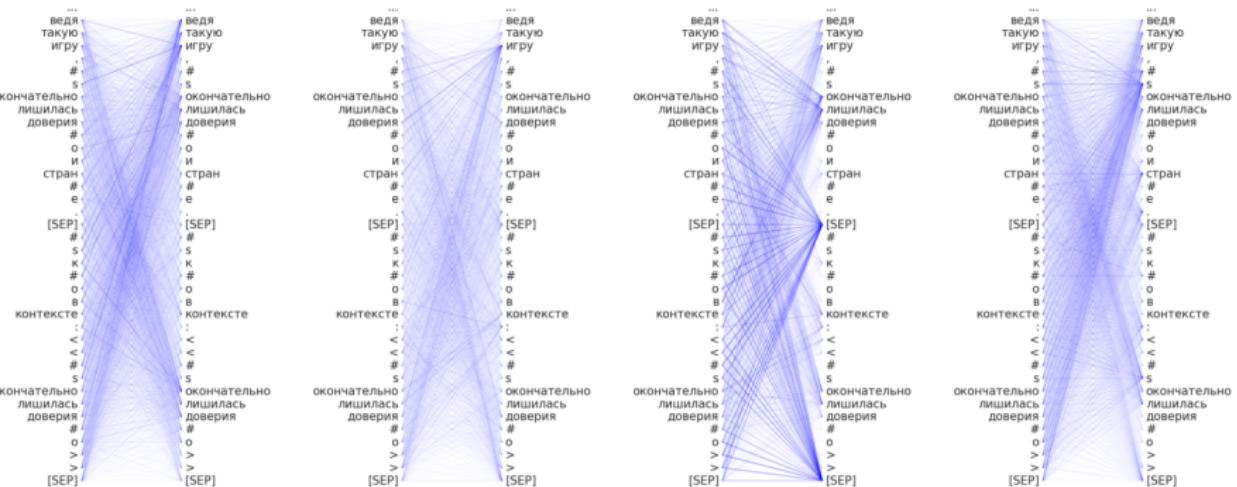
\* Предобучены на RuAttitudes<sup>6</sup>   \*\* Для MPQA-3.0,  $F_1 = 36.0$ <sup>[6]</sup>

6 <https://github.com/nicolay-r/RuAttitudes/tree/v2.0>

[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.

# Эволюция механизма внимания SentenceRuBERT

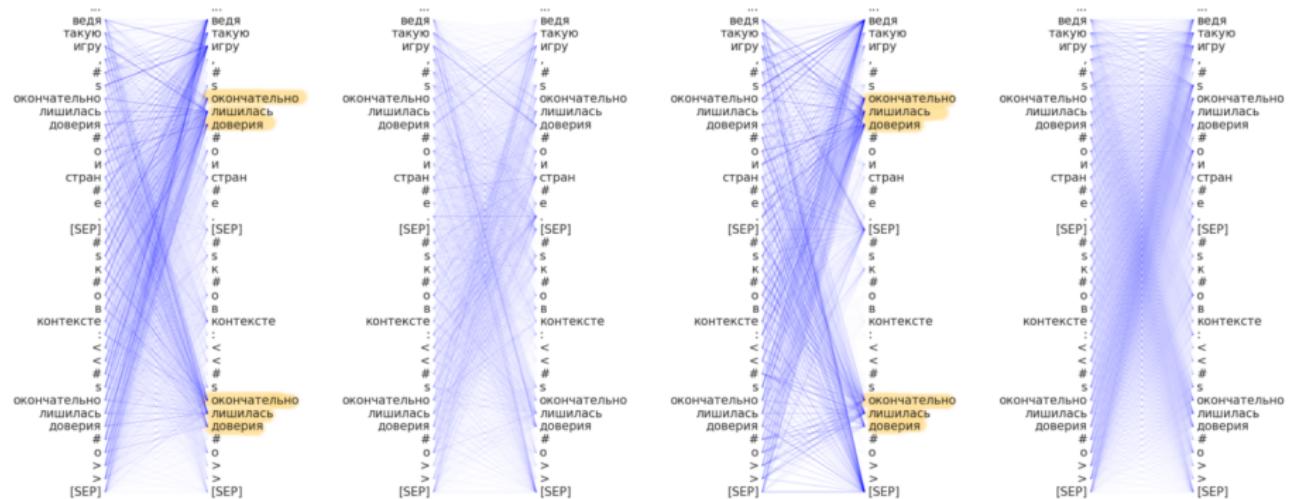
## SentenceRuBERT (голова 2, слои слева-направо: 2, 4, 8, 11)<sup>7</sup>



7 пример: *ведя такую игру, #S окончательно лишилась доверия #0 и стран #E. [SEP]*  
*#S к #0 в контексте: « #S окончательно лишилась доверия #0 » [SEP]*

# Эволюция механизма внимания SentenceRuBERT

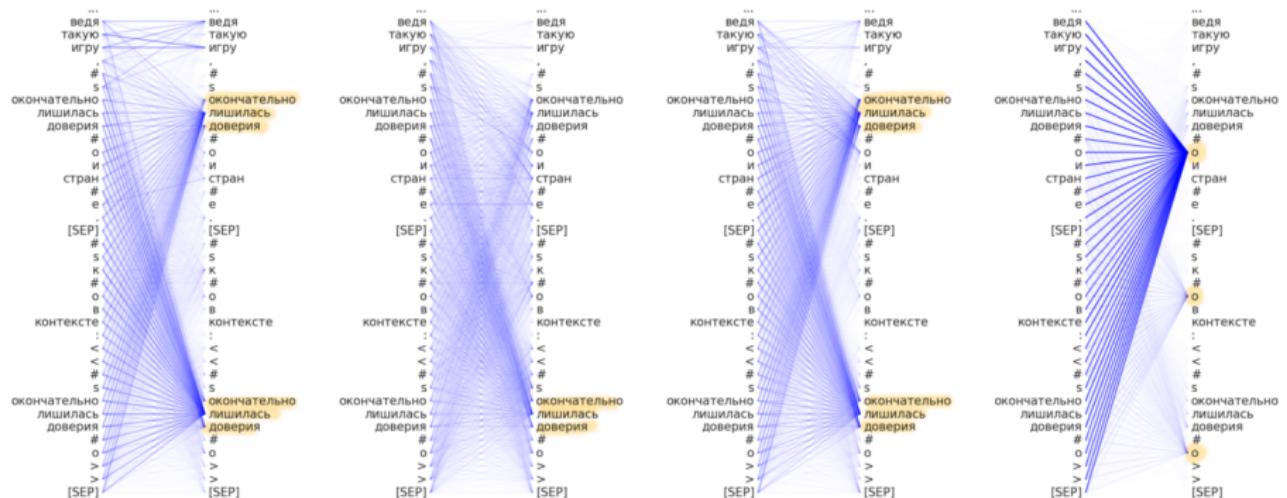
## SentenceRuBERT + 4 эпохи на RuAttitudes (NLI<sub>pretrain</sub>)<sup>8</sup>



8 повышено внимание к фреймам: *окончательно, лишиться, доверия* на слоях 2 (слева) и 8

# Эволюция механизма внимания SentenceRuBERT

SentenceRuBERT (NLI<sub>pretrain</sub>) + 4 эпохи дообучения на RuSentRel<sup>9</sup>



9 Еще большее внимание к фреймам (*окончательно, лишиться, доверия*), а также к участникам отношения #S и #0 (верхний слой)

ARElight

# ARElight страница проекта<sup>10</sup>

Document Attitudes  
Annotation App for your  
Relation Extraction task

**Infer attitudes** from large Mass-media  
documents or **sample texts** for your  
Machine Learning models applications

Get Started

Powered by AREkit

The ARElight logo is prominently displayed in the bottom right corner, featuring the word "ARE" in white on a purple background and "light" in yellow with a purple outline.

10 <https://nicolay-r.github.io/arelight-page/>

# Sample Texts – возможность ARElight

В случае нейронных сетей (фреймы, признаки):

id	doc_id	label	text_a	s_ind	t_ind	sent_ind	entity_values	entity_types	frames	frame_conncts	uint	syn_subjs	syn_objs	entities	pos_tags
00_I0_	0	0	0< >< >< >	5	19	1	МОСКВА,НАТО,россия	GPE,ORG,GPE			5	19.5,19,25	15,15,15,11,13,13,15,2,14,15,13		
01_I0_	0	0	0< >< >< >	5	25	1	МОСКВА,НАТО,россия	GPE,ORG,GPE			5	25.5,19,25	15,15,15,11,13,15,2,14,15,13		
02_I0_	0	0	0< >< >< >	19	5	1	***,НАТО,россия	GPE,ORG,GPE			19	5.5,19,25	15,15,15,11,13,15,2,14,15,13		
03_I0_	0	0	0< >< >< >	19	25	1	***,БИ,ДО,ПОД	GPE,ORG,GPE			19	25.5,19,25	15,15,15,11,13,15,2,14,15,13		
04_I0_	0	0	0 объект намер	4	0	0	ШАДЛОССО	GPE,VERB			4	0.0,4	15,12,14,11,15		
05_I0_	0	0	0< >< >< >	25	5	1	МОСКВА,НАТО,россия	GPE,ORG,GPE			25	5.5,19,25	15,15,15,11,13,15,2,14,15,13		
06_I0_	0	0	0< >< >< >	25	19	1	МОСКВА,НАТО,россия	GPE,ORG,GPE			25	19.5,19,25	15,15,15,11,13,15,2,14,15,13		
07_I0_	0	0	0 subject намерен ввести-санкции против object	0	4	0	сша,россия	GPE,PERSON			0	4.0,4	15,12,14,11,15		

В случае языковых моделей BERT (TEXTA, TEXTB):

id	doc_id	label	text_a	text_b	s_ind	t_ind	sent_ind	entity_values	entity_types	entities
00_I0_	0	0	0< > марта президент #0 И0 провел переговоры с лидерами стран #5 И0 в контексте : <<#5 провел переговоры с лидерами стран #0>>		4	10	3	ставка,байден,евросоюза	GPE,PERSON,ORG,GPE,GPE,GPE	3,4,10,12,21,24
01_I0_	0	0	0< > марта президент #0 И0 провел переговоры с лидерами стран #5 И0 в контексте : <<#5 провел переговоры с лидерами стран #0 в #E выразил		4	24	3	ставка,байден,евросоюза	GPE,PERSON,ORG,GPE,GPE,GPE	3,4,10,12,21,24
02_I0_	0	0	0< > марта президент #0 И0 провел переговоры с лидерами стран #5 И0 в контексте : <<#5 И0>>		4	3	3	ставка,байден,евросоюза	GPE,PERSON,ORG,GPE,GPE,GPE	3,4,10,12,21,24
03_I0_	0	0	0< > марта президент #0 И0 провел переговоры с лидерами стран #5 И0 в контексте : <<#5 провел переговоры с лидерами стран #E в #E выразил		6	21	3	ставка,байден,евросоюза	GPE,PERSON,ORG,GPE,GPE,GPE	3,4,10,12,21,24
04_I0_	0	0	0< > И0 крайне зависим от #0 в плане поставок нефти и газа.		0	4	4	европейский союз,россия	LOC,GPE	0.4
04_I1_	0	0	0< > Поскольку И0 является важным узлом транспортировки российской нефти и газа.		9	13	7	украина,европу,россия	GPE,GPE,GPE	1.9,13
05_I0_	0	0	0< > марта президент #0 И0 провел переговоры с лидерами стран #5 И0 в #0 в контексте : <<#5 ее концепция #0>>		10	4	4	ставка,байден,евросоюза	GPE,PERSON,ORG,GPE,GPE,GPE	3,4,10,12,21,24
06_I0_	0	0	0< > марта президент #0 И0 провел переговоры с лидерами стран #5 И0 в #0 в контексте : <<#5 провел переговоры с лидерами стран #0 в #E выразил		24	4	3	ставка,байден,евросоюза	GPE,PERSON,ORG,GPE,GPE,GPE	3,4,10,12,21,24
07_I0_	0	0	0< > марта президент #0 И0 провел переговоры с лидерами стран #5 И0 в #0 в контексте : <<#5 И0>>		10	3	3	ставка,байден,евросоюза	GPE,PERSON,ORG,GPE,GPE,GPE	3,4,10,12,21,24
08_I0_	0	0	0< > марта президент #0 И0 провел переговоры с лидерами стран #5 И0 в #0 в контексте : <<#5 в #E вызвал внимание рынка и предложил о тп		10	21	3	ставка,байден,евросоюза	GPE,PERSON,ORG,GPE,GPE,GPE	3,4,10,12,21,24
09_I0_	0	0	0< > марта президент #0 И0 провел переговоры с лидерами стран #5 И0 в #0 в контексте : <<#5 И0 провел переговоры с лидерами стран #0 в #E>>		24	3	3	ставка,байден,евросоюза	GPE,PERSON,ORG,GPE,GPE,GPE	3,4,10,12,21,24
09_I1_	0	0	0< > марта президент #0 И0 провел переговоры с лидерами стран #5 И0 в #0 в контексте : <<#5 удалось уговорить #0>>		24	21	3	ставка,байден,евросоюза	GPE,PERSON,ORG,GPE,GPE,GPE	3,4,10,12,21,24
10_I0_	0	0	0< > После начала роско-украинского конфликта страны #5 одни #5 И0 в контексте : одни из других вводят в отношении #0>>		5	12	0	ондара,россия	LOC,GPE	5.12
11_I0_	0	0	0< > В настоящем #0 И0 в контексте : <<#5 продолжает		5	7	0	российская,украинская	LOC,GPE	5.7
12_I0_	0	0	0< > Поскольку И0 является важным узлом транспортировки российской нефти и газа.		13	1	7	украина,европа,россия	GPE,LOC,GPE	1.9,13
13_I0_	0	0	0< > марта президент #0 И0 провел переговоры с лидерами стран #5 И0 в #0 в контексте : <<#5 ее концепция #0>>		4	6	4	европейский союз,россия	LOC,GPE	0.4
14_I0_	0	0	0< > Поскольку И0 является важным узлом транспортировки российской нефти и газа.		9	7	4	европейский союз,россия	GPE,GPE,GPE	1.9,13
15_I0_	0	0	0< > После начала роско-украинского конфликта страны #5 одни #5 И0 в контексте : <<#5 ее концепция #0>>		13	5	0	ондара,россия	LOC,GPE	5.12
16_I0_	0	0	0< > марта президент #0 И0 провел переговоры с лидерами стран #5 И0 в #0 в контексте : <<#5 одни из других вводят в отношении #0>>		3	4	3	ставка,байден,евросоюза	GPE,PERSON,ORG,GPE,GPE,GPE	3,4,10,12,21,24
17_I0_	0	0	0< > марта президент #0 И0 провел переговоры с лидерами стран #5 И0 в #0 в контексте : <<#5 одни из двух вводят в отношении #0>>		21	4	3	ставка,байден,евросоюза	GPE,PERSON,ORG,GPE,GPE,GPE	3,4,10,12,21,24
18_I0_	0	0	0< > марта президент #0 И0 провел переговоры с лидерами стран #5 И0 в #0 в контексте : <<#5 И0 провел переговоры с лидерами стран #0 в #E>>		3	10	3	ставка,байден,евросоюза	GPE,PERSON,ORG,GPE,GPE,GPE	3,4,10,12,21,24
19_I0_	0	0	0< > марта президент #0 И0 провел переговоры с лидерами стран #5 И0 в #0 в контексте : <<#5 в #E вызвал внимание рынка и предложил о тп		21	10	3	ставка,байден,евросоюза	GPE,PERSON,ORG,GPE,GPE,GPE	3,4,10,12,21,24
20_I0_	0	0	0< > марта президент #0 И0 провел переговоры с лидерами стран #5 И0 в #0 в контексте : <<#5 удалось уговорить #0>>		21	24	3	ставка,байден,евросоюза	GPE,PERSON,ORG,GPE,GPE,GPE	3,4,10,12,21,24
21_I0_	0	0	0< > В настоящем времена конфликт между #0 и #5 продолжает		7	5	0	российская,украинская	GPE,GPE	5.7
22_I0_	0	0	0< > Поскольку И0 является важным узлом транспортировки российской нефти и газа.		1	13	7	украина,европа,россия	GPE,GPE,GPE	1.9,13

## Особенности набора инструментов AREkit



AREkit<sup>11</sup> – набор инструментов на Python для работы с отношениями на уровне контекстов и документов с поддержкой синонимии именованных сущностей

- ① Аннотации контекстов с отношениями
- ② Экспорт контекстов с отношениями из коллекций новостей (ядро на Pandas)
- ③ Обучение и применение нейросетей (Tensorflow), contrib.

Схожие решения: OpenNRE<sup>12</sup>, DeRE<sup>13</sup>

---

11 <https://nicolay-r.github.io/arekit-page/>

12 <https://github.com/thunlp/OpenNRE>

13 <https://github.com/ims-tcl/DeRE>

## Набор инструментов для работы с текстом в AREkit

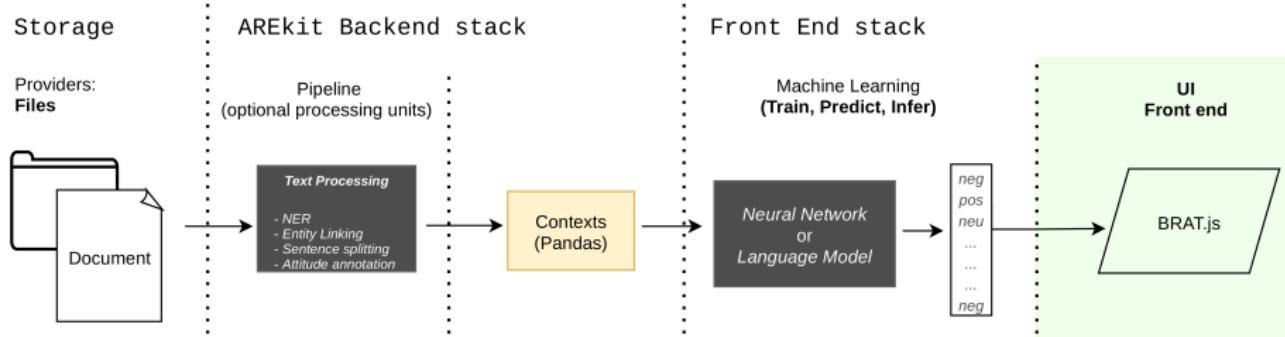
- `SynonymsCollection` – коллекция синонимов
  - `TextParser` – pipeline из разметки сущностей, фреймов, и пр.
  - `OpinionAnnotationAlgorithms` – алгоритм разметки пар
  - `DataFolding` – настройка разбиения данных (TRAIN/TEST/DEV)
  - `LabelsProvider` – обозначение классов
  - `EntitiesFormatter` – форматирование сущностей (#S, #O, #E)

13 <https://nicolay-r.github.io/blog/articles/2022-05/process-mass-media-relations-with-arekit>

## Реализация вывода отношений в ARElight

**Бэкенд:** AREkit<sup>14</sup> – обработка текстов, создание **итератора контекстов** с размеченными парами сущностей

**Фронтенд:** использование инструментов AREkit для подготовки данных; использование, обучение, применение моделей; визуализация результата с помощью brat;



14 <https://github.com/nicolay-r/AREkit/tree/0.22.0-rc>

## Docker версия ARElight для вывода отношений



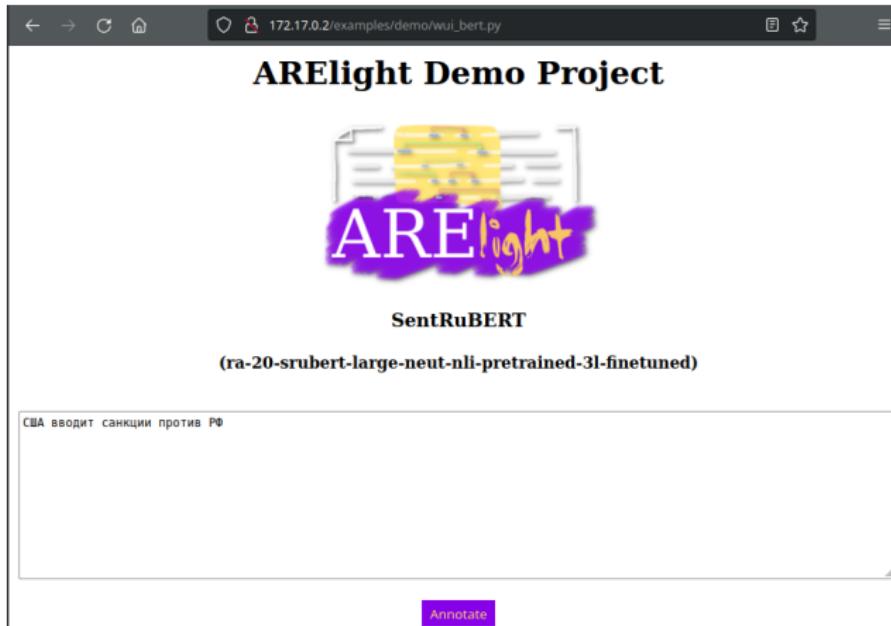
Docker + NVidia docker:

- DeepPavlov: NER (BERT<sub>ontonotes</sub>) и применение языковой модели SentenceRuBERT.
- Apache, CGI-based

Требования:

- 12 GB RAM
- 6 GB VRAM (NVidia GTX 1060 TI или выше)

Пример интерфейса в браузере<sup>15</sup>

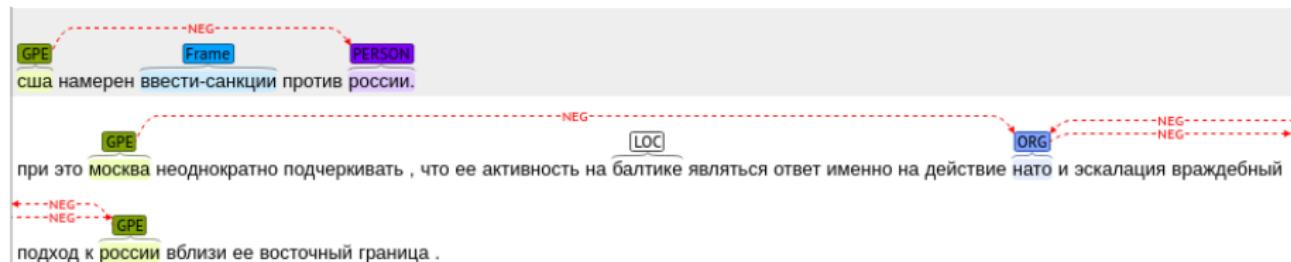


15 <https://github.com/nicolay-r/ARElight/tree/v0.22.0#readme>

## Infer Attitudes – возможность ARElight

PCNN<sup>16</sup>:

- Лемматизация текста
- Разметка фреймов на основе RuSentiFrames коллекции



16 [http://172.17.0.2/examples/demo/wui\\_nn.py](http://172.17.0.2/examples/demo/wui_nn.py)

# Infer Attitudes – возможность ARElight

Полный текст примера доступен по ссылке<sup>17</sup>.

SentenceRuBERT<sup>18</sup>:

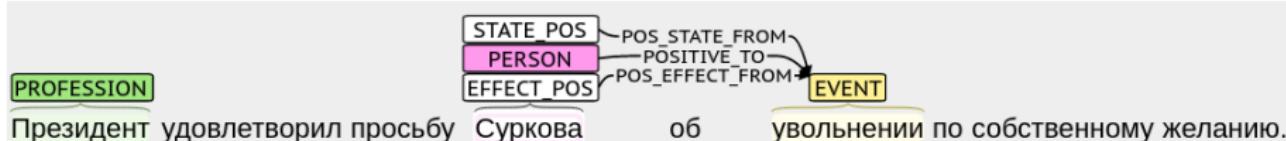
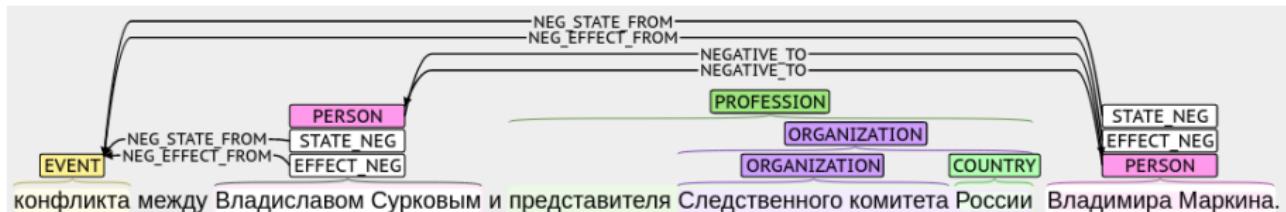


17 <https://raw.githubusercontent.com/nicolay-r/ARElight/main/data/texts-inosmi-rus/e1.txt>

18 [http://172.17.0.2/examples/demo/wui\\_bert.py](http://172.17.0.2/examples/demo/wui_bert.py)

## Повышение детализации анализа

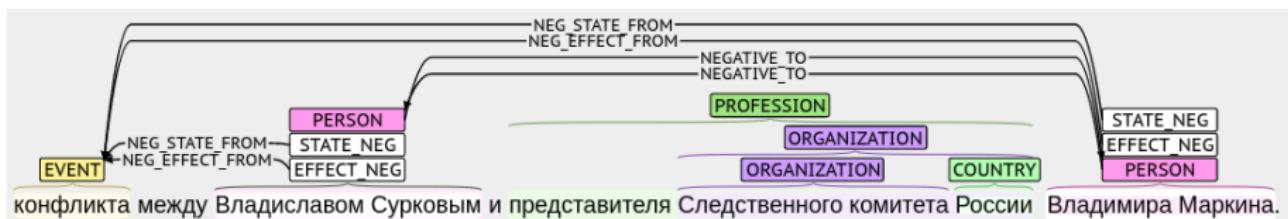
- Предсказание тональности к сущностям: PER, ORG, COUNTRY
- Разные направления тональность сущности: STATE, EFFECT.
- Не всегда сущность может быть источником, а например событие :<sup>19</sup>



19 например: конфликт, увольнение оказывают разный эффект

## Сравнение результатов

В разметке сторонней коллекции:



Результат демо-версии ARElight с SentRuBERT:



19 Для второго примера, сущность всего одна в тексте, поэтому отношение извлечено не будет

## Дальнейшее развитие



Ядро<sup>20</sup>, список нововведений в релизе 0.22.1:  
<https://github.com/nicolay-r/AREkit/issues/323>

- провайдеры для работы с коллекцией (NEREL<sup>[7]</sup>)
- поддержка вложенных объектов
- декларативное описание извлекаемых отношений



- Добавление EFFECT и STATE в разметке сущностей
- Ваши предложения

20 Текущая версия: <https://github.com/nicolay-r/AREkit/tree/0.22.0-rc>

[7] Natalia Loukachevitch и др. «NEREL: A Russian Dataset with Nested Named Entities, Relations and Events». В: *Proceedings of RANLP*. 2021, с. 876—885.

## Заключение

- Задача извлечения оценочных отношений<sup>21</sup> из аналитических текстов:
  - Актуальна в случае частого упоминания именованных сущностей в тексте и выражения мнения к ним
- Рассмотрены особенности реализации набора инструментов AREkit<sup>22</sup>:
  - **Возможность разметки оценочных пар** между сущностями;
  - **Поддержка синонимии** сущностей в представлениях отношений.
- Представлена демо версия проекта ARElight<sup>23</sup>.

---

21 <http://nlpprogress.com/russian/sentiment-analysis.html>

22 <https://github.com/nicolay-r/AREkit/tree/0.22.0-rc>

23 <https://github.com/nicolay-r/ARElight/tree/v0.22.0>

# Спасибо за внимание!



<https://nicolay-r.github.io>