Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский институт)

Представление на соискание учёной степени кандидата технических наук по специальности 05.13.11 Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Модели, методы и программные средства извлечения оценочных отношений на основе фреймовой базы знаний

Выступающий: Н. Л. Русначенко
Руководитель: д. т. н., проф. Н. В. Лукашевич

Москва, 2022

Тональность — это отношение (позитивное или негативное) некоторого лица относительно содержания текста или каких-то его аспектов;

- Общий анализ тональности:

  ... онлайн приложением банка остался доволен ... $\rightarrow$ POS

- Тональность текста к заданной сущности или ее аспектам

  ... *качество картинки* *фотокамеры*$_e$ *достаточно высокое ...* $\langle a, e \rangle \rightarrow$ POS

- Тональность отношений между упоминаемыми сущностями

  ... Москва$_e$ недовольна решением Варшавы$_e$ ... $\langle e_1, e_2 \rangle \rightarrow$ NEG

Актуальность. Задача извлечения оценочных отношений:

- нужна для более точного извлечения позиции автора к данным сущностям;
- оценочные отношения - это отдельный тип отношений для извлечения и анализа.

Проблемы:

- задача мало изучена:
- мало размеченных данных;

Извлечение оценочных отношений

Пример отношений в которых явно или неявно проявляется позитивное или негативное отношение сущностей друг к другу.

... При этом Москва$_e$ неоднократно подчеркивала, что ее активность на балтике$_e$ является ответом именно на действия **НАТО**$_e$ и эскалацию враждебного подхода к **России**$_e$ вблизи ее восточных границ ...
НАТО $\rightarrow$ Россия, Россия $\rightarrow$ НАТО Автор $\rightarrow$ Россия, Автор $\rightarrow$ НАТО

Другие направления[1] – таргетированный[2] анализ тональности:

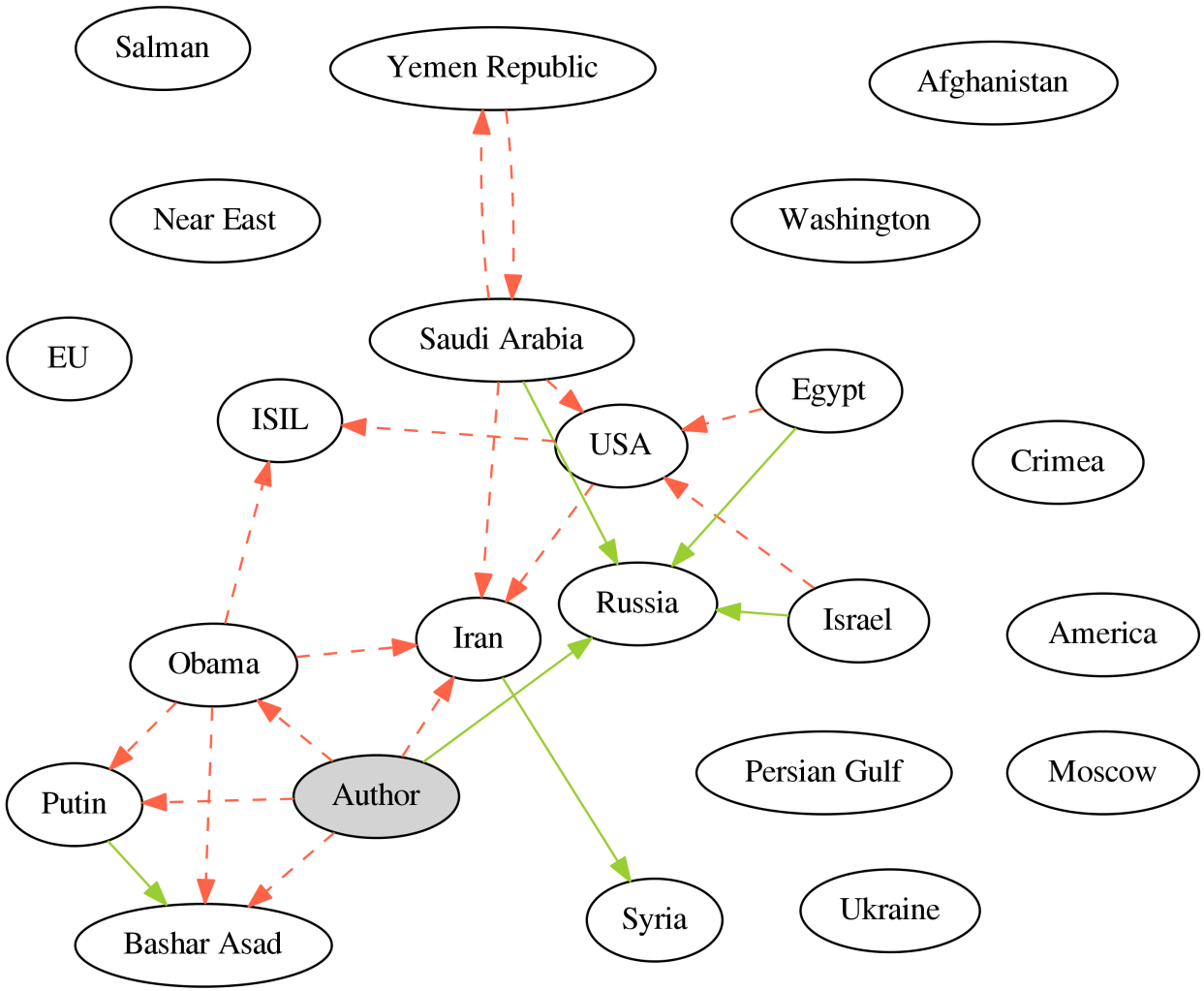
- между геополитическими сущностями (корпус NOW)[3];
- между сущностями в рамках корпуса MPQA[4] (F_1=0.36);
- к сущностям в сети Twitter[5];
- к аспектам[6];

---

[1] *Ana Marasovic* и др.: «... Discourse-Oriented Opinion Analysis», 2020 г.
[2] `https://tac.nist.gov/2014/KBP/Sentiment/index.html`
[3] ***Han* и др.: «No Permanent Friends or Enemies: Tracking...», 2018 г.**
[4] *Eunsol Choi* и др.: «Document-level sentiment inference with...», 2016 г.
[5] *Vo Duy-Ti* и др.: «Target-Dependent Twitter Sentiment Classification...», 2015 г.
[6] *Bagheri* и др.: «An unsupervised aspect detection model for sentiment...», 2013 г.

Извлечение оценочных отношений из текста

Текст: статья по международным отношениям (корпус RuSentRel)
Вершины: именованные сущности
Стрелки: оценочные отношения



Цель

- Разработать методы и программные средства для задачи извлечения оценочных *отношений* между именованными сущностями из текстов СМИ;
- Для достижения поставленной цели необходимо решить **задачи**:
  1. Исследовать методы машинного обучения для извлечения оценочных отношений между именованными сущностями из текстов СМИ;
  2. Исследовать методы порождения автоматически размеченных оценочных отношений на основе лексико-семантических ресурсов;
  3. Исследовать методы извлечения оценочных отношений на основе подхода опосредованного обучения[7] и комбинированной обучающей выборки (из ручной и автоматической разметки);
  4. Создать программные средства для обработки текстов СМИ, которые на основе текста статьи порождают список оценочных отношений между упомянутыми именованными сущностями.

[7] от англ. Distant Supervision

Извлечение оценочных отношений между именованными сущностями из текстов

- Дана текстовая коллекция C, состоящая из набора аналитических статей;
- Каждая статья включает: (1) документ D_i представленный последовательностью символов и (2) список упомянутых в документе 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, РФ, Российская Федерация), (Америка, США) ...

  задано отображение $\mathbf{G}$ для множества V всех вхождений именованных сущностей коллекции C на множество индексов групп G:

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

Постановка задачи II

- 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 (**a**), где l_j оценка пары (позитивная – POS, или негативная – NEG):

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

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

Классификация размеченных контекстов:

- **Контекст** – текстовый фрагмент, включающий две и более именованных сущностей предложения;
- **Размеченный контекст** – контекст с выделенной парой «субъект-объект» (a);

Выходная информация моделей:
3 класса $\{\text{POS},\text{NEG},\text{NEU}\}$
(извлечение оценочных отношений)

Обучение с учителем на размеченных контекстах:

1. Классические методы машинного обучения:
   - KNN, SVM, Naïve Bayes, Gradient Boosting
2. Сверточные и рекуррентные нейронные сети, сети с вниманием:
   - CNN, PCNN, LSTM, BiLSTM;
   - АттCNN$_e$, АттPCNN$_e$, IAN$_e$, BiLSTM, Атт-BLSTM;
3. Языковые модели:
   - mBERT, RuBERT, SentRuBERT.

Классификация контекстов: нейронные сети

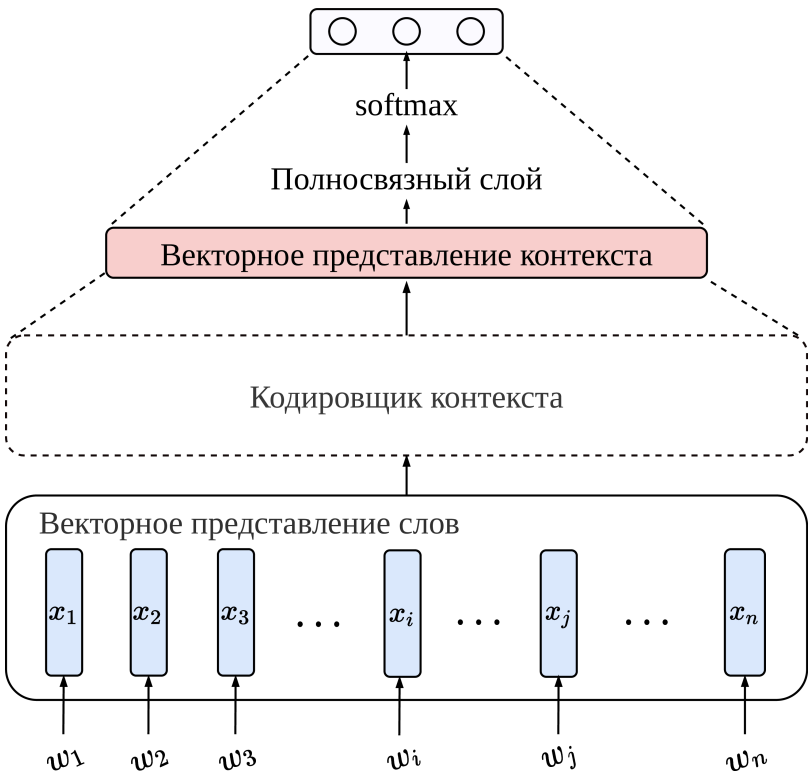


Рис.: Общая архитектура нейронных сетей.

Признаки векторного представления слов контекста:

1. Вектор терма из предобученной модели Word2Vec;
2. Вектор **расстояния** от терма контекста до каждого из участников пары a в отдельности;
3. Векторное представление **частей речи** слов контекста;
4. Признак вхождения терма в сторонний лексикон.

Оценка контекста: применение полносвязного слоя к векторному представлению контекста s:

$$o = W \cdot s + b$$

Классификация контекстов: языковые модели

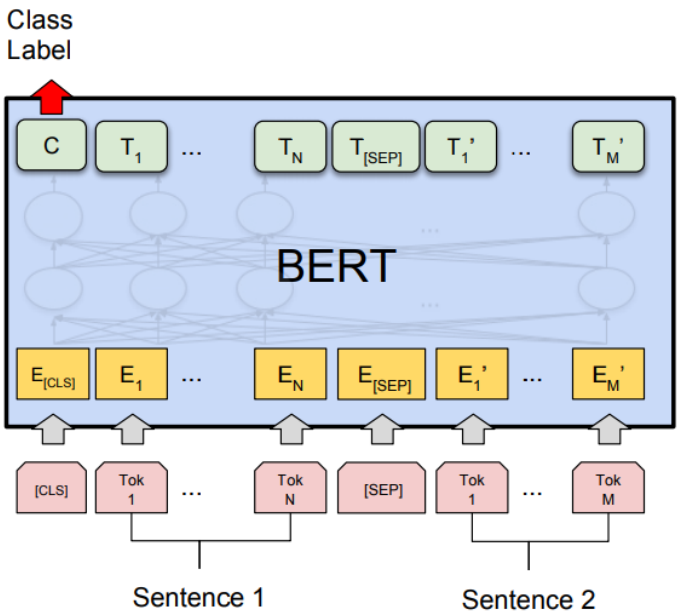


Рис.: Общая архитектура языковых моделей BERT [Девлин и др., 2018]

Значения последовательностей:

- TextA:
  - исходный контекст;
- TextB (опциально):
  - Вывод по контексту (NLI);
  - Вопрос;

Оценка контекста: применение полносвязного слоя к усредненной информации векторов

$$o = W \cdot s + b$$

$$s = \frac{\sum[C, T_1, ...T_N, T_{[sep]}, T'_1, ..., T'_M]}{N + M + 2}$$

где N, M – длины последовательностей TextA и TextB соответственно

Пример преобразования размеченных контекстов

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

↓

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

↓

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

Общие преобразования:

- Экранирование именованных сущностей.

- Исходная коллекция RuSentRel[8]: коллекция статей про международные отношения России;

Параметр (среднее)	TRAIN	TEST
Число документов	44	29
Предложений на документ	74.5	137
Сущностей на документ	194	300
Сущностей на документ (уникальных)	33.3	59.9
POS пар сущностей на документ	6.23	14.7
NEG пар сущностей на документ	9.33	15.6

- Статистика преобразованных контекстов:

Параметр	TRAIN	TEST
неразмеченных контекстов в среднем на документ	149	276

[8] https://github.com/nicolay-r/RuSentRel/tree/v1.1

Оценка отношений на уровне документа

Пусть $\mathbf{a}_c \subset \mathbf{a}$ – подмножество отношений с одинаковой оценкой $c \in \{\text{POS}, \text{NEG}\}$ для некоторого документа D.
Тогда при оценке $\mathbf{a}_c$ с эталонной разметкой, имеем:

- TP – число истинно-положительных отношений
- FP – число ложно-положительных отношений
- FN – число ложно-отрицательных отношений

1. Для класса c и документа D вычисляются: P (точность), R (полнота) и F_1-мера;
$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN} \qquad F_1 = \frac{2PR}{P + R}$$
2. $F^c_{1-macro}$ – макроусреднение по F_1 каждого из документов для c;
3. Фиксируется F^{PN}_{1-mean} учитывающее оценки по POS и NEG:
$$F^{PN}_{1-mean} = \frac{(F^P_{1-macro} + F^N_{1-macro})}{2} \cdot 100$$

Оценка результата:

- $F1^a_{cv}$ – усредненный результат **кросс-валидационного**[a] тестирования (CV-3);
- $F1_t$ – результат на TEST множестве.

‡ – подбор наилучших параметров по сетке.

Согласие экспертов: $F^{PN}_{1-mean} = 55.0$

Модель	$F1^a_{cv}$	$F1_t$
RuBERT (—)	**36.8**	37.6
RuBERT (вопрос)	32.0	35.3
RuBERT (NLI)	29.4	**39.6**
CNN	28.7	31.4
PCNN	**29.6**	**32.5**
LSTM	27.9	31.6
BiLSTM	<u>28.6</u>	32.4
AttCNN$_e$	27.6	29.7
AttPCNN$_e$	<u>29.9</u>	**32.6**
IAN$_e$	**30.8**	32.2
Att-BLSTM	27.5	<u>32.3</u>
Gradient boosting ‡	**20.3**	**28.0**

[a]Тестирование вводится ввиду разного объема документов множеств TRAIN и TEST

[9]https://github.com/nicolay-r/RuSentRel-Leaderboard

Методы опосредованного обучения в задаче извлечения оценочных отношений

Опосредованное обучение — автоматическая разметка обучающих данных большого объема за счет доп. ресурсов;
даны:

1. размеченная коллекция статей статьи представлены: документом (D_i), сущностями (E_i), списком отношений $(\mathbf{a}_i)$:
$$C = \left[\langle D_i, E_i, \mathbf{a}_i \rangle \right]_{i=1}^{|C|}$$
2. *модель извлечения оценочных отношений* (M) – отображение пары $\langle D_i, E_i \rangle$ в список оценочных отношений:
$$M : \langle D_i, E_i \rangle \to \hat{\mathbf{a}}_i$$
3. коллекция **неразмеченных** текстов СМИ:
$$N = \{D'_1, \ldots, D'_{|N|}\}$$

Опосредованное обучение модели M – итеративный процесс оптимизации параметров с применением алгоритма-посредника $\mathcal{A}$ для разметки оценочных отношений в $D'_i \in N$:

$$\mathcal{A}(D'_i) = \langle E'_i, \mathbf{a}_i \rangle$$

Необходимо разработать алгоритм $\mathcal{A}$ разметки оценочных отношений в N для организации опосредованного обучения M.

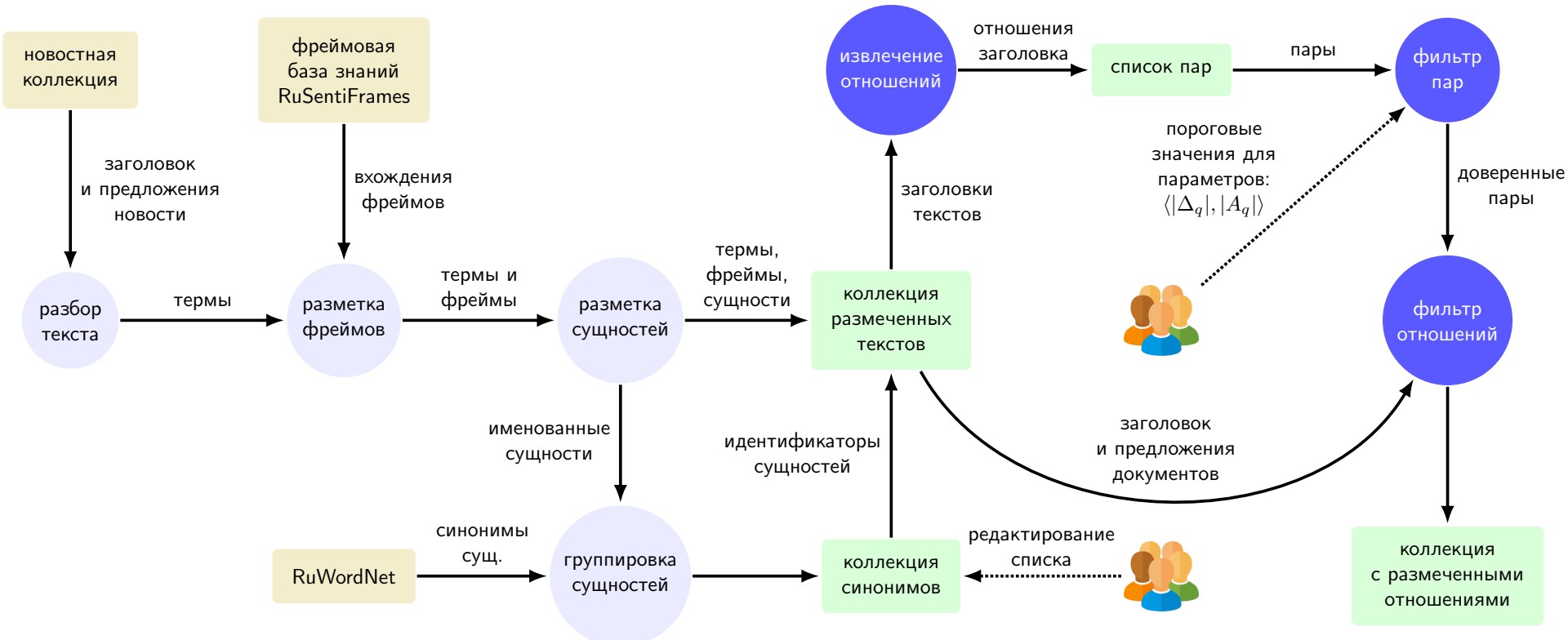
Используемые ресурсы для обработки документов

1. **Новостная коллекция**: русскоязычные статьи и новости крупных новостных источников (8.8 млн. новостей);
2. **База знаний RuSentiFrames**[10]: Описывает оценочную ассоциацию, передаваемую *предикатом* в формате глагола или сущ. (311 фреймов, связанных с 7034 лексическими единицами)
   - **Роли**: A0 (агент), A1 (тема);
   - **Размерности**: отношение автора текста к упомянутым участникам; **polarity** – оценка между участниками; **effects** – эффект воздействия на участников; психическое состояние (**states**) участников к описанной ситуации.

Фрейм «хвалиться»	Описание
ВХОЖДЕНИЯ	хвалиться, хвастовство, похвалиться
ROLES	A0: тот, кто хвалится A1: то, чем хвалятся
POLARITY	A0→A1, POS author→A0, NEG
EFFECT	A1, POS
STATE	A0, POS

[10] https://github.com/nicolay-r/RuSentiFrames

Архитектура системы обработки документов



Заполнение списка пар

1. Из заголовка D'_i извлекаются варианты фреймов RuSentiFrames:

$$F_i = \left[f_1, ..., f_{|F|} \right] \qquad f = \langle c_b, c_e, v, l_{\texttt{A0}\to\texttt{A1}} \rangle$$

   - $l_{\texttt{A0}\to\texttt{A1}}$ **инвертируется** если перед v присутствует частица «не»;
   - $l_{\texttt{A0}\to\texttt{A1}} = \text{NEU}$, если A0→A1 не определено;

2. Из E'_i составляется множество **пар** $\{\langle e'_i, e'_j \rangle\}$:
   - e'_i упомянута раньше e'_j;
   - Типы, как участников (t_i, t_j), так и всех сущностей между ними принадлежат $O_{valid} = \{\texttt{GPE}, \texttt{ORG}, \texttt{PER}\}$;
   - Все фреймы, входящие между участниками отношений $l_{\texttt{A0}\to\texttt{A1}} \in \{\text{POS}, \text{NEG}\}$.
   - Перед участниками отношений отсутствуют предлоги «в»/«на»;

3. Заполняется список пар:
   - Достоверные пары заголовка отправляется в список пар в формате кортежа: $a = \langle g_i, g_j, \mathbf{l} \rangle \qquad g_i \neq g_j \qquad g_i, g_j \in G$
   - $\mathbf{l}$ назначается следующим образом: POS (все $\{f_k\}$ внутри пары имеют POS оценку для A0→A1); NEG (иначе).

- Пусть $A = \{a_1, a_2, \ldots a_{|A|}\}$ множество достоверных пар, извлеченных в результате анализа всей коллекции N;
- Тогда, для некоторой пары $\langle g_i, g_j \rangle$ составим A_{ij} – подмножество соответствующих достоверных отношений A;
- **Ориентация пары** к классу $c \in \{\text{POS}, \text{NEG}\}$:

$$p(\langle g_i, g_j \rangle | c) = \frac{|\{\langle g_i, g_j, \mathbf{l} \rangle | \mathbf{l} = c\}|}{|A_{ij}|}$$

- **Наиболее оценочно ориентированные пары**: ограничения по $|A_{ij}|$ и $\Delta_A = p(A|\text{POS}) - p(A|\text{NEG})$;

A0	A1	Δ_q
МВД России	Российская Федерация	1.00
Путин	Министерство Внутренних дел	0.91
Канада	Украина	0.90
Израиль	ООН	-0.85
Азербайджан	Армения	-0.93
Карабах	Азербайджан	-0.94
ЕС	Siemens	-1.00

- **Извлечено**: 2372 пары (1%) из 247876; $\langle |\Delta_q| \geq 0.3, |A_q| \geq 25 \rangle$

Извлечение отношений из заголовка и текста новостей I

Нагорный Карабах$_e$ обвинил Азербайджан$_e$ в 200 обстрелах за неделю

1. Извлеченное из заголовка отношение $\langle g_i, g_j, 1 \rangle$ сопоставляется с аналогичной (доверенной) парой $\langle g_i, g_j \rangle$ из списка синонимов.
   - Отбираются только такие отношения, которые совпадают с аналогичными парами; т.е. 1 соответствует оценочной ориентации пары, определяемой знаком Δ_A: POS $(\Delta_A > 0)$, NEG $(\Delta_A < 0)$.
2. Извлеченное отношение $\langle g_i, g_j, 1 \rangle$ передается на этап **фильтрации предложений** для поиска таких же пар в предложениях новости.

RuAttitudes	Новостей	**134442**
	Отношений на новость	2.26
	Предложений на новость	0.88

Извлечение отношений из заголовка и текста новостей II

Заголовок
Тиллерсон$_e$: США$_e$ не снимут ***санкции***$_{neg}$ с РФ$_e$ до возвращения Крыма$_e$

$\downarrow$ сша$\rightarrow$россия$_{neg}$, сша$\rightarrow$крым$_{neg}$

Наиболее оценочно ориентированные пары[11]	
Запрос	Результат поиска
сша$\rightarrow$россия$_{neg}$	пара найдена, оценки совпадают; POS: 32%, NEG: 68%
сша$\rightarrow$крым$_{neg}$	пара не найдена

$\downarrow$ США$\rightarrow$РФ$_{neg}$

Предложение
Госсекретарь США$_e$ Рекс Тиллерсон$_e$, выступая в Брюсселе$_e$ на встрече глав МИД$_e$, входящих в входящих в состав НАТО$_e$, заявил, что санкции с России$_e$ будут сняты только после возвращения Крыма$_e$, сообщает CNN$_e$.

[11] $\langle |\Delta_q| \geq 0.3, |A_q| \geq 25 \rangle$

Этап разметки нейтральных отношений I

Пара $\langle e'_1, e'_2 \rangle$ заголовка или предложения считается **нейтральной**:

Объект отношения (e'_2) не является столицей/страной, и при этом тип объекта $t_2 = \texttt{LOC}$;

Дополнительные ограничения:

1. Типы, как участников (t_1, t_2), так и всех сущностей между ними принадлежат $\{\texttt{ORG}, \texttt{PER}, \texttt{GPE}\}$;
2. e'_1 упомянута раньше e'_2, участники e'_1 и e'_2 не принадлежат одной синонимичной группе, а также $\langle e'_1, e'_2 \rangle$ и $\langle e'_2, e'_1 \rangle$ не являются оценочными.

A0	A1	Вхождений
КНДР	корейский полуостров	301
Россия	ближний восток	232
США	Баренцево море	204
Япония	Курилы	172
Волгоград	река волга	155

Форматы проведения:

- 2 класса $\{\text{POS,NEG}\}$ – классификация оценочных отношений
- 3 класса $\{\text{POS,NEG,NEU}\}$ – извлечение оценочных отношений

Форматы обучения моделей:

- Предварительное – RuAttitudes $\rightarrow$ RuSentRel;
- Объединенное[12] – RuSentRel $\cup$ RuAttitudes;

Для результатов моделей, обученных с применением опосредованного обучения ($F1_a$) и обучения с учителем ($F1_b$) вычисляется коэффициент прироста качества (Δ) вычисляется по формуле:

$$\Delta(F1) = \left(\frac{F1_a}{F1_b} - 1\right) \cdot 100$$

[12] Для сверточных и рекуррентных нейронных сетей

Результаты (Нейронные сети)

Модель	RuAttitudes	Объединенное обучение			
		2 класса		3 класса	
		$F1^a_{cv}$	$F1_t$	$F1^a_{cv}$	$F1_t$
CNN	✓	**70.0**	**74.3**	**32.8**	39.6
CNN	—	63.6	65.9	28.7	31.4
PCNN	✓	**69.5**	70.5	31.6	**39.7**
PCNN	—	64.4	63.3	29.6	32.5
LSTM	✓	**68.0**	**75.4**	31.6	39.5
LSTM	—	61.9	65.3	27.9	31.6
BiLSTM	✓	**71.2**	68.4	32.0	38.8
BiLSTM	—	62.3	**71.2**	28.6	32.4
АттCNN$_e$	✓	66.8	72.7	30.9	39.9
АттCNN$_e$	—	65.0	66.2	27.6	29.7
АттPCNN$_e$	✓	**70.2**	67.8	**32.2**	39.9
АттPCNN$_e$	—	64.3	63.3	29.9	32.6
IAN$_e$	✓	69.1	**72.6**	30.7	36.7
IAN$_e$	—	60.8	63.5	30.8	32.2
Att-BLSTM	✓	66.2	71.2	31.0	37.3
Att-BLSTM	—	65.7	68.2	27.5	32.3
Среднее-$\Delta(F1)$	✓	**+8.7%**	**+8.9%**	**+10.6%**	**+23.4%**
Среднее	—	63.5	65.9	28.8	31.8

Результаты (Языковые модели)

		2 класса		3 класса	
Модель	RuAttitudes	$F1^a_{cv}$	$F1_t$	$F1^a_{cv}$	$F1_t$
mBERT (NLI_P + без TextB)	✓	68.9	67.7	**30.5**	31.1
mBERT (без TextB)	—	67.0	68.9	26.9	30.0
mBERT (NLI_P + TextB_{QA})	✓	69.6	65.2	**30.1**	**35.5**
mBERT (TextB_{QA})	—	66.5	65.4	28.6	33.8
mBERT (NLI_P + $\text{TextB}_{\text{NLI}}$)	✓	69.4	68.2	**33.6**	36.0
mBERT ($\text{TextB}_{\text{NLI}}$)	—	67.8	58.4	29.2	37.0
RuBERT (NLI_P + без TextB)	✓	**70.0**	69.8	35.6	35.4
RuBERT (без TextB)	—	67.8	66.2	**36.8**	37.6
RuBERT (NLI_P + TextB_{QA})	✓	**69.6**	68.2	34.8	37.0
RuBERT (TextB_{QA})	—	69.5	66.2	32.0	35.3
RuBERT (NLI_P + $\text{TextB}_{\text{NLI}}$)	✓	**71.0**	68.6	**36.8**	39.9
RuBERT ($\text{TextB}_{\text{NLI}}$)	—	68.9	66.4	29.4	39.6
SentRuBERT (NLI_P + без TextB)	✓	70.0	69.8	37.9	**39.8**
SentRuBERT (без TextB)	—	69.3	65.5	34.0	35.2
SentRuBERT (NLI_P + TextB_{QA})	✓	69.6	64.2	**38.4**	**41.9**
SentRuBERT (TextB_{QA})	—	70.2	67.1	34.3	38.9
SentRuBERT (NLI_P + $\text{TextB}_{\text{NLI}}$)	✓	70.2	67.7	**39.0**	38.0
SentRuBERT ($\text{TextB}_{\text{NLI}}$)	—	69.8	67.6	33.4	32.7
Среднее-$\Delta(F1)$	✓	+1.8%	+3.7%	**+13.5%**	+10.0%
Среднее	—	68.5	65.7	31.6	35.6

Визуализация весов внимания (нейронные сети)

Теплокарта назначения весов при различных форматах обучения:

Атт-BLSTM (Обучение с учителем)
вести такую игра , $\underline{\text{E}}_{subj}$ окончательно ***лишаться***$_{pos}$ ***доверие***$_{pos}$ $\underline{\text{E}}_{obj}$ и страна E
...
Но $\underline{\text{E}}_{subj}$ последовательно подчеркивать свой ***интерес***$_{pos}$ к ***нормализация***$_{pos}$ отношение с $\underline{\text{E}}_{obj}$ (<NUM> февраль <NUM> г . состояться визит E в E и его ***переговоры***$_{pos}$ с духовный лидер E и с президент E)
Атт-BLSTM (Применение опосредованного обучения)
вести такую игра , $\underline{\text{E}}_{subj}$ окончательно ***лишаться***$_{pos}$ ***доверие***$_{pos}$ $\underline{\text{E}}_{obj}$ и страна E
...
Но $\underline{\text{E}}_{subj}$ последовательно подчеркивать свой ***интерес***$_{pos}$ к ***нормализация***$_{pos}$ отношение с $\underline{\text{E}}_{obj}$ (<NUM> февраль <NUM> г . состояться визит E в E и его ***переговоры***$_{pos}$ с духовный лидер E и с президент E)

При использовании опосредованного обучения наблюдается особенность применения заголовков в обучении моделей:

$$\ldots \quad \text{Subject}_e \quad \ldots \quad \{frame_{A0 \rightarrow A1}\}_k \quad \ldots \quad \text{Object}_e \quad \ldots$$

Ядро

AREkit – набор инструментов[13] для разбора документов с аннотацией (1) именованных сущностей, (2) фреймов и (3) отношений на контекстном и документном уровнях

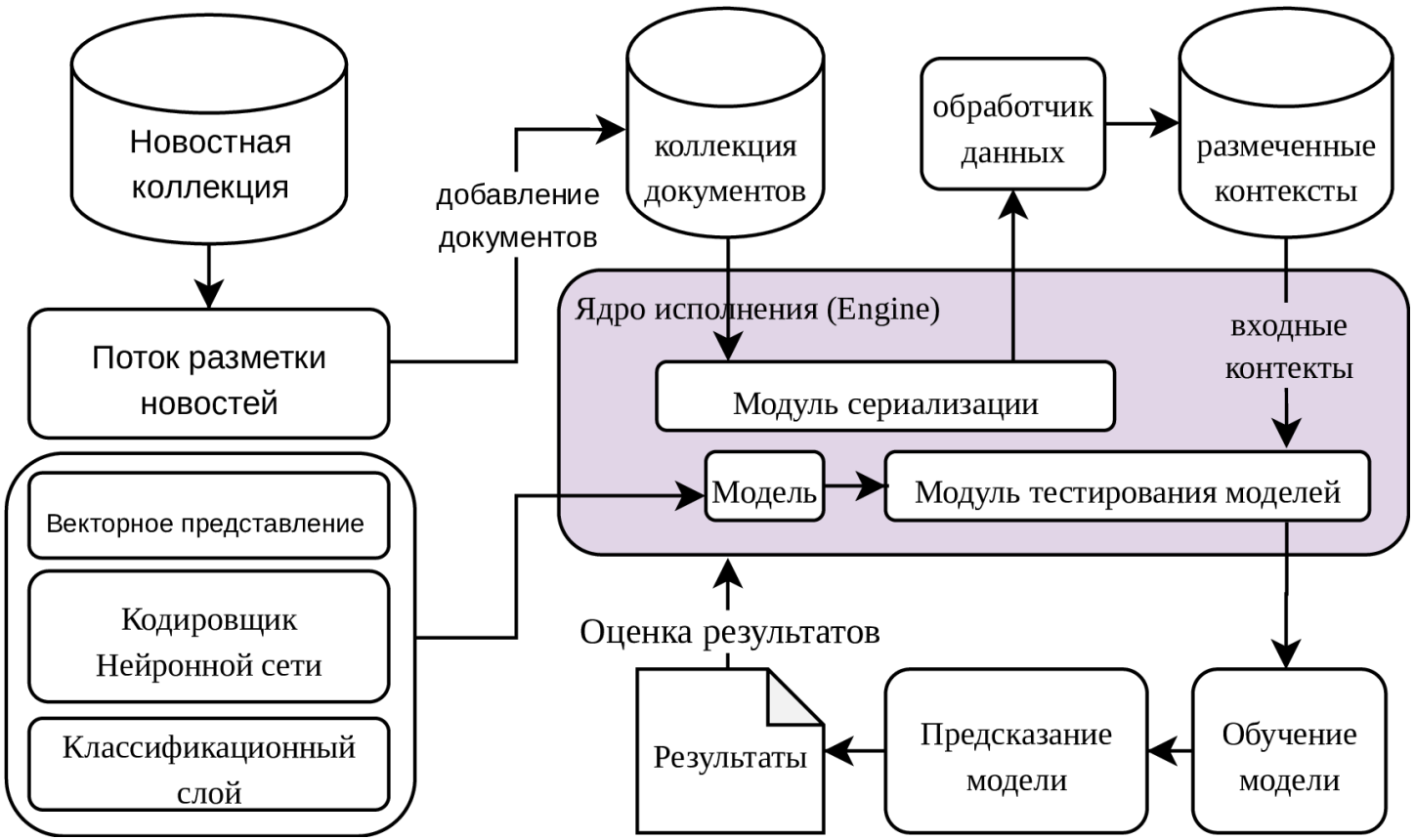
Приложения

Потока обработки новостных текстов[14]

Приложения для Deep Learning экспериментов в рамках корпуса RuSentRel[15,16];

[13] https://github.com/nicolay-r/AREkit
[14] https://github.com/nicolay-r/frame-based-attitude-extraction-workflow/tree/v2.0
[15] https://github.com/nicolay-r/neural-networks-for-attitude-extraction/tree/0.20.5
[16] https://github.com/nicolay-r/bert-for-attitude-extraction-with-ds/

Общая архитектура программного комплекса



Результаты работы

- Дана постановка задачи извлечения оценочных отношений на уровне документов. На основе корпуса RuSentRel были исследованы подходы к извлечению оценочных отношений на основе нейронных сетей с механизмом внимания, а также языковых моделей.
- Предложен метод автоматического порождения обучающей коллекции на основе оценочных фреймов лексикона RuSentiFrames. Применение опосредованного обучения повысило качество языковых моделей на 10-13% и на 25% при сравнении с наилучшими результатами моделей на основе нейронных сетей.

Научная новизна:

- Предложена структура фреймовой базы знаний RuSentiFrames для описания тональностей, ассоциирующихся со словами и выражениями русского языка, включая тональность отношений между участниками ситуации, отношение автора к участникам ситуации, позитивные и негативные эффекты, связанные с ситуацией. Такая база знаний описывает значительно более сложную структуру тональностей, ассоциированных с словом, в отличие от обычных списков оценочных слов с оценками тональностей п. паспорта: 4
- Впервые для русского языка поставлена задача и выполнено исследование методов извлечения тональности отношений между именованными сущностями, упомянутыми в текстах СМИ

Научная и практическая значимость II

- Для обучения моделей извлечения оценочных отношений предложен новый метод автоматического порождения обучающей коллекции на основе оценочных фреймов нового лексикона RuSentiFrames и использования структуры новостных текстов. Применение опосредованного обучения с использованием RuAttitudes-2.0 повысило качество языковых моделей BERT на 10-13% по метрике F1, и на 25% при сравнении с наилучшими результатами остальных моделей на основе оценочных фреймов нового лексикона RuSentiFrames и структуры новостных текстов п. паспорта: 4, 7

Практическая значимость. Разработаны модели извлечения оценочных отношений, а также методы автоматической обработки внешних новостных источников информации. Предложенные методы извлечения оценочных отношений могут быть использованы при решении прикладных задач текстового анализа: составление статистической оценки новостного источника, глубокого понимания текста с точки зрения установления связей между упомянутыми объектами.

Основные публикации I

Основные результаты по теме диссертации изложены в 13 печатных изданиях, 2 из которых изданы в журналах, рекомендованных ВАК РФ, 8 – в изданиях, индексируемых в системах Web of Science и Scopus, 1 – в тезисах докладов. Зарегистрирован 1 патент.

1. Николай Русначенко. — «Программный комплекс для извлечения оценочных отношений между именованными сущностями из коллекций новостных документов». — Номер и дата поступления заявки: 2021662348 08.08.2021, Дата регистрации: 13.08.2021 ПАТЕНТ
2. Nicolay Rusnachenko и Natalia Loukachevitch. — «Studying Attention Models in Sentiment Attitude Extraction Task». — В: Proceedings of the 25th International Conference on Natural Language and Information Systems. Lecture Notes in Computer Science (LNCS), vol. 12089, pp. 157-169, — 2020. SCOPUS ЖУРНАЛ
3. Nicolay Rusnachenko и Natalia Loukachevitch. — «Neural Network Approach for Extracting Aggregated Opinions from Analytical Articles». — В: International Conference on Data Analytics and Management in Data Intensive Domains. Communications in Computer and Information Science (CCIS), vol. 1003, pp. 167-179, — 2018. SCOPUS ЖУРНАЛ
4. Nicolay Rusnachenko и Natalia Loukachevitch. — «Sentiment Attitudes and Their Extraction from Analytical Texts». — В: International Conference on Text, Speech, and Dialogue. Lecture Notes in Computer Science (LNCS), vol. 11107, pp. 41-49, — 2018. SCOPUS ЖУРНАЛ

Основные публикации II

5. Nicolay Rusnachenko, Natalia Loukachevitch и Elena Tutubalina. — «Distant Supervision for Sentiment Attitude Extraction». — В: Proceedings of Recent Advances in Natural Language Processing Conference. pp. 1022–1030, — 2019. SCOPUS
6. Nicolay Rusnachenko и Natalia Loukachevitch. — «Extracting Sentiment Attitudes from Analytical Texts via Piecewise Convolutional Neural Network». — В: In Proceedings of CEUR Workshop, DAMDID-2018 Conference. pp. 186-192, — 2018. SCOPUS
7. Nicolay Rusnachenko и Natalia Loukachevitch. — «Attention-Based Neural Networks for Sentiment Attitude Extraction using Distant Supervision». — В: The 10th International Conference on Web Intelligence, Mining and Semantics (WIMS 2020), pp. 159–168, June 30-July 3, 2020, Biarritz, France. — 2020. SCOPUS
8. 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). pp. 459-468, — 2018. SCOPUS
9. Natalia Loukachevitch и Nicolay Rusnachenko – «Sentiment Frames for Attitude Extraction in Russian» — В: Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialogue-2020 — 2019. SCOPUS

Основные публикации III

10. Николай Леонидович Русначенко. — «Применение языковых моделей в задаче извлечения оценочных отношений.» — В: Труды Института системного программирования РАН;33(3): с. 199-222, — 2021. ВАК ЖУРНАЛ
11. Николай Леонидович Русначенко и Наталья Валентиновна Лукашевич. — «Методы интеграции лексиконов в машинное обучение для систем анализа тональности». — В: Искусственный интеллект и принятие решений. с. 78–89, — 2017. ВАК ЖУРНАЛ
12. Nicolay Rusnachenko и Natalia Loukachevitch. — «Using convolutional neural networks for sentiment attitude extraction from analytical texts». — В: EPiC Series in Language and Linguistics. pp. 1-10, — 2019.
13. Лукашевич Н.В., Карнаухова В.А., Русначенко Н.Л. – «Автоматическое извлечение имплицитных оценок из текстов» – В: Труды международной конференции по компьютерной и когнитивной лингвистике TEL-2018 С. 169-179, – 2018.

Поддержка выполненных работ

Работы поддержаны тремя грантами РФФИ:

- **16-29-09606 (офи_м)**: Автоматические методы выявления среды распространения терроризма и экстремизма в социальных сетях;
- **19-37-50001 (мол_нр)**: Автоматическое порождение обучающей коллекции из русскоязычных новостных источников и социальных сетей для извлечения оценочных отношений из аналитических текстов личный;
- **20-07-01059 (А)**: Автоматический анализ тональности текстов с множественными оценками на основе оценочных фреймов.

Регистрация программы ЭВМ[17]

РОССИЙСКАЯ ФЕДЕРАЦИЯ

RU **2021663268**

ФЕДЕРАЛЬНАЯ СЛУЖБА
ПО ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ

(12) **ГОСУДАРСТВЕННАЯ РЕГИСТРАЦИЯ ПРОГРАММЫ ДЛЯ ЭВМ**

Номер регистрации (свидетельства):
2021663268

Дата регистрации: **13.08.2021**

Номер и дата поступления заявки:
2021662348 08.08.2021

Дата публикации: **13.08.2021**

Контактные реквизиты:
89162429919, kolyarus@yandex.ru

Автор:
Русначенко Николай Леонидович (RU)

Правообладатель:
Русначенко Николай Леонидович (RU)

Название программы для ЭВМ:
Программный комплекс для извлечения оценочных отношений между именованными сущностями из коллекций новостных документов

Реферат:
Программа предназначена для обработки новостных и аналитических текстов. Может использоваться в качестве потока/модуля (workflow) программного-аппаратного комплекса обработки текстовой информации новостного характера. Функциональные возможности программы: возможность извлечения оценочных отношений между именованными из коллекции аналитических текстов с размеченными именованными сущностями; аннотация новостных текстов с выполнением разметки именованных сущностей и извлечением оценочных отношений посредством использования лексикона RuSentiFrames; возможность применения методов машинного обучения на основе нейронных сетей и языковых моделей BERT к коллекциям аналитических текстов. Тип ЭВМ: ПК. ОС: Ubuntu Linux 18.0.4.

Язык программирования: Python

Объем программы для ЭВМ: 13,2 МБ

[17] https://new.fips.ru/registers-doc-view/fips_servlet?DB=EVM&DocNumber=2021663268&TypeFile=html

Участие в конференциях

- Международная Конференция «Диалог» (Россия, Москва, РГГУ, 2018);
- 20-ая Международная Конференция Data Analytics and Management in Data Intensive Domains (Россия, Москва, МГУ, 2018);
- 21-ая Международная Конференция Text-Speech-Dialog (Чехия, Брно, 2020);
- 12-ая Международная Конференция Recent Advances in Natural Language Processing (Болгария, Варна, 2020);
- 25-ая Международная Конференция Natural Language & Information Systems (Германия, Саарбрюккен, 2020);
- 10-ая Международная Конференция Web Intelligence, Mining and Semantics (Франция, Биаритц, 2020);

Соответствие паспорту специальности 05.13.11

Направление	**Результат работы**
4. Системы управления базами данных и знаний	Впервые для русского языка поставлена задача и выполнено исследование методов извлечения тональности отношений между именованными сущностями, упомянутыми в текстах СМИ
7. Человеко-машинные интерфейсы; модели, методы, алгоритмы и программные средства машинной графики, визуализации, обработки изображений, систем виртуальной реальности, мультимедийного общения	Предложен и реализован новый метод автоматического порождения обучающей коллекции для классификации оценочных отношений по двум и трем классам на основе словаря оценочных фреймов RuSentiFrames; человек-оператор может вручную управлять параметрами отбора оценочных отношений, и таким образом управлять процессом автоматической разметки

Спасибо за внимание!

Замечания

1. Корпус RuSentRel рассматривается как сторонний ресурс, детали построения не рассматриваются в тексте.
2. Полная картина результатов:

   `https://github.com/nicolay-r/RuSentRel-Leaderboard`

AREight – приложение обработки текстов СМИ в рамках задачи извлечения отношений[18]
Зависимости: Python, AREkit, DeepPavlov, brat

`Docker + NVidia docker`:

- Модели NER ($\text{BERT}_{\text{ontonotes}}$) и BERT.
- Apache, CGI

Визуализация:

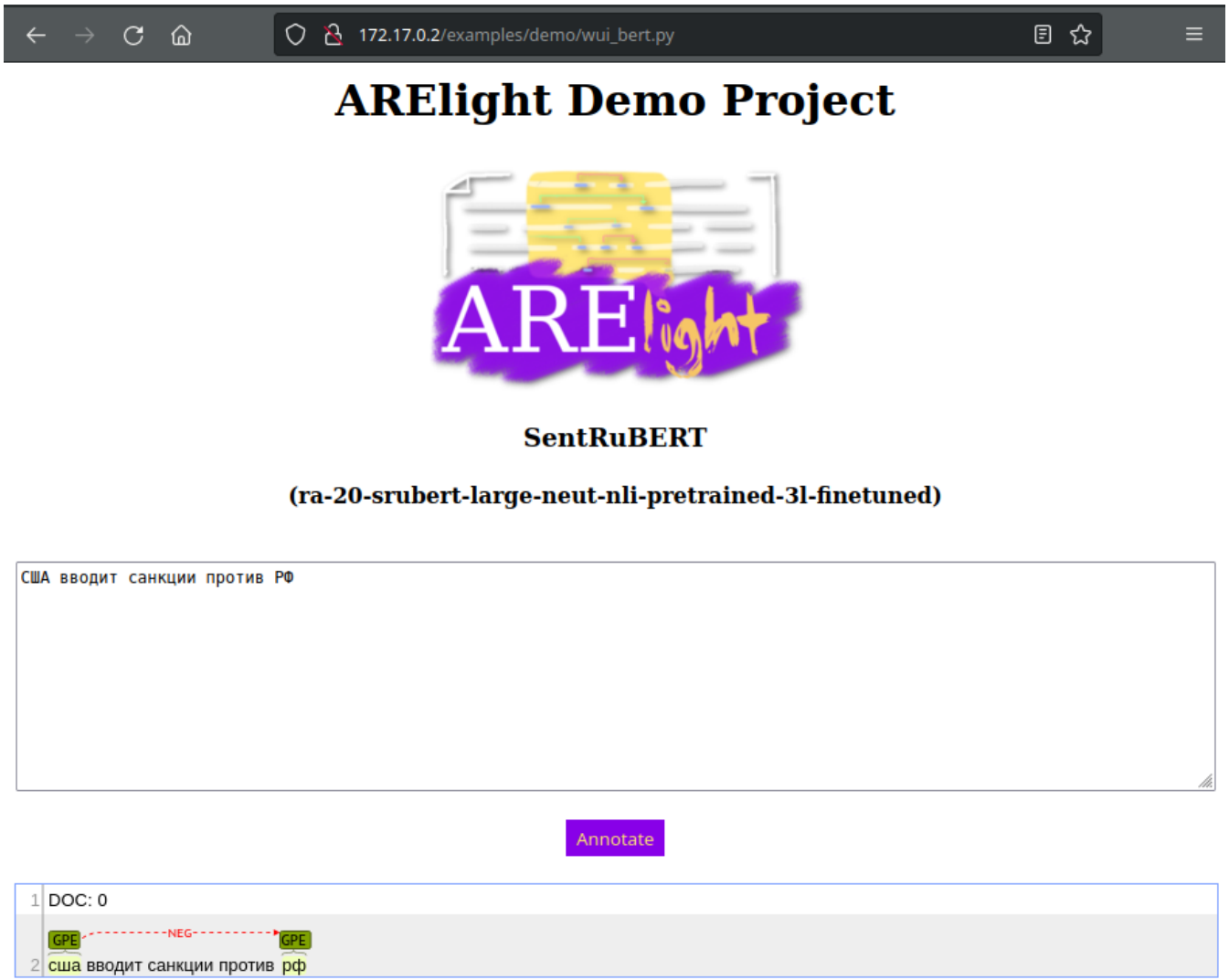
- Brat

Технические требования:

- 12 Гб ОЗУ, 6 Гб VRAM (NVidia GTX 1060 TI +), CUDA

[18] `https://nicolay-r.github.io/arelight-page/`

Шаблон демо проекта извлечения оценочных отношений из текста новостного документа



Основные положения выносимые на защиту

1. Предложена структура фреймовой базы знаний RuSentiFrames для описания тональностей, ассоциирующихся со словами и выражениями русского языка, включая тональность отношений между участниками ситуации, отношение автора к участникам ситуации, позитивные и негативные эффекты, связанные с ситуацией
2. Предложен и реализован новый метод автоматического порождения обучающей коллекции для классификации оценочных отношений по двум и трем классам на основе словаря оценочных фреймов RuSentiFrames
3. Программный комплекс AREkit для создания автоматически размеченной обучающей коллекции для извлечения оценочных отношений, с программным интерфейсом для задания настроек пользователем, а также обучения методов на основе нейронных сетей

Обработка размеченного контекста

Все слова разделены на *группы термов*:

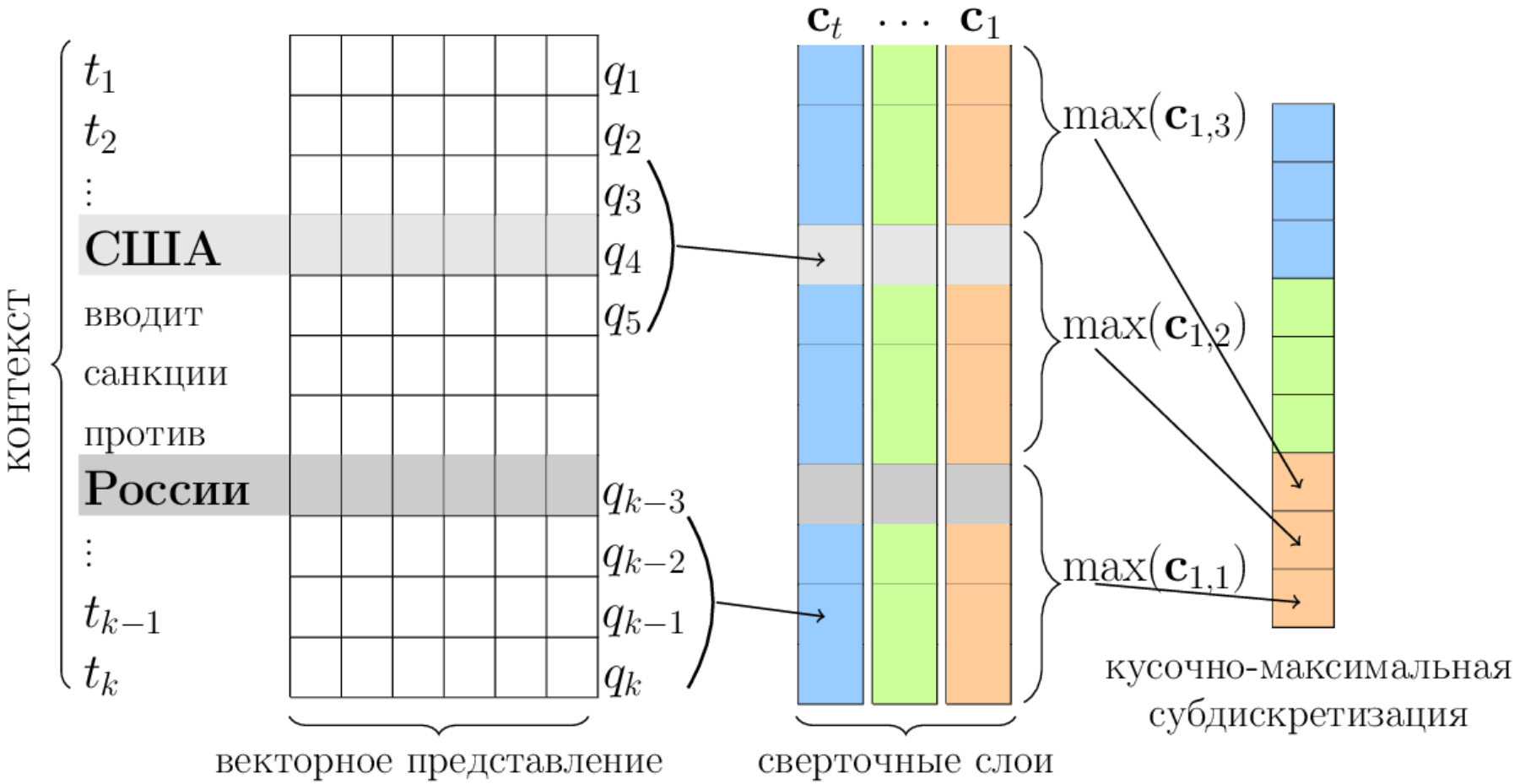
- ENTITIES – именованные сущности ($\underline{\text{E}}_e$), участники пары субъект-объект ($\underline{\text{E}}_{subj}$, $\underline{\text{E}}_{obj}$);
- TOKENS – знаки препинания (<COMMA>, <DOT>), цифры (<NUMBER>), URL-ссылки, и т.д.; полный список составляет 17 вхождений;
- WORDS – прочие слова контекста, представленные в **лемматизированной форме** и нижнем регистре.

Размеченный контекст
«Президент $\underline{\text{Путин}}_{obj}$ несколько раз намекал на это в интервью»: – вице-премьер $\underline{\text{Дмитрий Рогозин}}_{subj}$ написал это в статье в 2012 году.

↓

Термы входного контекста
<QUOTE> $\underline{\text{E}}_{obj}$ несколько раз намекать на это в интервью <QUOTE> <COLON> <DASH> вице-премьер $\underline{\text{E}}_{subj}$ написать это в статья в <NUMBER> год <DOT>

Архитектура кодировщика модели PCNN



Кусочно-максимальная субдискретизация

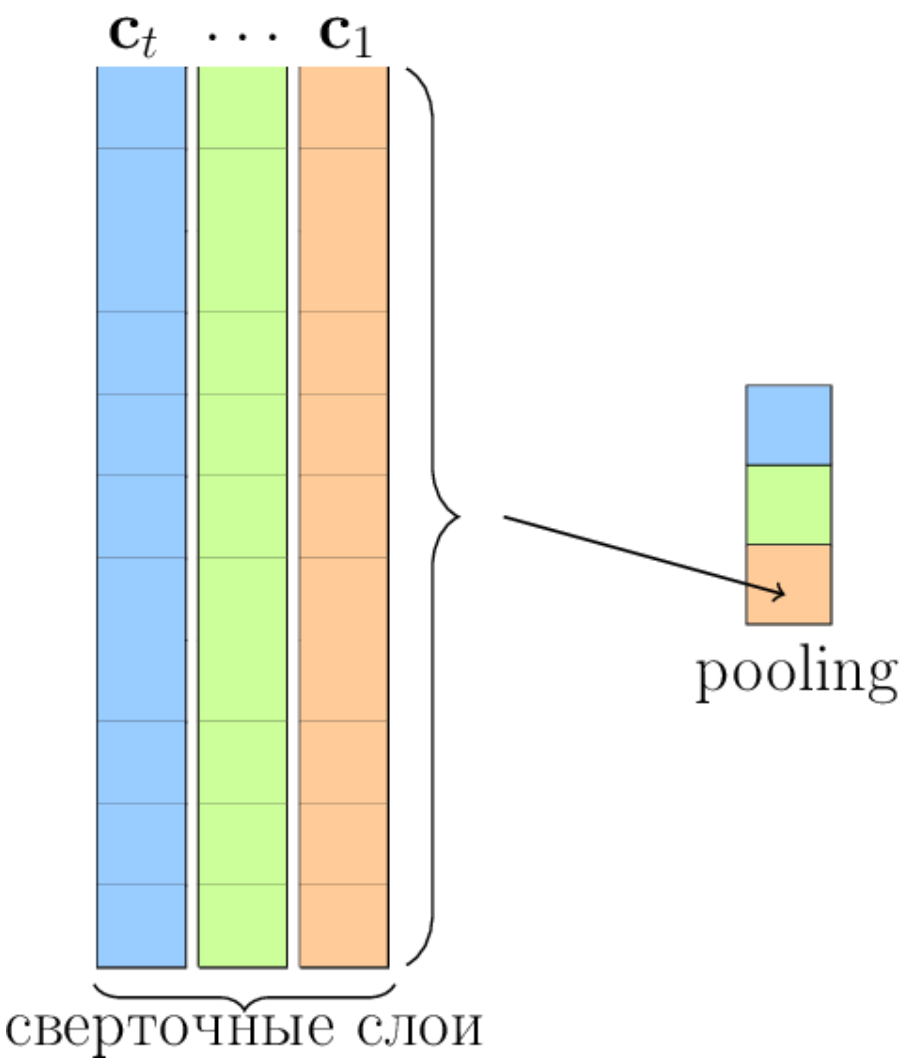


Рис.: Классическая архитектура CNN (свертка целого размеченного контекста)

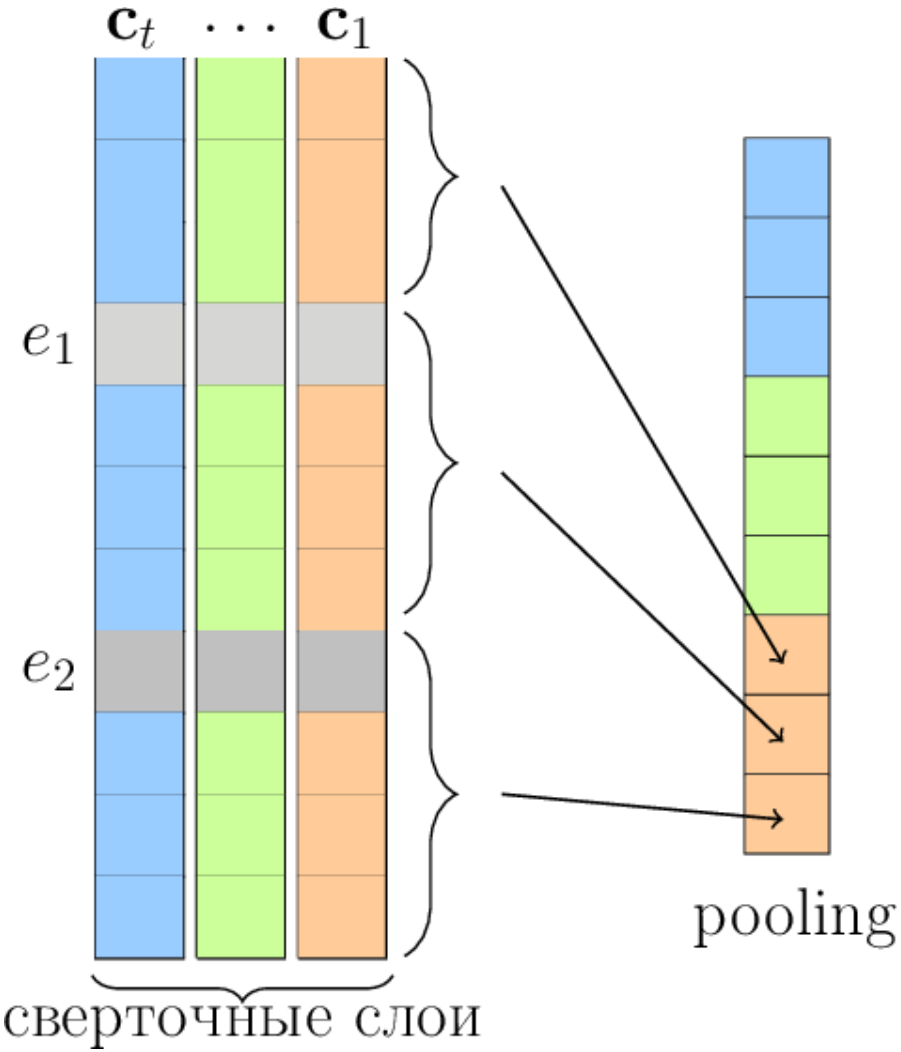


Рис.: Специализированная под извлечение отношений архитектура PCNN

Операция свертки векторного представления размеченного контекста

Длина окна:

$$\omega = 3$$

Размер вектора:

$$m = 4 + 2 = 6$$

Сверточный фильтр:

$$\mathbf{w} \in \mathbb{R}^{\omega \cdot m}$$

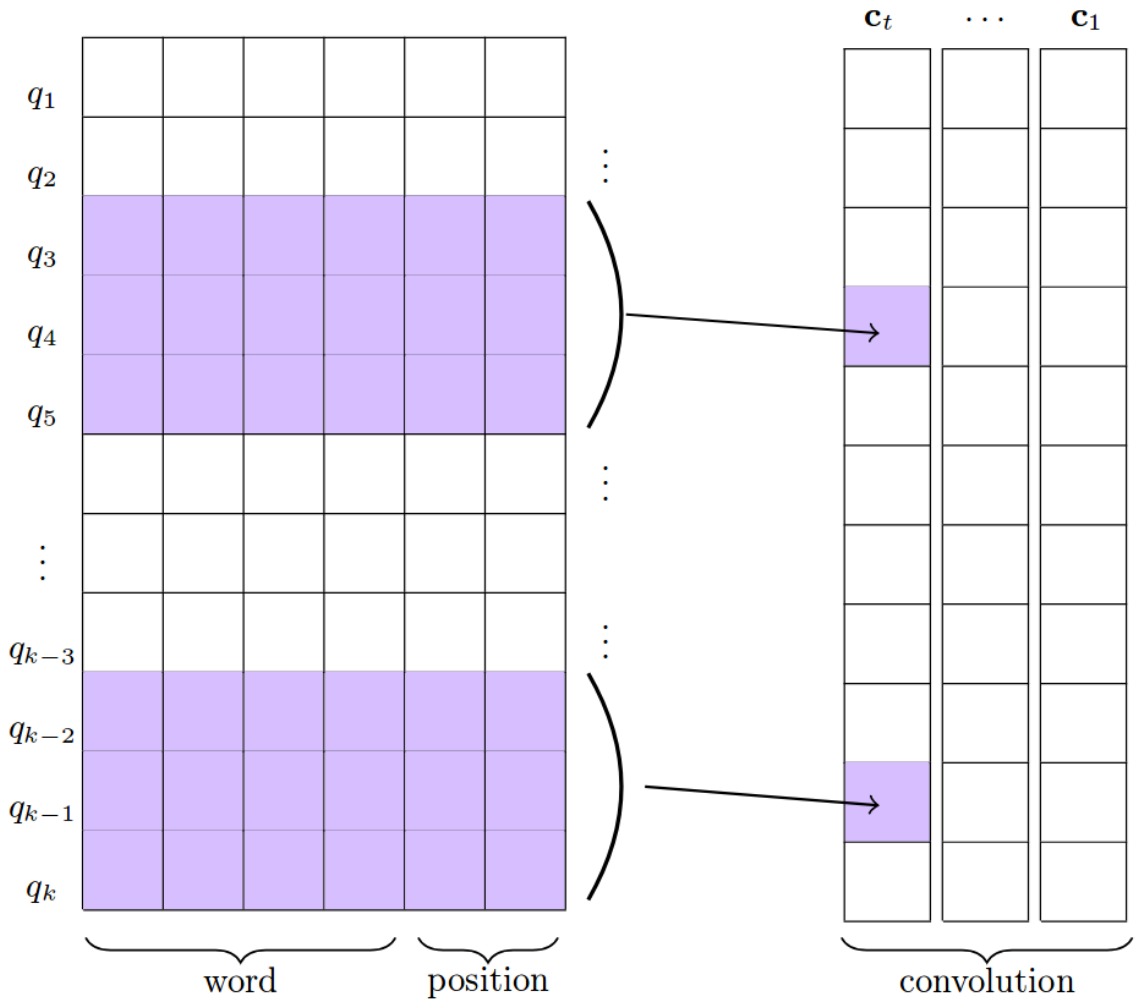
Свертка векторов:

$$W = \{\mathbf{w_1} \dots \mathbf{w_t}\}$$

$$c_j = \mathbf{w} q_{j-w+1:j}$$

Множество сверток:

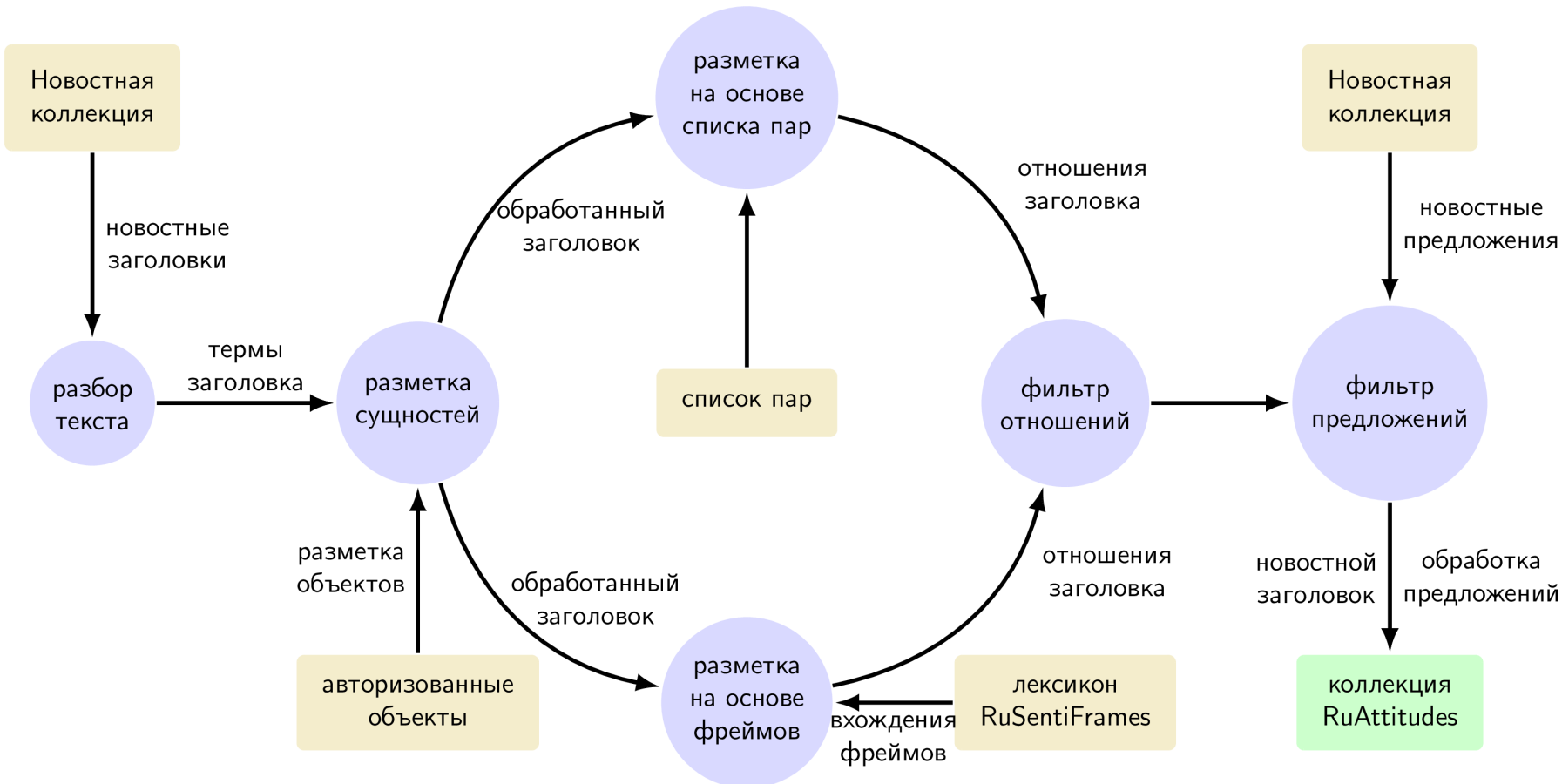
$$C = \{\mathbf{c}_1, \dots, \mathbf{c}_t\}$$



Выводы к разделу 2

- В разделе вводится постановка задачи извлечения оценочных отношений между упомянутыми именованными сущностями на уровне документов.
- Основой выполнения автоматической разметки участников текста являются *контексты*, в которых содержится упоминание пары сущностей – кандидатов на роль оценочного отношения в контексте.
- Для проведения анализа таких контекстов предложена модель на основе архитектуры нейронных сетей сверточного типа. с выполнением операции субдискретизации по частям контекста относительно участников отношения.
- Результаты применения такой модели превосходят классические методы на основе ручных признаков более чем на 15%.
- Наибольший результат в $F^{PN}_{1-mean} = 0.33$ достигается при агрегации всех контекстов при формировании оценки на уровне документов

Архитектура алгоритма ($\mathcal{A}$) обработки новостных документов I



1. Для некоторого документа D'_i выполняется разбор текста с извлечением именованных сущностей:

$$E'_i = \left[e'_1, \ldots, e'_{|E'|}\right] \qquad e' = \langle c_b, c_e, v, t\rangle \qquad t \in T$$

   - Используется предобученная модель $\text{BERT}_{\text{Mult-OntoNotes}}$ библиотеки DeepPavlov[19]. Модель обучена на коллекции OntoNotes, разметка которой включает 19 типов сущностей ($|T| = 19$).

2. Заполнение списка синонимов для сущностей:

$$\mathbf{G} : V \to G$$

   Сущности e'_1 и e'_2 принадлежат одной группе $g \in G$, если совпадают их *нормальные формы*.
   Для e' в качестве нормальной формы используется:
   1. Пакет `Yandex Mystem` для получения лемматизированной формы e'.
   2. Ресурс RuWordNet для получения названия синонимичной группы (если значение найдено).

[19] `http://docs.deeppavlov.ai/en/0.11.0/features/models/ner.html`

Статистика этапов применения алгоритма $\mathcal{A}$

Этап	Параметр	Версии процессов обработки новостей 1.0	2.0	
Коллекция	Тип	$\text{NEWS}_{\text{Base}}$	$\text{NEWS}_{\text{Base}}$	$\text{NEWS}_{\text{Large}}$
	Документы	$2.8 \cdot 10^6$		$8.8 \cdot 10^6$
	Предложений на новость	13.24		14.10
Разметка сущностей	Метод	DeepNER	$\text{BERT}_{\text{Mult-OntoNotes}}$	
	Авторизация объектов	да	нет	
Разметка на основе фреймов	Отношений с участниками между объектами	—	867481	2481426
	Отношений без фреймов между участниками	—	38%	39%
	Отношений без A0→A1	—	12%	12%
	Отношений, перед участниками которых предлоги «в» и «на»	—	15%	15%
	Отношений из заголовков	55566	302319	843799
Список пар	Число пар	246	100329	247876
	Доверенных пар	—	887	2372
	$\langle \lvert\Delta_A\rvert \geq 0.30, \lvert A\rvert \geq 25 \rangle$	—	1%	1%
Разметка на основе списка пар	Отношений из заголовков	60788	—	—
Фильтрация отношений заголовка	Извлечено	22589	65588	200009
	- Разная оценка	7929	13583	42627
		35%	21%	21%
	- Одинаковая оценка	14 660	52005	157382
		65%	79%	79%
Фильтрация предложений	Извлечено	19569	39152	117791
Коллекция RuAttitudes	Версия	1.0-Base	2.0-Base	2.0-Large
	Новостей	13450	44017	134442
	Отношений на новость	1.08	2.28	2.26
	Предложений на новость	1.45	0.89	0.88
Нейтральные отношения	Версия	—	2.0-Base	✓
	Добавлено отношений	—	5428	17790
		—	5.72%	6.23%
	Отношений на новость	—	0.12	0.13
	Отношений на предложение	—	0.03	0.03

Кодировщик контекста с механизмом внимания на основе признаков

1. Признак f: – объект/субъект отношения (маскированный):

$$h_i = [x_i, f]$$

2. Вычисление значимости слова (его веса):

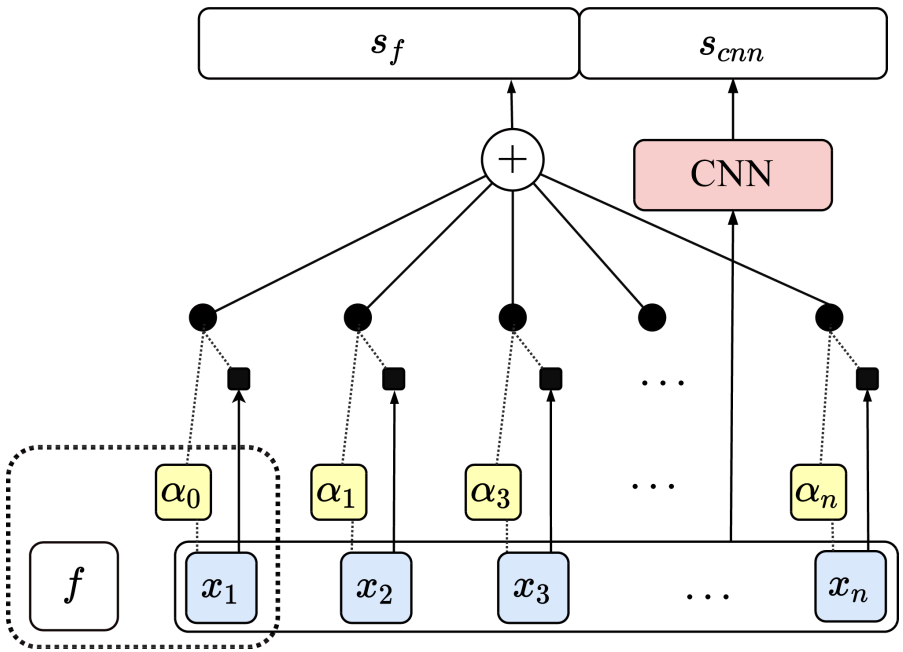
$$u_i = W_a \left(\tanh(W_{we} \cdot h_i + b_{we})\right) + b_a$$

3. Вектор внимания:

$$\alpha = softmax(u)$$

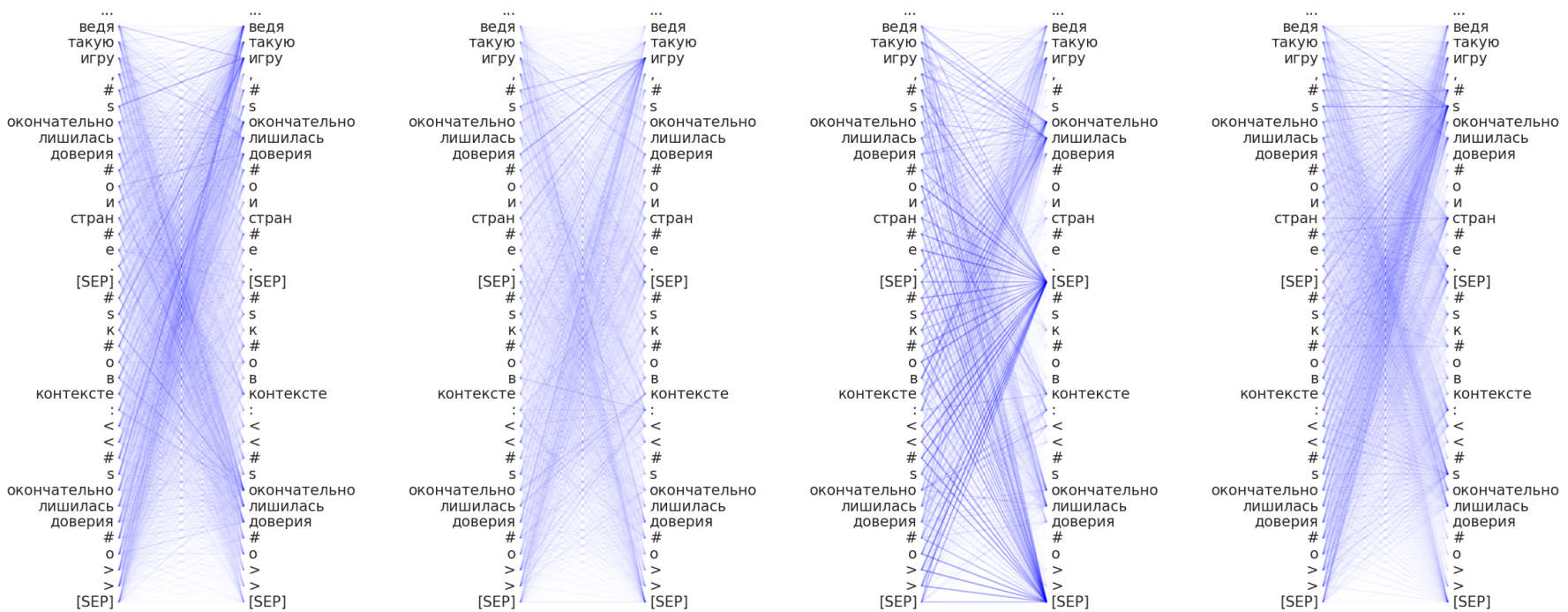
$$\hat{s} = \sum_{i=1}^{n} x_i \cdot \alpha_i$$

$$s_f = avg_{j=1..k}(\hat{s})$$



Визуализация весов внимания моделей BERT I

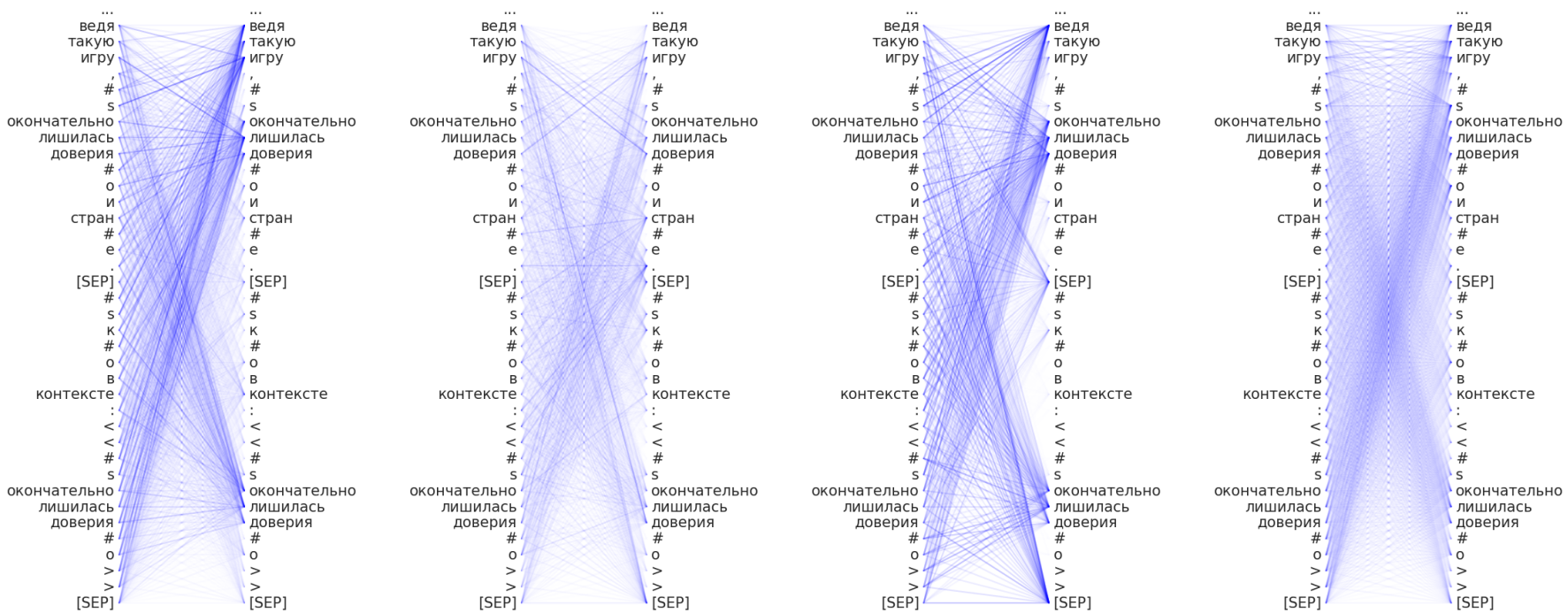
SentRuBERT
голова №2



Визуализация весов внимания моделей BERT II

SentRuBERT-NLI$_P$

голова №2

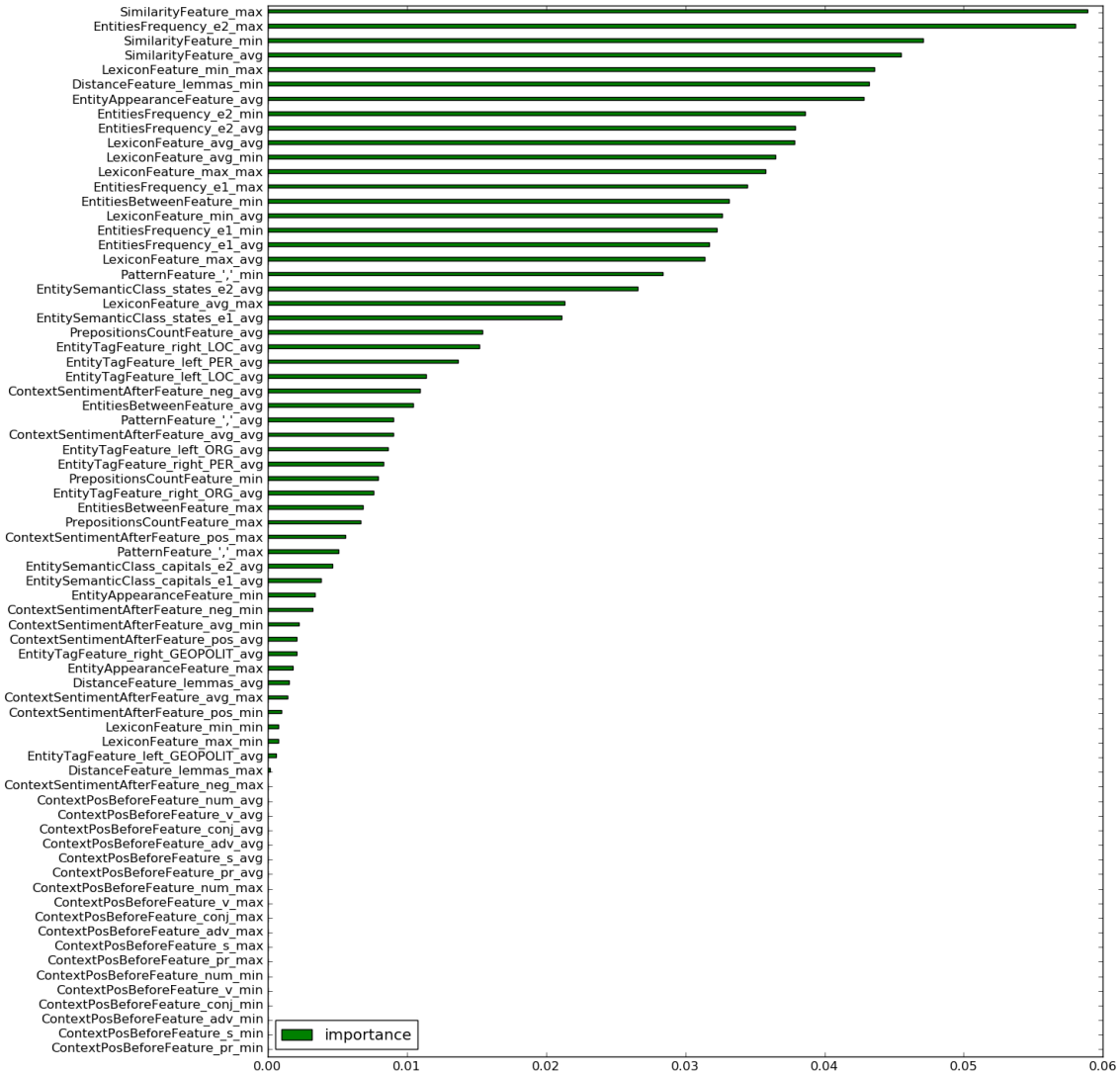


Выводы к разделу 3

- В данной главе вводится постановка задачи опосредованного обучения моделей в задаче извлечении оценочных отношений.
- Основой опосредованного обучения является алгоритм обработки новостных документов, результатом применения которого стала коллекция RuAttitudes. Применение такой коллекции в обучении моделей выполнялось в форматах: (1) предобучения (2) совместного обучения с размеченной коллекцией аналитических текстов.
- Применение опосредованного обучения повысило качество моделей нейронных сетей на 3% при дообучении, и 7-16% при совместном обучении. Прирост качества в обучении языковых моделей BERT составил 10-13% и 25% при сравнении с аналогичными результатами моделей нейронных сетей.
- Предложенные механизмы внимания и исследуемые языковые модели предоставляют информацию для визуального указания наиболее значимых слов текста относительно рассматриваемых в нем участников отношения.

Оценка значимости признаков

Оценка значимости признаков моделью Random Forest (используется 10 деревьев, веса классов равны) по параметру Mean Decrease in Impurity, MDI;
модель предварительно обучена на обучающем множестве коллекции RuSentRel



Статистика разметки нейтральных отношений и оценка качества разметки RuAttitudes

Нейтральные отношения:

Добавлено отношений	**17790 (6.23%)**
Отношений на новость	0.13
Отношений на предложение	0.03

Оценка качества составленной разметки RuAttitudes:

Корпус	Точность
На основе пар сущностей	67.0
На основе фреймов	62.5
Множество с одинаковыми оценками	89.0

Характеристики сервера: 2 x Intel® Xeon® CPU E5-2670 v2 2.50 Ггц, 80 Гб ОЗУ (DDR-3), 2 x Nvidia GeForce GTX 1080 Ti (11.2 Гб); Ubuntu 18.0.4; обучение в контейнерах Docker (19.03.5).

		с учителем	предобучение	дообучение	объединенное
Модель	Версия RA	Время$_{\text{эпох}}$	Время$_{\text{эпох}}$	Время$_{\text{эпох}}$	Время$_{\text{эпох}}$
CNN	✓	—	00:59:42$\|_{18}$	00:04:18$\|_{43}$	01:32:52$\|_{28}$
CNN	—/1.0-Base	00:00:35$\|_{5}$	00:02:20$\|_{07}$	00:02:48$\|_{24}$	00:08:15$\|_{15}$
PCNN	✓	—	03:19:20$\|_{20}$	00:11:30$\|_{30}$	04:47:44$\|_{26}$
PCNN	—/1.0-Base	00:02:10$\|_{5}$	00:09:04$\|_{08}$	00:14:48$\|_{37}$	00:24:44$\|_{14}$
LSTM	✓	—	1$_{\text{д.}}$09:00:45$|_{95}$	00:44:12$\|_{51}$	15:56:48$\|_{46}$
LSTM	—/1.0-Base	00:11:11$\|_{11}$	00:36:16$\|_{16}$	01:20:45$\|_{85}$	00:58:45$\|_{15}$
BiLSTM	✓	—	2$_{\text{д.}}$23:41:10$|_{131}$	04:19:12$\|_{192}$	2$_{\text{д.}}$13:48:20$|_{100}$
BiLSTM	—/1.0-Base	00:19:48$\|_{11}$	01:25:48$\|_{22}$	01:33:45$\|_{73}$	01:34:40$\|_{16}$
AttCNN$_e$	✓	—	21:42:27$\|_{19}$	00:56:15$\|_{25}$	17:05:52$\|_{16}$
AttCNN$_e$	—/1.0-Base	00:14:25$\|_{5}$	00:47:42$\|_{06}$	00:26:10$\|_{10}$	02:50:00$\|_{15}$
AttPCNN$_e$	✓	—	1$_{\text{д.}}$00:14:27$|_{19}$	01:16:00$\|_{30}$	19:50:56$\|_{16}$
AttPCNN$_e$	—/1.0-Base	00:16:15$\|_{5}$	00:59:23$\|_{07}$	00:46:45$\|_{17}$	03:09:45$\|_{15}$
IAN$_e$	✓	—	2$_{\text{д.}}$09:45:48$|_{78}$	04:47:51$\|_{171}$	2$_{\text{д.}}$23:18:30$|_{86}$
IAN$_e$	—/1.0-Base	00:35:06$\|_{18}$	02:47:32$\|_{28}$	02:43:20$\|_{98}$	02:35:20$\|_{20}$
Att-BLSTM	✓	—	3$_{\text{д.}}$03:22:30$|_{134}$	02:45:00$\|_{100}$	15:04:22$\|_{26}$
Att-BLSTM	—/1.0-Base	00:19:15$\|_{11}$	01:01:04$\|_{16}$	03:11:34$\|_{121}$	01:27:30$\|_{14}$
mBERT	✓	—	08:40:14$\|_{04}$	00:10:32$\|_{14}$	—
mBERT	—/1.0-Base	00:10:32$\|_{35}$	00:58:14$\|_{04}$	00:10:32$\|_{14}$	—
RuBERT	✓	—	06:30:11$\|_{03}$	00:06:10$\|_{7}$	—
RuBERT	—/1.0-Base	00:06:10$\|_{12}$	00:43:41$\|_{03}$	00:06:10$\|_{7}$	—

Преобразование оценок с уровня контекстов на уровень отношений

Составление списка отношений документа (**a**):

- для списка контекстов соответствующей субъектно-объектной пары выполняется агрегация оценок **методом голосования**.

Кодировщик контекста с механизмом внимания типа Self-Attention

1. Скрытое состояние **w**.

$$h_i = \overrightarrow{h_i} + \overleftarrow{h_i},\ i \in \overline{1..n}.$$

2. Вычисление значимости слова:

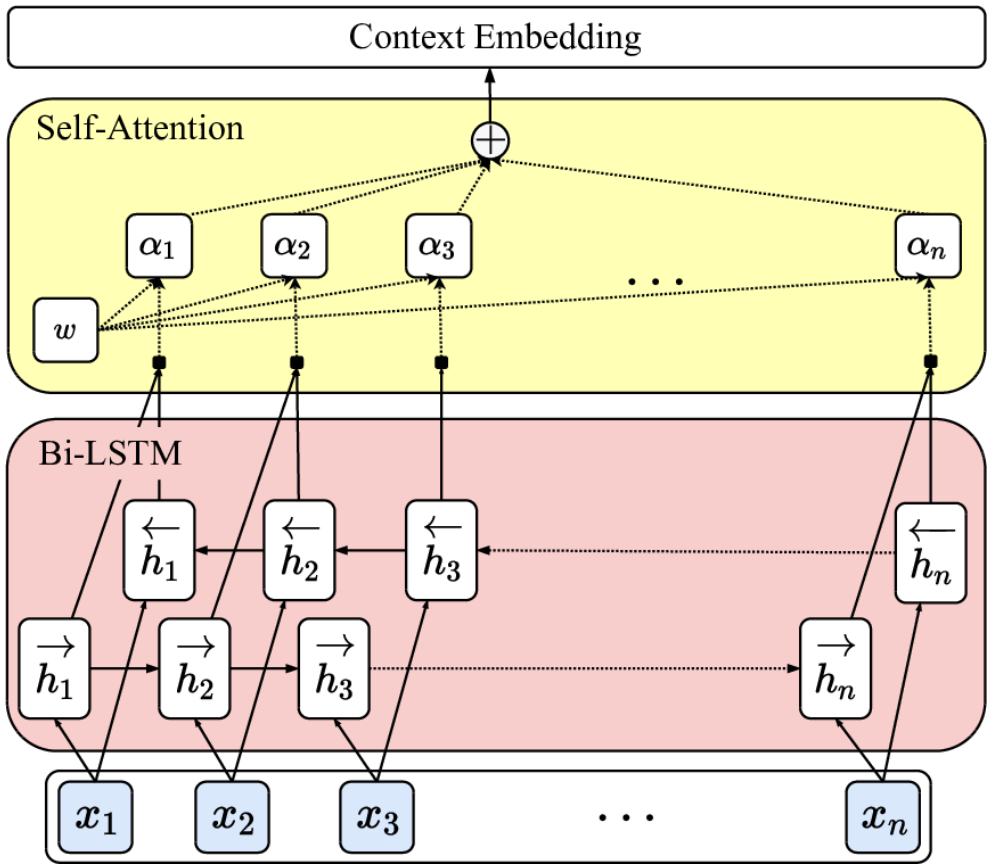
$$m_i = \tanh(h_i)$$

$$u_i = m_i^T \cdot \mathbf{w}$$

3. Вектор внимания:

$$\alpha = softmax(u)$$

$$s = \tanh(H \cdot \alpha)$$



Подбор числа фильтров для модели CNN/PCNN

CNN vs. PCNN