Набор инструментов AREkit для извлечения оценочных отношений из новостных текстов

Русначенко Николай

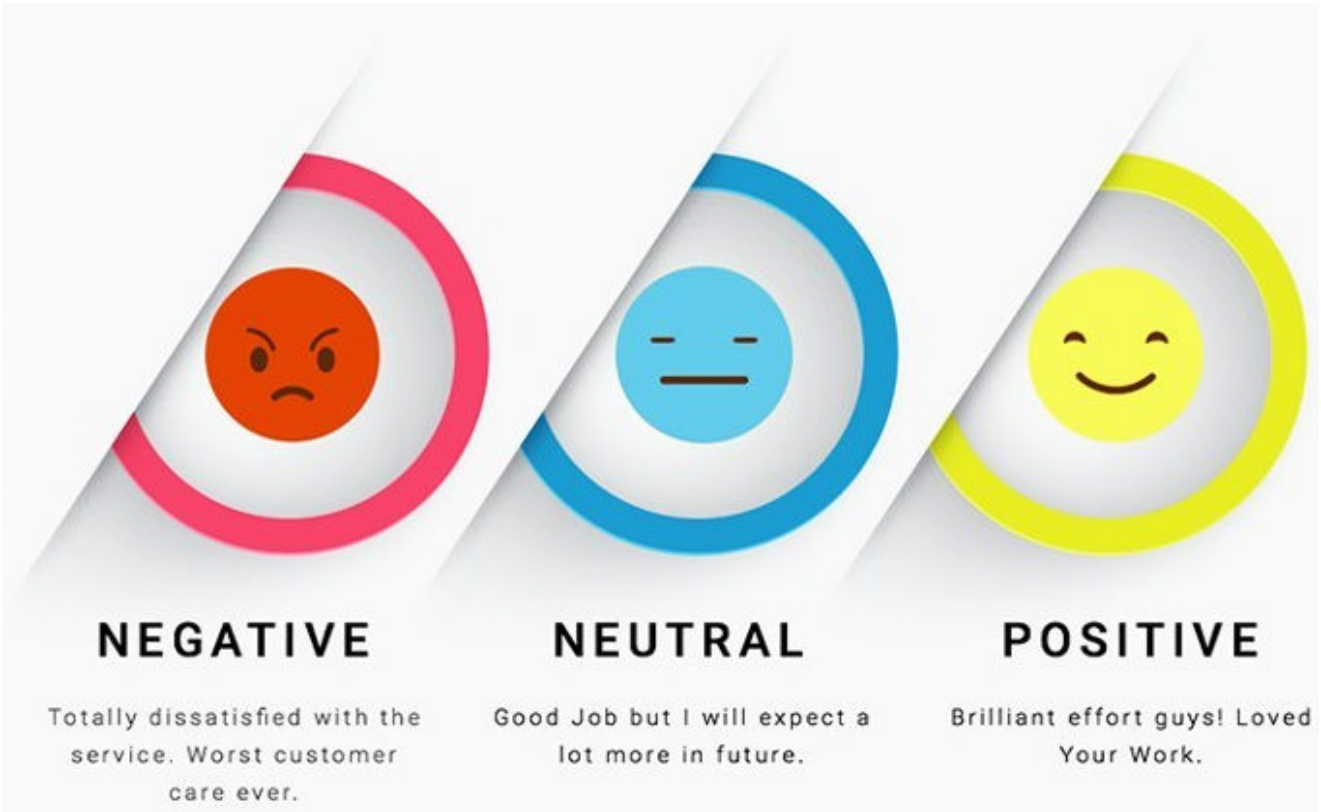
МГТУ им. Н.Э. Баумана, Москва, Россия

kolyarus@yandex.ru

nicolay-r.github.io

November 26'th, 2020

Анализ тональности



Text classification

Первая попытка предложения постановки задачи[1]:

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

d – документ
c – класс {positive, negative}

"Качество картинки этой камеры в ночное время – потрясающее"

$$\langle d \rangle \rightarrow positive$$

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

Targeted sentiment analysis

Предполагает указание сущности/объекта в качестве параметра[2]:

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

e_j – Объект, либо сущность

"Качество снимков такой камеры$_e$ в ночное время просто потрясающее, особенно если пользоваться штативом$_e$"

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

[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[3]:

1. Извлечение аспекта;
2. Аспектно-ориентированный анализ тональности:

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

a_k – аспект (характеристика объекта)

"Качество картинки этой камеры$_e$ – потрясающее . . ."[3]

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

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

Opinion Definition

Определение мнения (от англ. Opinion) согласно работам[3,4] определяется пятеркой элементов:

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

h_t – автор

t_l – время

[4] Bing Liu и др. «Sentiment analysis and subjectivity». в: *Handbook of natural language processing* 2 (2010), с. 627—666.

Attitude Definition

Мнения образуемые между упомянутыми именованными сущностями[1]:

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

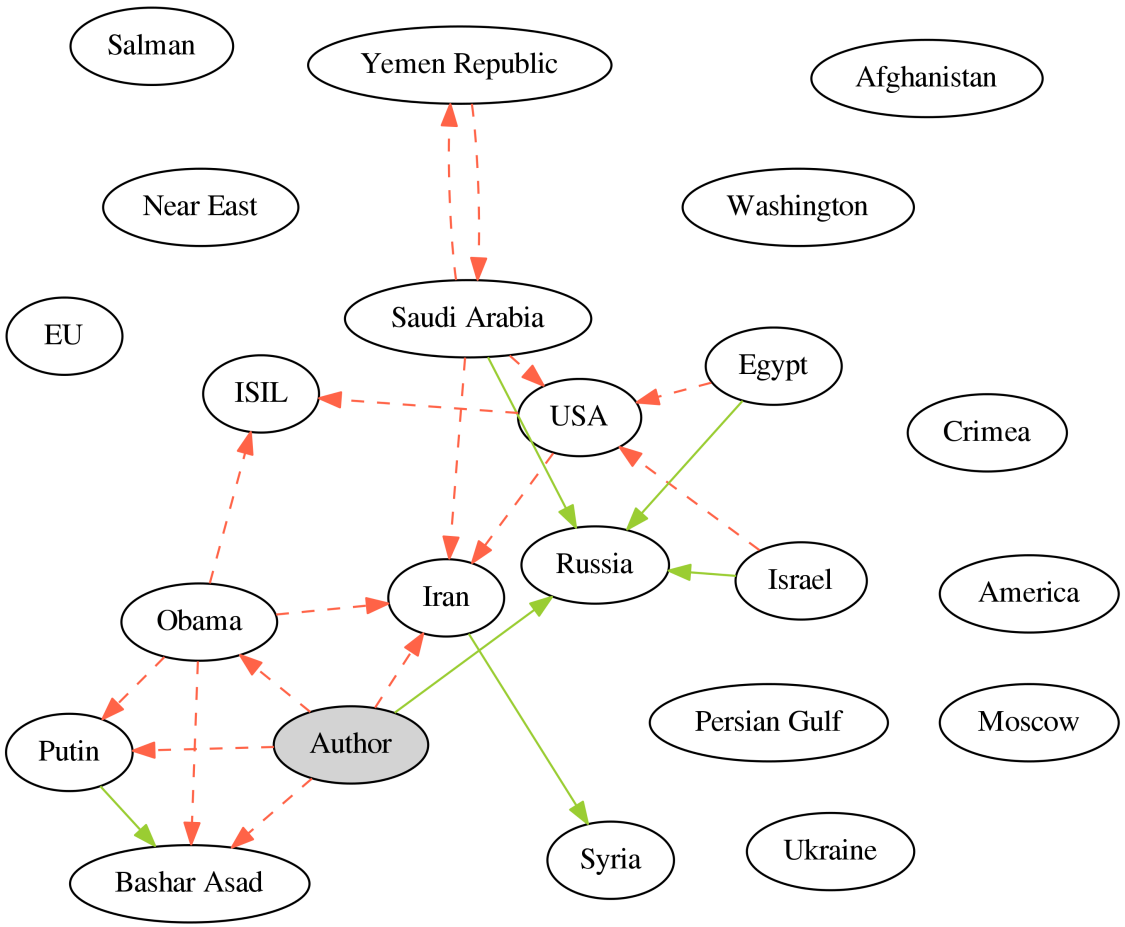
e_m – Субъект
e_j – Объект

(Субъект $\rightarrow$ Объект)

“ ... Москва$_e$ недовольна решением Варшавы$_e$ ... ”
$\langle e_m, e_j \rangle \rightarrow \text{NEG}$

1 Теперь вместо *автора* рассматривается *субъект*

Представление отношений между именованными сущностями в документе



Задача извлечения оценочных отношений

Постановка задачи I

- Дана коллекция C, состоящая из набора аналитических статей;
- Каждая статья включает: (1) документ D_i – последовательность символов, (2) список упомянутых именованных сущностей E_i:

$$D_i = \{c_1, \ldots, c_{|D_i|}\} \quad E_i = [e_1, \ldots, e_{|E_i|}]$$

- Под *именованной сущностью* (E) понимается слово или словосочетание (v), указывающее на объект реальности:

$$e = \langle v, t \rangle \quad v = [c_b ... c_e] \quad t \in T$$

- Для синонимичных упоминаний:

... (Россия$_e$, РФ$_e$, Российская Федерация$_e$) ...

Постановка задачи II

задано отображение G для множества V всех вхождений именованных сущностей коллекции C на множество индексов групп G:

$$\mathrm{G} : V \to G \quad G = \{g_1, \ldots, g_{|G|}\} \quad g_i \in \mathbb{N}$$

- Обозначим $a = \langle v_i, v_j \rangle$ *парой субъект-объект*, где v_i и v_j вхождения сущностей e_i и e_j соответственно;
- Необходимо для каждого $D_i \in C$ составить список оценочных отношений[5] $(\mathbf{a})$, где l_j оценка пары (позитивная – POS, или негативная – NEG):

$$\mathbf{a} = \left[\langle a_j, l_j \rangle \right]_{j=1}^{|\mathbf{a}|}$$

[5] 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.

Примеры

In: При этом Москва$_e$ неоднократно подчеркивала, что ее активность на балтике$_e$ является ответом на действия **НАТО**$_e$ и эскалацию враждебного подхода к **России**$_e$ вблизи ее восточных границ ...
Out: (НАТО, Россия, NEG), (Россия, НАТО, NEG)

In[2]: Трамп$_e$ обвинил Китай$_e$ и Россию$_e$ в «игре деноминации валют»
Out: (Трамп, Китай, NEG) (Трамп, Россия, NEG)

In: Говорить о разделении кавказского региона$_e$ из-за конфронтации **России**$_e$ и **Турции**$_e$ пока не приходится, хотя опасность есть.
Out: (Турция, Россия, NEG) (Россия, Турция, NEG)

2 Сложность структуры, сущности «Россия» и «Китай» нейтральны друг к другу

Основная идея

- Критерий наличия отношения: относительно короткое расстояние между сущностями в тексте, т.е. в **контексте**.
- **Размеченный контекст** – контекст с выделенной парой «субъект-объект» (a);
- Извлечение отношений – разметка POS и NEG среди множества *нейтрально отмеченных* контекстов.

Подходы

Классические методы машинного обучения (ручные признаки)[5]:

- KNN, SVM, Naïve Bayes, Gradient Boosting

Обучение с учителем на размеченных контекстах:

1. Сверточные и рекуррентные нейронные сети, сети с вниманием:
   - CNN, PCNN, LSTM, BiLSTM;
   - AttCNN$_e$, AttPCNN$_e$, IAN$_{ends}$, BiLSTM, Att-BLSTM;
2. Языковые модели[6]:
   - mBERT, RuBERT, SentRuBERT.

---

[6] 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.

Набор данных

1. **RuSentRel**[3]: статьи про международные отношения России;

Параметр (среднее)	TRAIN	TEST
Число документов	44	29
Предложений на документ	74.5	137
Сущностей на документ	194	300
POS пар сущностей на документ	6.23	14.7
NEG пар сущностей на документ	9.33	15.6

2. **RuAttitudes**[4]: автоматически размеченные отношения с помощью лексикона RuSentiFrames.

RuAttitudes	Версия	2.0-LARGE
	Новостей	**134442**
	Отношений на новость	2.26
	Предложений на новость	0.88

3 https://github.com/nicolay-r/RuSentRel/tree/v1.1
4 https://github.com/nicolay-r/RuAttitudes/tree/v2.0

Представление контекстов

Контекст
Говорить о разделении кавказского региона$_e$ из-за конфронтации России$_{subj}$ и Турции$_{obj}$ пока не приходится, хотя опасность есть.

↓

Для нейронных сетей
говорить о разделение E из-за ***конфронтация***$_{neg}$ $\underline{E}_{subj}$ и $\underline{E}_{obj}$ ***не-приходиться***$_{neg}$ COMMA хотя опасность есть DOT

↓

Для языковых моделей
TextA: Говорить о разделении E из-за конфронтации $\underline{E}_{subj}$ и $\underline{E}_{obj}$ пока не-приходится , хотя опасность есть .
TextB$_{\text{NLI}}$: $\underline{E}_{subj}$ к $\underline{E}_{obj}$ в контексте «$\underline{E}_{subj}$ и $\underline{E}_{obj}$» TextB$_{\text{QA}}$: Что вы думаете по поводу отношения $\underline{E}_{subj}$ к $\underline{E}_{obj}$ в контексте : «$\underline{E}_{subj}$ и $\underline{E}_{obj}$» ?

4 Для нейронных сетей используется лексикон RuSentiFrames

Результаты в рамках корпуса RuSentRel[6]

Модель	$F1_{cv\text{-}3}$	$F1_{test}$
SentenceRuBERT ($NLI_{pretrain}$ + NLI_{ft})*	39.0	38.0
SentenceRuBERT ($NLI_{pretrain}$ + QA_{ft})*	38.4	41.9
PCNNends*	32.2	39.9
IAN_{ends}	30.8	32.2
$AttPCNN_{ends}$	29.9	32.6
PCNN	29.6	32.5
Gradient Boosting (Grid Search)	20.3	28.0
Random Forest (Grid search)	19.1	27.0
Согласие экспертов	55.0	55.0

* в предобучении/обучении используется RuAttitudes.

** в рамках корпуса MPQA-3.0, F1 = 36.0[7]

[7] 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.

Программные средства

Введение

1. AREkit[5] – набор инструментов для разбора документов с аннотацией (1) именованных сущностей, (2) фреймов и (3) отношений на контекстном и документном уровнях

2. Приложений для Deep Learing экспериментов в рамках корпуса RuSentRel[6,7];

---

5 https://github.com/nicolay-r/AREkit

6 https://github.com/nicolay-r/neural-networks-for-attitude-extraction/tree/0.20.5

7 https://github.com/nicolay-r/bert-for-attitude-extraction-with-ds/

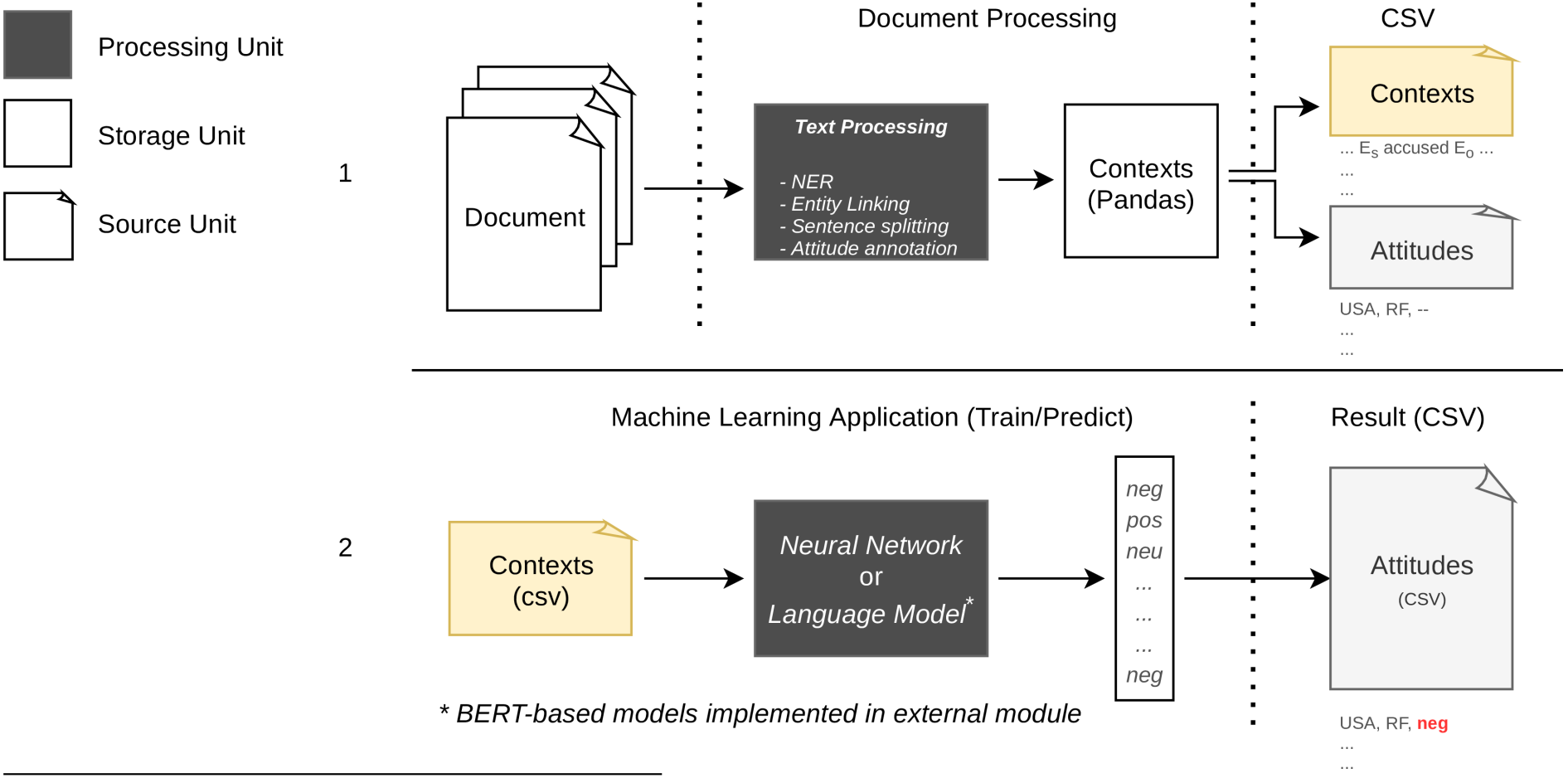
Обзор смежных фреймворков

- OpenNRE[8] – предлагает основу для моделей нейронных сетей для извлечения отношений (Backend для подготовленных данных);
- DeRE[9] – предлагает декларативный подход извлечения отношений (Backend для моделей машинного обучения);

8 https://github.com/thunlp/OpenNRE
9 https://github.com/ims-tcl/DeRE

Предоставляемые возможности

Текущая версия[10]: Поддержка двух сценариев

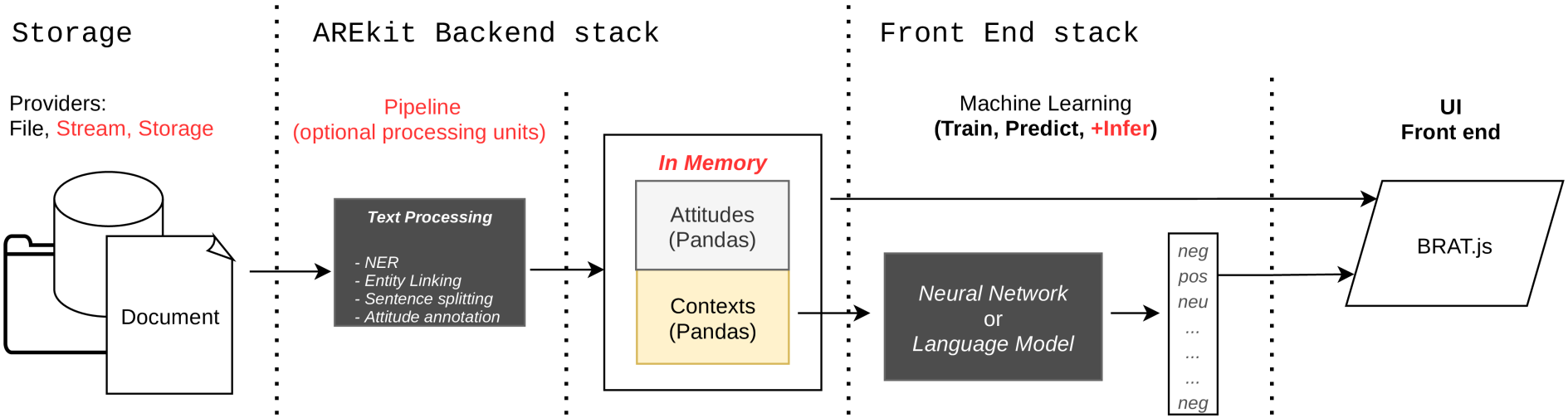


10 https://github.com/nicolay-r/AREkit/tree/0.21.0-rc

Создание демонстративной версии

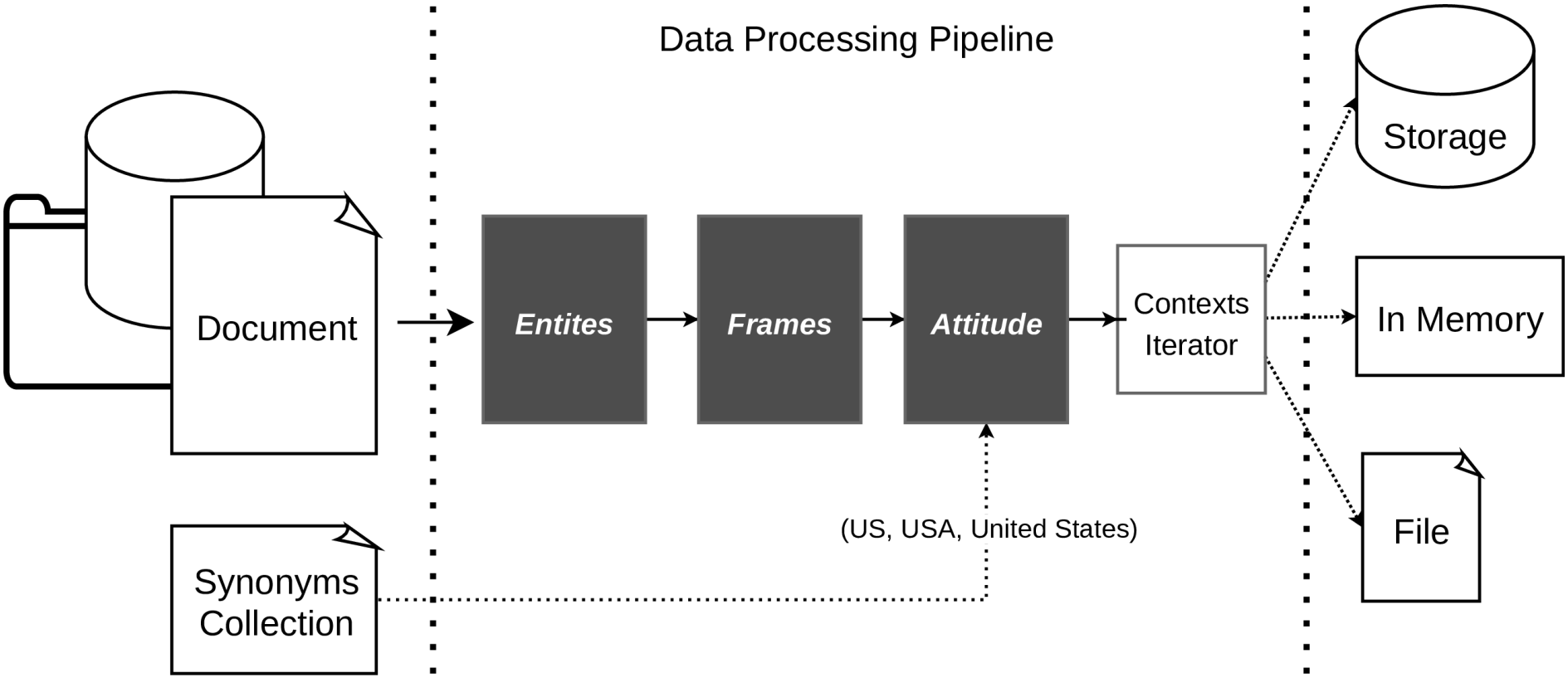
В разработке[11]:

- Поддержка различных источников данных;
- Pipelines.

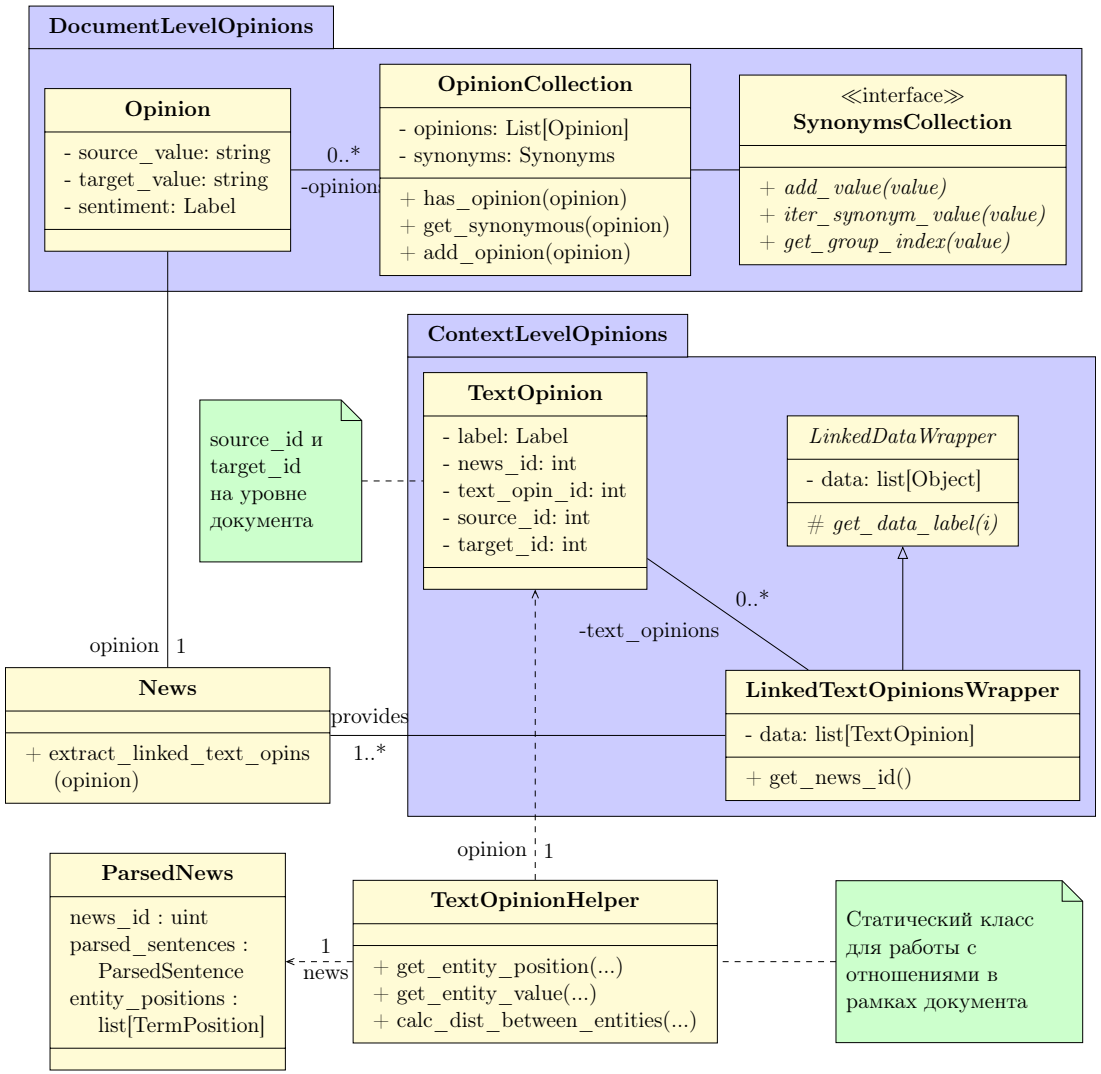


11 https://github.com/nicolay-r/AREkit/tree/0.21.1-rc

AREkit Backend

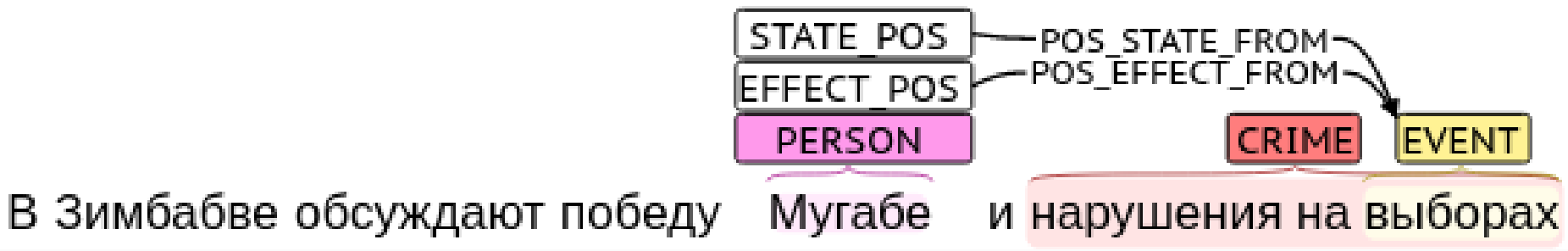


AREkit Backend: Attitude (Opinions) architecture



Добавление декларативного описания задачи[8]

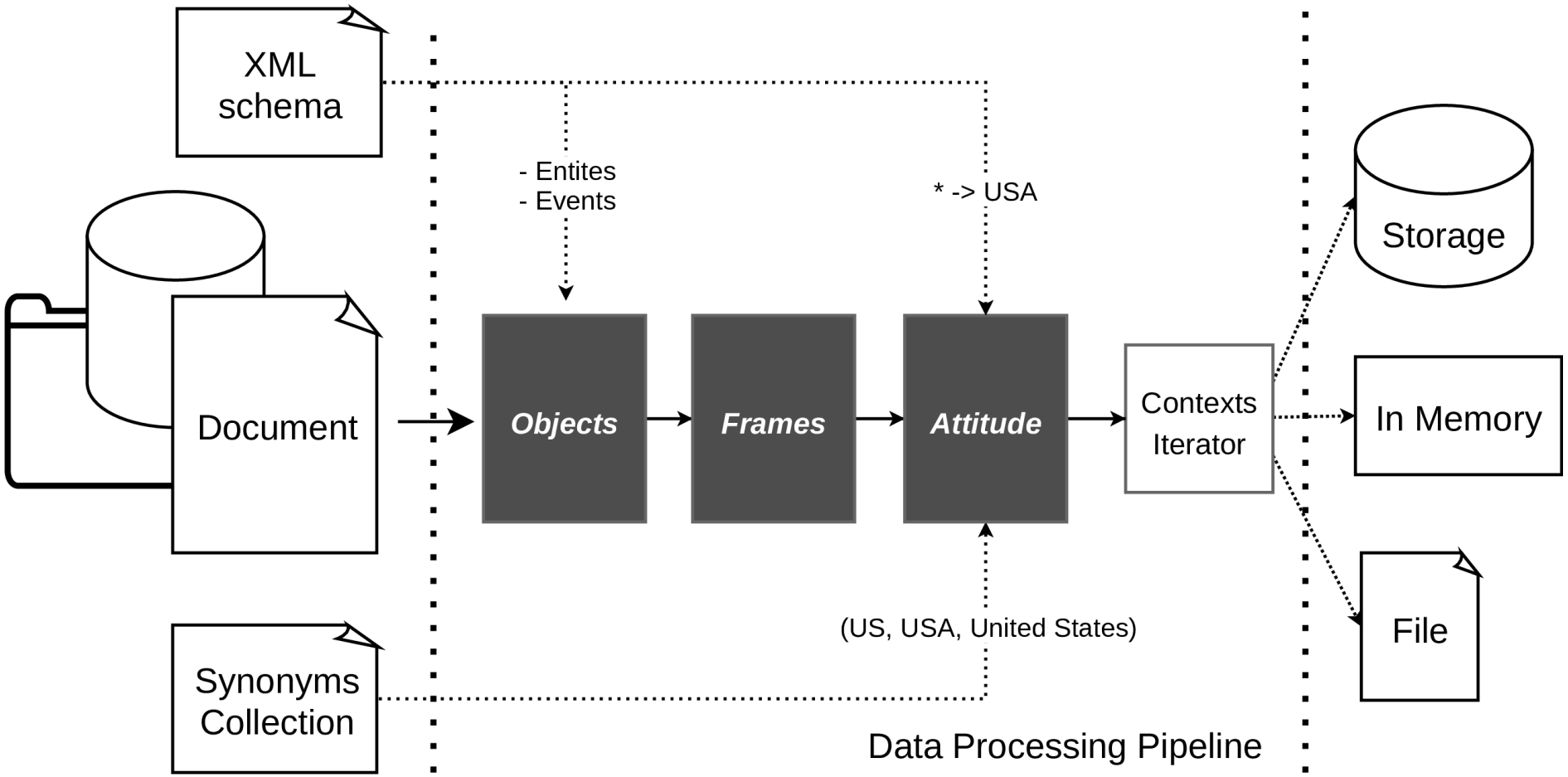
В случае наличия богатой аннотации в тексте:



- Декларированное описание извлекаемых отношений:
  США$_e$ $\rightarrow$ *
  * $\rightarrow$ РФ$_e$
  * $\rightarrow$ [EVENT [Евросоюз$_e$]]

[8] Heike Adel и др. «DERE: A task and domain-independent slot filling framework for declarative relation extraction». в: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2018, с. 42—47.

Добавление декларативного описания задачи



Заключение

- Задача извлечения оценочных отношений из аналитических текстов:
  - Актуальна в случае частого упоминания именованных сущностей в тексте и выражения мнения к ним;
- Рассмотрены особенности реализации набора инструментов AREkit;
  - **Возможность извлечения оценочных отношений** между сущностями на *контекстном* и *документном* уровнях;
  - **Поддержка синонимии** сущностей в представлениях отношений.

Спасибо за внимание!

https://nicolay-r.github.io

https://github.com/nicolay-r/ARElkit