DOI: 10.15514/ISPRAS-2018-1(2)-33

Language Models Application in Sentiment Attitude Extraction Task

[1]*Rusnachenko N. L. <kolyarus@yandex.ru>*
[1] *Bauman Moscow State Technical University,*
5, Building 1, 2-nd Baumanskaya Str., Moscow, 105005, Russia

Abstract. Large text can convey various forms of sentiment information including the author's position, positive or negative effects of some events, attitides of mentioned entities towards to each other. In this paper, we experiment with BERT based language models for extracting sentiment attitudes between named entities. Given a mass media article and list of mentioned named entities, the task is to extract positive or negative attitudes between them. Efficiency of language model methods depends on the amount of training data. To enrich training data, we adopt distant supervision method, which provide automatic annotation of unlabeled texts using an additional lexical resource. The proposed approach is subdivided into two stages FRAME-BASED: (1) sentiment pairs list completion (PAIR-BASED), (2) document annotations using PAIR-BASED and FRAME-BASED factors. Being applied towards a large news collection, the methods generates RuAttitudes2017 automatically annotated collection. We evaluate the approach on RuSentRel-1.0, consisted of mass media articles written in Russian. Adopting RuAttitudes2017 in the training process results in 10-13% quality improvement by F1-measure over supervised learning and by 25% over the top neural network based model results.

Keywords: Sentiment Analysis, Relation Extraction, Distant Supervision, Neural Networks, Language Models

For citation: Rusnachenko N. L. Language Models Application in Sentiment Attitude Extraction Task. Trudy ISP RAN/Proc. ISP RAS, 2018, vol. 1, issue 2, pp. 3–4. 10.15514/ISPRAS-2018-1(2)-33

Acknowledgments: This work was supported by a grant from the RFBR 20-07-01059

Применение языковых моделей в задаче извлечения оценочных отношений

[1]*Русначенко Н. Л. <kolyarus@yandex.ru>*
[1] *Московский государственный технический институт им. Н.Э.Баумана, 105005б Россия, Москва, 2-я Бауманская ул., д. 5, стр. 1*

Аннотация. Объемные тексты могут содержать источники взаимосвязанной информации различных типов, передаваемых посредством отношений, некоторые из которых могут быть оценочными. Проведение анализа таких текстов требует установление подобных связей, определении их участников: события, сущности, и т.д. В данной работе исследуется применение языковых моделей BERT в задаче извлечения оценочных отношений. Для произвольного документа и списка размеченных в нем именованных сущностей, такая задача предполагает составление списка оценочных отношений между ними. Эффективность применения языковых моделей напрямую зависит от объема обучающих данных. Для увеличения объема обучающего множества применяется подход опосредованного обучения. Такое обучение подразумевает применение алгоритма автоматической разметки оценочных отношений из сторонних источников. Предложенный подход разметки оценочных отношений основан на двуэтапном применении FRAME-BASED фактора в анализе новостных документов, для: (1) составления списка оценочных пар (PAIR-BASED), (2) разметки документов с использованием PAIR-BASED и FRAME-BASED факторов. Полученная на основе такого алгоритма коллекция получила название RuAttitudes2017. Для проведения экспериментов с моделями использовался корпус новостных текстов на русском языке RuSentRel-1.0. Применение опосредованного обучения с использованием коллекции RuAttitudes2017 повысило качество моделей на 10-13% по метрике F1, и на 25% при сравнении с наилучшими результатами моделей на основе нейронных сетей.

Ключевые слова: анализ тональности, извлечение отношений, опосредованное обучение, нейронные сети, языковые модели

Для цитирования: Русначенко Н. Л. Применение языковых моделей в задаче извлечения оценочных отношений. Труды ИСП РАН, 2018, том 1 вып. 2, с. 3–4. 10.15514/ISPRAS-2018-1(2)-33

Благодарности: Данная работа выполнена при поддержке гранта РФФИ 20-07-01059

1. Введение

Анализ тональности, т.е. выделение мнения автора к предмету обсуждения в тексте, является одним из наиболее востребованных приложений автоматической обработки текстов за последние годы. Одной из подзадач анализа тональности является задача извлечения оценочных отношений [1], которая предполагает

классификацию взаимоотношений между упоминаемыми в тексте именованными сущностями. Извлечение оценочных отношений существенно для анализа тональности новостных и аналитических текстов, поскольку сложным образом влияет на анализ авторской позиции в тексте. В следующем примере приводится фрагмент новостного сообщения, оценочные отношения возникают между сущностями «Россия» и «НАТО» (сущности подчеркнуты): ... *Москва$_e$ неоднократно подчеркивала, что ее активность на Балтике$_e$ является ответом именно на действия НАТО$_e$ и эскалацию враждебного подхода к Россию$_e$ вблизи ее восточных границ ...*

Многие задачи анализа тональности решаются на основе методов машинного обучения, которые, однако, требуют значительного объема обучающих данных. Одним из подходов, направленным на снижение объема ручной разметки данных, является подход *опосредованного обучения* (от англ. Distant Supervision). Опосредованное обучение предполагает выполнение автоматической разметки объемных текстовых коллекций [2] на основе некоторых дополнительных ресурсов, полученная размеченная коллекция далее используется в качестве данных для методов машинного обучения. Несмотря на большое число проведенных исследований подобного подхода разметка документов [3, 4] для задачи анализа тональности и извлечения отношений, область остается изученной лишь частично [5].

В данной работе исследуется применение языковых моделей для извлечения оценочных отношений, предобученных на основе большого автоматического размеченного корпуса извлеченных оценочных отношений по методу опосредованного обучения. Подход основан на использовании лексикона RuSentiFrames [6], который содержит описание оценочных отношений между аргументами слов-предикатов русского языка.

Таким образом, вклад настоящей работы следующий:

- Исследованы методы машинного обучения для извлечения оценочных отношений из русскоязычных аналитических текстов на уровне документа.
- Предложен подход к автоматическому порождению обучающей коллекции для извлечения оценочных отношений, включающий: (1) предварительный этап обработки коллекции для автоматического порождения списка оценочных пар, (2) автоматическую разметку нейтральных отношений.
- Проведены исследования извлечения отношений на основе корпуса RuSentRel-1.0 для языковых моделей BERT [7] с применением предложенного подхода порождения обучающей коллекции; согласно полученным результатам исследования, применение опосредованного обучения улучшает качество извлечения оценочных отношений языковыми моделями на 10-13% (трехклассовая классификация) по F1-мере и на 25% при сравнении результатов русскоязычных языковых моделей с аналогичными результатами других архитектур нейронных сетей

(кросс-валидационное тестирование).

2. Языковые модели для извлечения отношений

Появление архитектуры *трансформера* [8] оказало огромное влияние в решении многих задач автоматической обработки естественного языка. Эта архитектура основана на независимом применении кодировщика [7] и декодировщика [9] трансформера. Применение таких компонентов подразумевает выполнение этапов: (1) предварительного обучения на большом объеме неразмеченных данных и (2) дообучение под конкретную задачу обработки текстов на естественном языке. По завершении первого этапа, модели на основе таких компонентов могут быть интерпретированы как *языковые модели* – вероятностные распределения над последовательностями слов.

Основополагающей моделью на основе *декодировщика* стала GPT [9]. Актуальной на настоящий момент версией является модель GPT-3 [10]. Дообученная версия такой модели на русскоязычных данных получила название **ruGPT-3**[1]. В работе [11] авторы представляют модель для извлечения отношений, которая основывается на классической архитектуре трансформера [8] и дообучении GPT [9], что привело к модели под названием **TRE** [11].

В случае *кодировщиков*, основополагающей моделью стала **BERT** [7]. Такая модель предполагает в качестве входной информации последовательность, опциально разделенную специальным символом [SEP] на две независимые последовательности: TextA и TextB. Учет всех слов контекста достигается благодаря введению задачи *предсказания маскированных токенов* (от англ. Masking Language Modeling): предсказание случайного маскированного слова входной последовательности. Дополнительной задачей, призванной установить связь между TextA и TextB, стала *Natural Language Inference* (NLI), в которой требуется определить[2], является ли TextB продолжением TextA. Применение языковых моделей BERT в задачах классификации выполняется с введением *классификационного слоя*, отвечающего за сопоставление входной последовательности множеству выходных классов задачи.

В области аспектно-ориентированного анализа тональности, авторы [12] предлагают подход применения модели BERT в **вопросно-ответной** постановке (Q/A, составление вопроса в TextB для последовательности TextA), и в постановке **вывода по тексту** NLI (указание ожидаемой информации в TextB, которая должна быть выведена из TextA).

Одним из направлений в развитии BERT-производных архитектур стала публикация предобученных моделей. Исходно доступный набор предобученных моделей[3] делится на: ориентированные под конкретные языки (английский,

[1] `https://github.com/sberbank-ai/ru-gpts`

[2] Для проведения классификации в модели вводится предусмотрен специальный токен [CLS] перед началом входной последовательности

[3] `https://github.com/google-research/bert`

китайский) и мультиязыковые. Из множества мультиязыковых моделей выделим модель **mBERT**[4], которая предобучена на текстовых данных 104 языков и поддержкой регистра букв в представлении входных последовательностей [7]. Модель mBERT доступна и распространена только в формате base. Для русского языка, авторами проекта DeepPavlov опубликована модель **RuBERT** [13] – дообученная версия mBERT на русскоязычных новостных данных и статьях энциклопедии «Википедия». Модель **SentRuBERT**[5] является дообученной версией RuBERT коллекциями: (1) переведенных на русский язык текстами корпуса SNLI [14] сервисом Google-Translate; (2) русскоязычных текстов корпуса XNLI [15].

Другим направлением в развитии BERT-архитектур стала модификация используемых задач этапа предобучения. В модели **Electra** [16] задача предсказания маскированного слова модифицирована в задачу выявления в контексте специально подмененного слова. **RoBERTa** [17] представляет собой улучшение предобучения (задач на этапе предобучения моделей). Построение модели кросс-языкового кодировщика предложений на основе набора текстов ста различных языков [18] стало одним из применением модели RoBERTa. Такая модель получила название **XLM-R**[6]. **SpanBERT** [19] представляет собой модификацию BERT, ориентированную под задачу извлечения отношений (от англ. Relation Extraction) [20], посредством изменения алгоритма маскирования частей текста последовательности.

Архитектура классификационного слоя может быть также специфична для конкретной задачи. Например, для задачи извлечения отношений модель **R-BERT** [21] учитывает информацию об объектах отношения входной последовательности.

3. Используемые ресурсы

В п. 3.1–3.2 рассматриваются ресурсы, которые использовались для разметки коллекции с целью проведения опосредованного обучения моделей. Основная идея подхода состоит в следующем. Лексикон оценочной лексики RuSentiFrames [6] используется для автоматической разметки оценочных отношений в заголовках большой неразмеченной новостной коллекции. Извлечение отношений производится из заголовков, поскольку они обычно короче, содержат меньше именованных сущностей. Далее размеченные отношения в заголовках фильтруются и используются для разметки отношений и внутри текстов новостей.

[4] https://huggingface.co/bert-base-multilingual-cased
[5] https://huggingface.co/DeepPavlov/rubert-base-cased-sentence
[6] https://github.com/pytorch/fairseq/tree/master/examples/xlmr

Таблица 1: Пример описания фрейма «Одобрить» в лексиконе RuSentiFrames

Измерения фрейма «Одобрить»	Описание
roles	A0: тот, кто одобряет A1: то, что одобряется
polarity	A0→A1 , pos, 1.0 A1→A0, pos, 0.7
effect	A1, pos, 1.0
state	A0, pos, 1.0 A1, pos, 1.0

Таблица 2: Распределение вхождений отношений в лексиконе RuSentiFrames-2.0

Полярность	Класс тональности	Количество
A0→A1	pos	2558
A0→A1	neg	3289
author→A0	pos	170
author→A0	neg	1581
author→A1	pos	92
author→A1	neg	249

3.1 Лексикон фреймов RuSentiFrames

Лексикон RuSentiFrames-2.0 описывает оценки и коннотации, передаваемые предикатом в форме отдельного слова (существительного, глагола) или словосочетания. Структура фреймов включает в себя набор специфичных для предикатов ролей и набор различных измерений (характеристик) для описания фреймов. Для обозначения ролей семантические аргументы предикатов нумеруются, начиная с нуля. Для конкретного предиката Arg0 – это, как правило, аргумент (Agent), демонстрирующий свойства агента (активного участника) ситуации [22], в то время как Arg1 это объект (Theme).

В основной части лексикона представлены следующие *измерения*:

- Отношение автора текста к указанным участникам (Roles);
- Polarity – положительная или отрицательная оценка между участниками отношений;
- Effect – положительный или отрицательный эффект для участников;
- State – положительное или отрицательное эмоциональное состояние участников, связанных с описанной ситуацией.

Все утверждения включают доверительную оценку, которая в настоящее время имеет два значения: 1 – утверждение почти всегда верно, или 0.7 – разметка по-умолчанию. Утверждения о нейтральной оценке, эффекте или состоянии участников не учитываются в лексиконе.

Созданные фреймы связаны также с «семейством» слов и выражений (лексических единиц), которые имеют одинаковые тональности. Лексические единицы, связанные с фреймом, могут быть отдельными словами или словосочетаниями.

RuAttitudes-2.0 сохраняет общую структуру лексикона версии 1.0. В ресурсе описано 311 фреймов, связанных с 7034 лексическими единицами, среди которых 6788 уникальных, из которых 48% – глаголы, 14% – существительные и оставшиеся 38% – словосочетания. Число вхождений фреймов увеличено на 12% при сравнении с версией 1.0.

Пример формата описания фрейма «Одобрить» приведен в Таблице 1.

Таблица 2 предоставляет статистику различных типов отношений в RuSentiFrames. Для проведения автоматической разметки в методе опосредованного обучения моделей используется только отношения агента ситуации к объекту (A0→A1).

3.2 Новостные коллекции

Коллекции $NEWS_{Base}$ (2,8 млн. новостных текстов) и $NEWS_{Large}$ (8,8 млн. новостных текстов), используемые для извлечения отношений, состоят из русскоязычных статей и новостей крупных новостных источников, специализированных политических сайтов, опубликованных в 2017 году. Каждая статья разделена на заголовок и содержание.

4. Описание подхода

Основные предположения подхода состоят в следующем:

- отношения между сущностями, упоминаемыми в новости, в большинстве случаев наиболее четко и просто выражаются в заголовке новости;
- появление предиката из RuSentiFrames (FRAME-BASED) в заголовке позволяет достаточно надежно извлечь отношения между именованными сущностями;
- суммирование выделенных отношений по большой коллекции позволяет выделить основную тональность отношений между сущностями (PAIR-BASED фактор);
- для формирования автоматически размеченной коллекции выбираются заголовки новостей, в которых тональность отношений между сущностями, выделенная на основе фреймов (FRAME-BASED) совпадает с насчитанной тональностью по коллекции для этих сущностей (PAIR-BASED) – так называемые доверенные отношения;
- в размеченную коллекцию также включаются предложения из тела новости с выбранным заголовком, поскольку предполагается, что в среднем тональность отношения между сущностями внутри новости соответствует тональности отношения в заголовке. При этом предложения из тела новости имеют более разнообразную структуру.

Полученный набор данных с автоматически размеченными оценочными отношениями получил название RuAttitudes2017. Рис. 1 иллюстрирует процесс

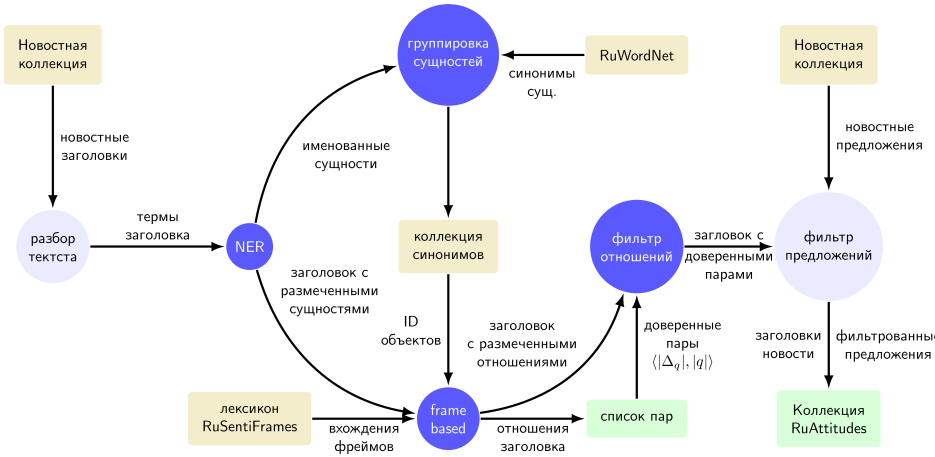

Рис. 1: Диаграмма рабочего процесса извлечения оценочных отношений; *прямоугольники* – источники информации; *кружки* – компоненты обработки потока данных; *стрелки* – передача информации между блоками с указанием ее типа в подписи; для произвольной пары q, $|\Delta_q|$ – абсолютная разница вероятностей принадлежности положительному и отрицательному классам ($|\Delta_q| \in [0, 1]$), и $|q|$ – число соответствующих отношений

автоматической разметки[7] новостной коллекции.

4.1 Извлечение оценочных отношений из новостных статей

Процесс извлечения оценочных отношений включает выполнения двух последовательных этапов обработки новостной коллекции:

- Этап 1 – автоматическое составление списка пар сущностей с превалирующей тональностью отношений из заголовков новостей из неразмеченного корпуса текстов;
- Этап 2 – применение собранного списка пар сущностей с выявленной тональностью отношений для отбора достоверных отношений из новостных заголовков и текстов новостей для формирования автоматически размеченной коллекции RuAttitudes.

Рассмотрим общие компоненты потока обработки информации обоих этапов для заголовка некоторого документа новостной коллекции.

Модуль разбора текста подразумевает выполнение преобразования заголовка представленного последовательностью символов в последовательность термов. Содержимое заголовка разбивается на слова с выделением знаков препинания.

Модуль NER выполняет задачу извлечения именованных сущностей из

[7] https://github.com/nicolay-r/RuAttitudes/tree/v2.0

последовательности термов. Для этого используется предобученная модель $\text{BERT}_{\text{Mult-OntoNotes}}$ библиотеки DeepPavlov[8]. Модель обучена на коллекции OntoNotes [23], разметка которой включает 19 типов сущностей. Результатом такого модуля обработки является список сущностей $E = [e_1, \ldots e_{|E|}]$, каждый элемент которого представлен последовательностью термов и типом.

Модуль *группировки сущностей* использует множество E для пополнения списка синонимов. Пара сущностей $e_i, e_j \in E, i \neq j$ являются синонимами, если совпадают их *нормальные формы*. Для получения нормальной формы именованной сущности используется:

1. Лемматизированная форма значения (последовательности термов)[9];
2. Ресурс RuWordNet [24] для получения совокупности синонимичных вариантов упоминания сущности (если вхождение сущности найдено).

В результате можно автоматически сгруппировать такие синонимы, как: (США, Соединенные Штаты), (Россия, Российская Федерация, РФ).

Модуль *Frame-Based* выполняет задачу извлечения достоверных отношений из новостного заголовка с использованием лексикона RuSentiFrames. Для этого на первом шаге из последовательности термов извлекаются вхождения фреймов. Далее составляется множество достоверных пар сущностей. Пара $\langle e_i, e_j \rangle$, где $e_i, e_j \in E$ считается *достоверной*, если выполнены следующие условия:

1. Именованная сущность e_i упомянута раньше e_j;
2. Участники e_i и e_j не являются синонимами;
3. Участники и все именованные сущности между ними принадлежат множеству O_{valid}, которое включает: организации (`ORG`), людей (`PER`), геополитические сущности (`GPE`);
4. Для всех фреймов, входящих между участниками отношений, определена полярность типа `A0`→`A1`;
5. Отсутствуют предлоги[10] «в» и «на» перед участниками отношений.

Касательно условия п. 5, наличие предлогов «в» и «на» в большинстве случаев связано с месторасположением, которое обычно не является субъектом или объектом оценочного отношения: *Крым$_e$ бросит вызов Киеву$_e$: «в» ООН$_e$ представят резолюцию о преступлениях против людей «на» Украине$_e$, включающая убийства и похищения.*

Этап 1. Заполнение списка пар. Для некоторого новостного заголовка с множеством размеченных в нем оценочных фреймов, пусть P – множество извлеченных достоверных отношений (результат применения модуля Frame-Based, рис. 1). Тогда, каждая пара $\langle e_i, e_j \rangle \in P$ отправляется в *список пар* в следующем

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

[9] Применяется пакет `Yandex Mystem`

[10] Условие является результатом проведения дополнительного анализа ошибочных результатов.

Таблица 3: Список доверенных пар, извлеченных из новостной коллекции $\text{NEWS}_{\text{Large}}$, при ограничениях $|\Delta_q| \geq 0.8$, $|A_q| \geq 150$, где q – произвольная достоверная пара; содержимое упорядочено по Δ_q; пары с одинаковым значением Δ_q упорядочены относительно $|A_q|$: по-убыванию ($\Delta_q > 0$), по-возрастанию ($\Delta_q < 0$)

A0	A1	Δ_q	$\lvert A_q\rvert_{pos}$	$\lvert A_q\rvert_{neg}$
МВД России	Российская Федерация	1.00	256	0
Путин	Министерство Внутренних дел	0.91	150	7
Канада	Украина	0.90	218	11
Пентагон	Украина	0.90	147	8
Порошенко	НАТО	0.88	244	16
Порошенко	Совет Национальной Безопасности и Обороны	0.87	173	12
Путин	Макрон	0.86	186	14
Афганистан	Россия	0.85	166	13
Европейский Парламент	Украина	0.84	273	23
Украина	МВФ	0.80	204	23
Трамп	ИГИЛ	-0.79	24	200
Россия	ИГИЛ	-0.79	60	516
Гройсман	Донбасс	-0.82	23	232
Турция	ИГИЛ	-0.83	14	149
Россия	Siemens	-0.83	19	204
Израиль	ООН	-0.85	12	144
Азербайджан	Армения	-0.93	14	412
Карабах	Азербайджан	-0.94	10	346
ЕС	Siemens	-1.00	0	186

формате:

$$a = \langle d, g_i, g_j, l \rangle \tag{1}$$

где $d \in N$ – индекс документа рассматриваемого заголовка в новостной коллекции, $g_i, g_j \in N$ – индексы синонимичных групп участников в списке синонимов, а l – оценка пары, которая назначается следующим образом: pos (если для всех вхождений фреймов между e_i и e_j оценка `A0`→`A1` одинакова, и равна pos), neg (иначе).

Таким образом, результирующий *список пар* (см. рис. 1) представляет собой множество достоверных пар $A = \{a_1, a_2, \ldots a_{|A|}\}$, извлеченных из заголовков всех документов новостной коллекции.

Извлечение доверенных пар. Из составленного списка пар (см. рис. 1), представленного множеством A, можно выделить наиболее положительно и отрицательно ориентированные пары. *Ориентация* некоторой пары $q = \langle g_i, g_j \rangle$ к классу $c \in \{\text{pos}, \text{neg}\}$ вычисляется по формуле:

$$p(q|c) = \frac{|\{\langle d, g_i, g_j, l \rangle \, | l = c\}|}{|A_q|} \tag{2}$$

где A_q – подмножество множества A, элементы которого соответствуют паре q.

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

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

Список доверенных пар $\langle \lvert\Delta_q\rvert \geq 0.3, \lvert A_q\rvert \geq 25 \rangle$	
Запрос	Результат поиска
сша$\rightarrow$россия$_{neg}$	пара найдена, оценки совпадают: «сша» $\rightarrow$ «россия» (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$_e$.

Рис. 2: Применение метода Pair-Based для извлечения достоверных пар из заголовка с последующим выполнением фильтрации отношений и поиском доверенных пар (США$\rightarrow$РФ$_{neg}$) в предложениях новости

Оценочная ориентация пары q определяется по формуле:

$$\Delta_q = p(q|\text{pos}) - p(q|\text{neg}) \tag{3}$$

Результирующая оценка для q определяется знаком выражения формулы 3: pos ($\Delta_q > 0$), neg ($\Delta_q < 0$). Таким образом, для извлечения и составления множества доверенных пар A' необходимо задать пороговые значения для $|\Delta_q|$ и $|A_q|$. Формат представления доверенной пары q в множестве A' следующий:

$$q = \langle g_i, g_j, \Delta_q \rangle \tag{4}$$

В таблице 3 приведены примеры доверенных пар в результате анализа новостной коллекции $\text{NEWS}_{\text{Large}}$ при $|\Delta_q| \geq 0.8$ и $|A_q| \geq 150$.

Этап 2. Разметка оценочных отношений. Для некоторого новостного заголовка, пусть P – множество извлеченных достоверных отношений (результат применения модуля Frame-Based, рис. 1). Модуль *фильтрации отношений* выполняет отбор оценочных отношений среди множества достоверных пар P. Пара $\langle e_i, e_j \rangle \in P$ считается оценочным отношением, если $\langle g_i, g_j \rangle$ содержится в множестве доверенных пар A' и оценка $\langle e_i, e_j \rangle$ совпадает с оценочной ориентацией доверенной пары.

Отобранные оценочные отношения далее передаются на вход модулю *фильтрации предложений* для поиска таких же отношений в предложениях новости. Оценочное отношение заголовка присутствует в предложении новости, если предложение содержит упоминание обеих участников. На рисунке 2 рассмотрено применение второго этапа процесса извлечения оценочных отношений для заголовка: *«США$_e$ не снимут* ***санкции***$_{neg}$ *с РФ$_e$ до возвращения Крыма$_e$»*.

Для каждого документа новостной коллекции дополнительно проводится разметка нейтральных отношений. Для некоторого документа с множеством размеченных именованных сущностей E, пара именованных сущностей $\langle e_1, e_2 \rangle \in E$ заголовка или предложения считается нейтральной, если выполнены следующие условия:

- Сущность e_1 упомянута в тексте перед e_2 и имеет тип из множества O_{valid};
- Сущность e_2 имеет тип LOC и не находится в списке стран/столиц;
- Участники e_1 и e_2 не принадлежат одной синонимичной группе, а также отношения $\langle e_1, e_2 \rangle$ и $\langle e_2, e_1 \rangle$ не содержатся в разметке оценочных отношений.

4.2 Автоматическая разметка отношений и анализ результатов

Поэтапная оценка количества извлеченных данных в результате применения потока обработки (см. рис. 1) к новостным коллекциям приведена таблице 4. Результатом применения подхода автоматической разметки новостных статей стали коллекции RuAttitudes2017, созданные независимо в результате обработки $\text{NEWS}_{\text{Base}}$ и $\text{NEWS}_{\text{Large}}$. Рассмотрим подробнее каждый этап обработки новостных текстов.

На первом этапе список пар заполняется отношениями, которые были извлечены методом Frame-Based. Среди всех заголовков отбираются отношения, участники которых имеют тип из множества O_{valid}. Далее, процент отвергнутых отношений относительно такого числа составил 65%, из которых: 38% отношений без вхождений фреймов между сущностями, 12% отношений, для которых существуют вхождения фреймов с неопределенной полярностью A0→A1, и 15% с наличием предлогов «в» и «на». Таким образом, 35% отношений от изначального числа были отобраны как «достоверные» и переданы в *список пар*.

На втором этапе производится фильтрация отношений из заголовков и предложений новостей (см. Таблицу 4). Для извлечения доверенных пар были выбраны параметры: $|\Delta_q| \geq 0.3$, $|q| \geq 25$. В результате, 22-24% достоверных отношений из заголовков были сопоставлены с доверенными парами, среди которых 79% отношений совпадали с оценочной ориентацией соответствующих пар. Новости с такими отношениями в заголовках передавались на этап фильтрации предложений. Дополнительный выбор новостных предложений позволил увеличить объем разметки на 89%.

Объем нейтрально размеченных отношений составил 5-6% от общего числа оценочных отношений коллекций RuAttitudes2017. Расширенные версии коллекций получили названия 2017-Base и 2017-Large для $\text{NEWS}_{\text{Base}}$ и $\text{NEWS}_{\text{Large}}$ соответственно. Среди объектов таких пар, в большинстве случаев, к сущности типа LOC относятся: моря, озера, острова, реки, и т.д (см. таблицу 5).

5. Эксперименты

5.1 Корпус RuSentRel

Корпус предоставляет собой 75 больших аналитических текстов по международной политике с портала ИНОСМИ[11], размеченных с выделением

[11] inosmi.ru

Таблица 4: Количественная оценка параметров автоматической разметки текстов новостных коллекций $\text{NEWS}_{\text{Base}}$ и $\text{NEWS}_{\text{Large}}$; выделенные зеленым цветом результаты соответствуют количественной оценке ресурсов в результате обработки новостных коллекций двумя этапами

Этап	Параметр		
Коллекция	Тип новостной коллекции	$\text{NEWS}_{\text{Base}}$	$\text{NEWS}_{\text{Large}}$
	Документы	$2.8 \cdot 10^6$	$8.8 \cdot 10^6$
FRAME-BASED	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	302319	843799
Список пар	Число пар	100329	247876
	Доверенных пар	887	2372
	$\langle \lvert\Delta_q\rvert \geq 0.3, \lvert A_q\rvert \geq 25 \rangle$	1%	1%
	Отношений сопоставленных с доверенными парами	65588	200009
		22%	24%
Фильтрация отношений заголовка	Извлечено	65588	200009
	- Разная оценка	13583	42627
		21%	21%
	- Одинаковая оценка	52005	157382
		79%	79%
Фильтрация	Извлечено предложений	39152	117791
RuAttitudes	Версия	2017-Base	2017-Large
	Новостей	44017	134442
	Отношений на новость	2.28	2.26
	Предложений на новость	0.89	0.88
Нейтральные отношения	Версия	2017-Base	2017-Large
	Добавлено отношений	5428	17790
		5.72%	6.23%
	Отношений на новость	0.12	0.13
	Отношений на предложение	0.03	0.03

порядка 2000 оценочных отношений между упомянутыми в текстах сущностями. Общая статистика корпуса по фиксированным разделением документов на обучающее и тестовое множества, приведена в таблице 6. В текстах статей автоматически размечены именованные сущности по четырем классам: личности (PER), организации (ORG), места (LOC), геополитические сущности (GEO). Общее число размеченных именованных сущностей составляет 15.5 тысяч.

Разметка отношений поделена на два типа: (1) отношение автора к упомянутой именованной сущности; (2) отношения субъектов, переданное от одних именованных сущностей к другим именованным сущностям. Отношения фиксируются тройками, и рассматриваются не для каждого предложения, а для документа в целом. Оценка отношения может быть отрицательной (neg), либо положительной (pos); например: (Автор, США, neg), (США, Россия, neg). Нейтральные а также отсутствующие отношения в корпусе не зафиксированы.

Таблица 5: Примеры наиболее частотных, нейтрально размеченных отношений из корпуса $\text{RuAttitudes2017}_{\text{Large}}$

A0	A1	Вхождений	Процент
КНДР	корейский полуостров	301	1.7%
Россия	ближний восток	232	1.3%
США	Баренцево море	204	1.1%
Иран	ближний восток	189	1.1%
Япония	Курилы	172	1.0%
США	ближний восток	166	0.9%
РФ	Курилы	163	0.9%
Волгоград	река волга	155	0.9%
Правительство РФ	волга	120	0.7%
Япония	Южный Курилы	115	0.6%
КНДР	Тихий Океан	103	0.6%
Сирия	Тивериадский Озеро	93	0.5%
НАТО	Североатлантический	92	0.5%
Гуам	Тихий Океан	79	0.4%
Израиль	Тивериадский Озеро	74	0.4%
Россия	Арктика	73	0.4%

Таблица 6: Параметры корпуса RuSentRel-1.0 с фиксированным разбиением на обучающую и тестовые коллекции

Коллекция	Обучающая	Тестовая
Документов	44	29
Предложений (ср./док.)	74.5	137
Упомянутых сущностей (*NE*) (ср./док.)	194	300
Сущностей (ср. на документ)	33.3	59.9
Положительных пар сущностей (ср./док.)	7.23	14.7
Негативных пар (ср./док.)	9.33	15.6
Расстояние между *NE* в предложении (в словах)	10.2	10.2
Нейтральных пар (ср./док.)	120	276

5.2 Описание эксперимента

Пусть задано подмножество документов коллекции RuSentRel, в котором каждый документ представлен парой: (1) текст, (2) список выделенных именованных сущностей E. Используя методы машинного обучения, для каждого документа требуется составить список оценочных отношений между парами сущностей множества E. Оценка отношения может быть отрицательной (neg), либо положительной (pos) (согласно п. 5.1). Составление списка выполняется в двух независимых экспериментах:

1. *Двуклассовый* [5] – необходимо определить оценки заведомо известных пар;
2. *Трехклассовый* – необходимо извлечь оценочные отношения из документа.

Описание подхода. Основное предположение о наличии оценочного отношения между парой сущностей в тексте документа – относительно короткое расстояние между ними. *Контекст* – ограниченный по длине фрагмент предложения,

содержащий не менее двух именованных сущностей, в котором выделена пара $\langle e_s, e_o \rangle$ сущностей «субъект→объект». Таким образом, для некоторой пары сущностей можно составить множество контекстов. Контекст рассматривается как *оценочный*, если соответствующая пара $\langle e_s, e_o \rangle$, для которой такой контекст был составлен, присутствует в разметке документа. В противном случае контекст считается *нейтральным*.

Таким образом, процесс извлечения оценочных отношений может быть сведен к классификационной задаче на уровне контекстов с последующим отображением контекстных отношений на уровень документа. Оценка контекста с выделенным в нем парой $\langle e_s, e_o \rangle$ может быть отрицательной (neg), положительной (pos), или *нейтральной* (neu). Для отображения контекстных отношений на уровень документов используется вычисление среднего значения среди полученных оценок по всем контекстам рассматриваемого отношения *методом голосования* [25].

Обработка и извлечение контекстов. Пример обработки контекстов для подачи на вход языковым моделям BERT приведен на рис. 3. Входная последовательность может состоять из одной (TextA) или двух последовательностей (TextA+TextB), соединенных разделителем. Если основная часть (TextA) используется для форматированного представления исходного контекста, то дополнительная последовательность TextB может быть использована для передачи вспомогательной информации. В работе рассмотрены следующие форматы входных последовательностей [12]:

- C – использование последовательности без разделения (TextA);
- QA – дополнение TextA вопросом в TextB;
- NLI – дополнение TextA выводом отношения по контексту в TextB.

В случае нейронных сетей используется контекст без добавления вспомогательной информации. Для контекста применяются дополнительные преобразования: лемматизация термов, разметка знаков препинания, разметка вхождений фреймов [26, 27]. В целях устранения возможности принятия решения моделями на основе статистики и значений сущностей контекста, применяется *маскирование сущностей*. Используются следующие типы масок: $\underline{\text{E}}_{subj}$ (субъект и его синонимы), $\underline{\text{E}}_{obj}$ (объект и его синонимы), и $\underline{\text{E}}$ для остальных сущностей.

Таблица 7 приводит количественные данные для извлеченных контекстов[12] из коллекций RuSentRel и RuAttitudes. В опосредованном обучении используются две версии автоматически размеченных корпусов: 2017-Base и 2017-Large.

Оценка качества разметки. Для некоторого документа коллекции, оценка качества разметки основана на подсчете метрик точности (P), полноты (R), и F_1-меры для каждого из оценочных классов в отдельности. Для оценки результата

[12]Параметр, отвечающий за максимально допустимое расстояние в термах между участниками отношений контекста [26, 27] не рассматривается, так как такое ограничение оказывает влияние на результирующую разметку и отсутствие в ней некоторых контекстов.

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

↓

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

Рис. 3: Пример обработки контекста в последовательности (TextA) и представлений вспомогательной информации (TextB) для подачи на вход языковым моделям BERT; для TextB используются форматы: задание вопроса (QA), вывод по контексту (NLI)

Таблица 7: Число контекстов, извлеченных на этапе подготовки данных из коллекций RuAttitudes и обучающего множества коллекции RuSentRel; максимально допустимое число термов в контексте ограничено значением 50.

Коллекция	pos	neg	neu
RuSentRel (обучающее множество)	551	727	6530
RuAttitudes (2017-Base)	38809	55725	4723
RuAttitudes (2017-Large)	123281	161275	15429

на множестве документов размера n фиксируется показатель F^{PN}_{1-mean}, который в свою очередь основан на вычислении макро-усреднений $F_{1-macro}$ над документами по каждому из оценочных классов в отдельности:

$$F^{\text{pos}}_{1-macro} = \frac{1}{n}\sum_{i=1}^{|n|} F^{\text{pos}}_{1}(i) \qquad F^{\text{neg}}_{1-macro} = \frac{1}{n}\sum_{i=1}^{|n|} F^{\text{neg}}_{1}(i) \tag{5}$$

$$F^{PN}_{1-mean} = \frac{(F^{\text{pos}}_{1-macro} + F^{\text{neg}}_{1-macro})}{2} \cdot 100 \tag{6}$$

Форматы обучения моделей. Обучение моделей проводилось в следующих режимах:

1. *Обучение с учителем* – обучение на составленных контекстах оценочных отношений ручной разметки коллекции RuSentRel обучающего множества документов (см. таблицу 6);
2. *Применение опосредованного обучения* – обучение моделей на основе контекстов оценочных отношений коллекций RuSentRel (обучающее множество документов) и RuAttitudes.

Опосредованное обучение выполняется в форматах:

- *Предобучение с последующим дообучением* – модели изначально обучались с использованием опосредованного обучения RuAttitudes, после которого следует *дообучение* контекстами коллекции RuSentRel;

- *Объединенное обучение* – процесс обучения с объединенным набором данных коллекций RuAttitudes и RuSentRel (только нейронные сети);

Перед обучением применяется балансировка данных по числу контекстов классов тональности *методом дублирования* (от англ. Oversampling) для достижения объема, равного числу контекстов наибольшего класса[13].

Для объединенного обучения, алгоритм объединения зависит от формата оценки моделей. При *кросс-валидационном*, в каждом разбиении объединяется коллекция RuAttitudes с каждым обучающим блоком RuSentRel. При *фиксированном*, обучающий набор представляет собой комбинацию RuAttitudes с фиксированным обучающим множеством документов RuSentRel.

Параметры обучения моделей. Для нейронных сетей измерение средних значений *точности* проводилось каждые 5 эпох. Оценка моделей производится на основе результатов последней эпохи обучения. Процесс обучения завершается в случае превышения лимита в 200 эпох. Для избежания проблем переобучения моделей предусмотрено использование механизма *dropout*. В качестве параметров нейронных сетей используются настройки работы [26]. Выбор коэффициента скорости обучения зависел от формата обучения: 0.1 (объединенное и предварительное обучение), 0.01 (дообучение).

Предварительное обучение языковых моделей составляет 5 эпох. В качестве настроек обучения языковой модели используются параметры по-умолчанию [7], за исключением параметра прогрева модели (применение повышенной скорости обучения на начальном этапе). Значение такого коэффициента равно 1 на этапе предобучения модели, и 0.1 на этапе дообучения (по-умолчанию). Ограничение по длине входной последовательности выбрано в 128 токенов. Такое ограничение позволяет покрыть ≈95% примеров без проведения усечений длин контекстов.

5.3 Описание моделей и результаты их применения

Список моделей нейронных сетей, выбранных для экспериментов:

- **CNN**, **PCNN** – модели сверточных нейронных сетей [28];
- **AttCNN**$_e$**, AttPCNN**$_e$ – модели с кодировщиками на основе механизма внимания; e указывает на применения участников отношения ($\underline{\text{E}}_{obj}$, $\underline{\text{E}}_{subj}$) в качестве аспектов в механизме внимания [26];
- **LSTM**, **BiLSTM**, **Att-BLSTM** [26] – модели с кодировщиками на основе рекуррентных нейронных сетей LSTM [29].

Список используемых языковых моделей: mBERT [7], RuBERT, SentRuBERT. Обучение с учителем и дообучение моделей исследовалось форматов $\{\text{C}, \text{NLI}, \text{QA}\}$, рассмотренных в п. 5.2. Предобучение моделей выполнялось только контекстах, представленных в формате NLI (далее обозначено как NLI_P).

[13] В случае объединенного обучения на коллекциях RuSentRel и RuAttitudes, балансировка применяется после объединения извлеченных контекстов обеих коллекций.

Таблица 8: Результаты применения опосредованного обучения для моделей с кодировщиками на основе сверточных и рекуррентных нейронных сетей, а также моделей с механизмом внимания; результаты обучения с учителем отмечены прочерком в колонке «Версия RA»; наилучший результат по каждой модели выделен жирным шрифтом; результаты опосредованного обучения, превосходящие аналогичные при обучении с учителем отмечены подчеркиванием

		Дообучение				Объединенное обучение			
		Двухклассовая		Трехклассовая		Двухклассовая		Трехклассовая	
Модель	Версия RA	$F1^a_{cv}$	$F1_t$	$F1^a_{cv}$	$F1_t$	$F1^a_{cv}$	$F1_t$	$F1^a_{cv}$	$F1_t$
CNN	2017-Large	**68.2**	**69.8**	28.6	**36.1**	**70.0**	**74.3**	**32.8**	39.6
CNN	2017-Base	67.0	66.8	**29.8**	33.1	62.8	67.2	31.1	**40.3**
CNN	—	63.6	65.9	28.7	31.4	63.6	65.9	28.7	31.4
PCNN	2017-Large	66.1	**70.8**	29.8	32.1	**69.5**	70.5	31.6	**39.7**
PCNN	2017-Base	**66.9**	69.4	**30.5**	**33.6**	65.8	**71.2**	**31.9**	38.3
PCNN	—	64.4	63.3	29.6	32.5	64.4	63.3	29.6	32.5
LSTM	2017-Large	**69.9**	**70.4**	**30.5**	**33.7**	**68.0**	**75.4**	31.6	**39.5**
LSTM	2017-Base	66.1	64.6	27.6	32.7	65.2	69.9	31.5	37.2
LSTM	—	61.9	65.3	27.9	31.6	61.9	65.3	27.9	31.6
BiLSTM	2017-Large	62.1	64.0	28.4	**35.4**	**71.2**	68.4	32.0	38.8
BiLSTM	2017-Base	**65.6**	66.4	28.0	31.8	68.0	68.4	32.0	**39.5**
BiLSTM	—	62.3	**71.2**	**28.6**	32.4	62.3	**71.2**	28.6	32.4
AttCNN$_e$	2017-Large	**65.9**	**67.5**	28.0	**35.0**	66.8	**72.7**	30.9	**39.9**
AttCNN$_e$	2017-Base	62.6	65.7	**28.4**	33.5	**68.0**	69.2	**31.3**	37.6
AttCNN$_e$	—	65.0	66.2	27.6	29.7	65.0	66.2	27.6	29.7
AttPCNN$_e$	2017-Large	**66.8**	69.6	27.9	32.5	**70.2**	67.8	**32.2**	**39.9**
AttPCNN$_e$	2017-Base	63.3	**69.9**	**30.0**	34.8	68.2	68.9	31.8	38.9
AttPCNN$_e$	—	64.3	63.3	29.9	32.6	64.3	63.3	29.9	32.6
IAN$_e$	2017-Large	64.5	65.7	28.5	30.7	69.1	**72.6**	30.7	36.7
IAN$_e$	2017-Base	64.5	**66.9**	27.5	**33.5**	**69.8**	70.6	30.7	36.7
IAN$_e$	—	60.8	63.5	**30.8**	32.2	60.8	63.5	**30.8**	32.2
Att-BLSTM	2017-Large	**70.3**	67.0	**28.8**	33.3	66.2	**71.2**	31.0	37.3
Att-BLSTM	2017-Base	65.7	65.7	28.5	**33.7**	65.7	69.7	**31.8**	**40.1**
Att-BLSTM	—	65.7	68.2	27.5	32.3	65.7	68.2	27.5	32.3
Среднее $\Delta(F1)$	2017-Large	**+5.3%**	**+4.1%**	+0.4%	**+6.5%**	**+8.7%**	**+8.9%**	**+10.6%**	**+23.4%**
Среднее $\Delta(F1)$	2017-Base	+3.0%	+1.8%	+0.4%	+6.0%	+5.3%	+5.5%	+10.1%	+22.2%
Среднее	—	63.5	65.9	28.8	31.8	63.5	65.9	28.8	31.8

Результаты фиксировалась в следующих форматах:

- $F1^a_{cv}$ – усредненный показатель F^{PN}_{1-mean} в рамках 3-кратной кросс-валидационной проверки; разбиения проведены с точки зрения сохранения одинакового числа предложений в каждом из них;
- $F1_t$ – показатель F_{1-mean} на тестовом множестве (см. таблицу 6).

Для результатов моделей, обученных с применением опосредованного обучения ($F1_a$) и обучения с учителем ($F1_b$), оценка прироста качества опосредованного обучения подразумевает вычисление процентного соотношения по формуле:

$$\Delta(F1) = \left(\frac{F1_a}{F1_b} - 1\right) \cdot 100 \tag{7}$$

Результаты нейронных сетей. В таблице 8 представлены результаты

Таблица 9: Результаты применения опосредованного обучения для моделей BERT с различными представления контекста TextB; символ «P» указывает тип предобученной модели; результаты обучения с учителем отмечены прочерком в колонке «Версия RA»; наилучший результат по каждой модели выделен жирным шрифтом; результаты опосредованного обучения, превосходящие аналогичные при обучении с учителем отмечены подчеркиванием

		Дообучение			
		Двухклассовая		Трехклассовая	
Модель	Версия RA	$F1^a_{cv}$	$F1_t$	$F1^a_{cv}$	$F1_t$
mBERT (NLI$_P$ + C)	2017-Large	68.9	67.7	**30.5**	31.1
mBERT (NLI$_P$ + C)	2017-Base	**72.9**	**71.5**	30.3	**37.6**
mBERT (C)	—	67.0	68.9	26.9	30.0
mBERT (NLI$_P$ + QA)	2017-Large	69.6	65.2	**30.1**	**35.5**
mBERT (NLI$_P$ + QA)	2017-Base	**74.4**	**71.4**	29.5	32.4
mBERT (QA)	—	66.5	65.4	28.6	33.8
mBERT (NLI$_P$ + NLI)	2017-Large	**69.4**	68.2	**33.6**	36.0
mBERT (NLI$_P$ + NLI)	2017-Base	69.2	**69.6**	31.1	**37.5**
mBERT (NLI)	—	67.8	58.4	29.2	37.0
RuBERT (NLI$_P$ + C)	2017-Large	**70.0**	**69.8**	35.6	35.4
RuBERT (NLI$_P$ + C)	2017-Base	68.2	68.4	34.9	35.6
RuBERT (C)	—	67.8	66.2	**36.8**	**37.6**
RuBERT (NLI$_P$ + QA)	2017-Large	**69.6**	68.2	34.8	37.0
RuBERT (NLI$_P$ + QA)	2017-Base	68.6	**68.5**	**38.0**	**39.1**
RuBERT (QA)	—	69.5	66.2	32.0	35.3
RuBERT (NLI$_P$ + NLI)	2017-Large	**71.0**	**68.6**	**36.8**	39.9
RuBERT (NLI$_P$ + NLI)	2017-Base	67.0	66.9	36.1	39.4
RuBERT (NLI)	—	68.9	66.4	29.4	39.6
SentRuBERT (NLI$_P$ + C)	2017-Large	70.0	69.8	37.9	**39.8**
SentRuBERT (NLI$_P$ + C)	2017-Base	**70.3**	68.1	**38.5**	39.0
SentRuBERT (C)	—	69.3	65.5	34.0	35.2
SentRuBERT (NLI$_P$ + QA)	2017-Large	69.6	64.2	**38.4**	**41.9**
SentRuBERT (NLI$_P$ + QA)	2017-Base	68.6	67.5	35.5	33.6
SentRuBERT (QA)	—	70.2	67.1	34.3	38.9
SentRuBERT (NLI$_P$ + NLI)	2017-Large	70.2	67.7	**39.0**	38.0
SentRuBERT (NLI$_P$ + NLI)	2017-Base	**70.6**	**69.0**	35.4	**40.6**
SentRuBERT (NLI)	—	69.8	67.6	33.4	32.7
Среднее-$\Delta(F1)$	2017-Large	+1.8%	+3.7%	**+13.5%**	+10.0%
Среднее-$\Delta(F1)$	2017-Base	**+2.8%**	**+4.6%**	+11.4%	**+11.7%**
Среднее	—	68.5	65.7	31.6	35.6

экспериментов для моделей нейронных сетей[14]. Средний результат по всем моделям при обучении с учителем приведен в последнем ряду таблицы. Результаты на фиксированном $F1_t$ выше чем по метрике $F1^a_{cv}$ на 4% в случае двуклассового эксперимента на 10% при трехклассовой классификации. При дообучении моделей, прирост качества варьируется в диапазоне 2-5% и ≈0.4-7% для двух и трех-классовых форматов соответственно. При совместном обучении такой показатель увеличивается двое в случае двуклассовой классификации (5-9%) и более чем в 3 раза при трехклассовой классификации: ≈10.5% по $F1_{cv}$ и ≈23% по метрике $F1_t$. Наибольший прирост качества достигается при использовании RuAttitudes2017$_{\text{Large}}$.

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

Таблица 10: Оценка времени в трехклассовом эксперименте с фиксированным набором документов обучающей части коллекции RuSentRel при различных форматах обучения моделей; для языковых моделей приводится среднее время оценки по различным форматам представления входных данных

Модель	Версия RA	с учителем Время$_{эпох}$	предобучение Время$_{эпох}$	дообучение Время$_{эпох}$
Число используемых GPU		1	2	1
Контекстов в секунду		31	62	31
mBERT	2017-Large	—	$8{:}40{:}14|_{04}$	$00{:}10{:}32|_{14}$
mBERT	2017-Base	—	$2{:}59{:}45|_{04}$	$00{:}10{:}32|_{14}$
mBERT	—	$00{:}10{:}32|_{35}$	—	—
Контекстов в секунду		54	62	54
RuBERT/SentRuBERT	2017-Large	—	$6{:}30{:}11|_{03}$	$00{:}06{:}10|_{7}$
RuBERT/SentRuBERT	2017-Base	—	$2{:}14{:}47|_{03}$	$00{:}06{:}10|_{7}$
RuBERT/SentRuBERT	—/1.0-Base	$00{:}06{:}10|_{12}$	—	—

Языковые модели. Результаты экспериментов приведены в таблице 9. Средний результат по всем моделям при обучении с учителем приведен в последнем ряду таблицы. Сравнивая такие показатели с аналогичными из таблицы 8, замена нейронных сетей на языковые модели повышает качество оценки на 7.8% ($F1^a_{cv}$) в двуклассовом, и на 9.7% ($F1^a_{cv}$) и 12% ($F1_t$) в трехклассовом. Смена формата обучения в языковых моделях на опосредованное оказывает прирост в ≈2-5% в двуклассовом, и 10-13% в случае трехклассового эксперимента. Преимущество в использовании русскоязычно-ориентированных моделей перед mBERT особенно наблюдается в трехклассовых экспериментах: дообученная версия RuBERT показывает наилучший результат при использовании формата NLI для TextB. Модель SentRuBERT (NLI$_P$ + NLI) по качеству разметки сопоставима с качеством нейронных сетей объединенного формата обучения, при этом демонстрируя сохранение результата при переходе от фиксированного на кросс-валидационный формат тестирования (35.6-39.0). Такие результаты на 25% выше аналогичных результатов моделей нейронных сетей. Сохранение высоких оценок при разных форматах разбиения указывают на более высокую стабильность в результирующем состоянии в случае языковых моделей.

Оценка производительности языковых моделей. Обучение моделей проводилось на сервере с двумя процессорами Intel® Xeon® CPU E5-2670 v2 с частотой 2.50Ггц, 80 Гб ОЗУ (DDR-3), с двумя видеоускорителями Nvidia GeForce GTX 1080 Ti (11.2Гб); операционная система Ubuntu 18.0.4; обучение моделей выполнялось в контейнерах Docker версии 19.03.5. Применялись следующие параметры оценки: (1) общее время обучения модели; (2) общее число эпох. Оценка выполнялась при фиксированном формате разбиения документов[15]. В таблице 10 приводится средняя оценка времени по каждому из форматов представления входных данных языковых моделей. Во всех форматах обучения на

[15]Временная оценка при проведении кросс-валидационного тестирования была опущена ввиду схожести оценок по каждому из разбиений.

Таблица 11: Усредненная оценка вероятности внимания по головам языковой модели BERT по каждому из 12 слоев в отдельности для: токенов класса (CLS), разделителей (SEP), участникам отношения, всех сторонних токенов к FRAMES и SENTIMENT в отдельности; наибольшие значения в рядах отмечены жирным шрифтом

	номер слоя											
Группа термов	1	2	3	4	5	6	7	8	9	10	11	12
mBERT												
[CLS]	0.06	0.33	**0.36**	0.29	0.31	0.06	0.04	0.04	0.05	0.06	0.07	0.04
SEP	0.04	0.07	0.06	0.06	0.07	0.09	0.09	0.11	**0.12**	0.09	0.09	0.07
$\underline{\text{E}}_{subj}/\underline{\text{E}}_{obj}$	0.05	0.04	0.04	0.06	0.04	0.06	0.06	0.06	0.06	**0.07**	**0.07**	0.05
прочие→FRAMES	**0.07**	0.03	0.03	0.03	0.03	0.05	0.04	0.05	0.04	0.04	0.03	0.03
прочие→SENTIMENT	**0.08**	0.04	0.03	0.03	0.04	0.05	0.04	0.05	0.05	0.04	0.03	0.04
SentRuBERT												
[CLS]	0.03	0.27	0.33	0.30	**0.39**	0.09	0.02	0.03	0.03	0.05	0.04	0.02
SEP	0.05	0.06	0.03	0.04	0.04	0.15	0.22	**0.39**	0.28	0.29	0.07	0.04
$\underline{\text{E}}_{subj}/\underline{\text{E}}_{obj}$	0.10	0.06	0.07	0.07	0.05	0.06	0.08	0.04	0.06	0.05	0.11	**0.12**
прочие→FRAMES	0.05	0.03	0.03	0.03	0.03	0.04	0.04	0.03	0.05	0.05	**0.07**	0.06
прочие→SENTIMENT	0.06	0.03	0.03	0.03	0.03	0.04	0.04	0.05	0.06	0.06	**0.08**	**0.08**
SentRuBERT-NLI$_P$												
[CLS]	0.03	0.27	**0.36**	0.31	0.34	0.05	0.01	0.02	0.01	0.02	0.02	0.02
SEP	0.06	0.04	0.03	0.05	0.04	0.20	0.20	**0.28**	**0.28**	**0.28**	0.04	0.08
$\underline{\text{E}}_{subj}/\underline{\text{E}}_{obj}$	0.10	0.07	0.08	0.08	0.07	0.07	0.09	0.06	0.07	0.11	**0.28**	0.23
прочие→FRAMES	0.07	0.04	0.04	0.04	0.05	0.06	0.05	0.07	0.07	0.05	**0.10**	0.08
прочие→SENTIMENT	0.08	0.05	0.05	0.04	0.05	0.07	0.06	**0.09**	0.08	0.07	0.08	**0.09**

адаптацию русскоязычных моделей требуется меньше эпох при одинаковых настройках обучения: в 1.3 раза меньше на этапе дообучения, и в 2 раза меньше в остальных случаях. Замена mBERT на RuBERT или SentRuBERT сокращает время обучения в 3.5 раза.

5.4 Анализ влияния предварительного обучения на распределение весов механизма внимания в языковых моделях

Для анализа вклада различных элементов контекста в полученный результат часто производится сравнение весов механизма внимания. Для анализа были выбраны следующие состояния языковых моделей: mBERT, SentRuBERT и SentRuBERT-NLI$_P$ (предобученная версия SentRuBERT коллекцией RuAttitudes2017$_{\text{Large}}$). Среди всего множества контекстов рассматриваются только такие контексты, которые были извлечены дообученной моделью SentRuBERT (NLI$_P$ + NLI) из тестового множества коллекции RuSentRel. Таким образом было проанализировано 1032 контекста. В контекстах дополнительно размечены вхождения лексикона оценочных слов русского языка RuSentiLex [30] (SENTIMENT) и вхождения фреймов (FRAMES).

Для каждого входного контекста длиной в s токенов, вектор весов внимания $a \in R^{l \times h \times s \times s}$ содержит значения каждого слоя, по каждой голове модели BERT (l – число слоев языковой модели; h – число голов). Для произвольного слоя l' и

головы h', матрица $a_{l',h'} \in R^{s \times s}$ описывает веса связей токенов входных данных слоя l' с его выходными данными (токенами следующего слоя):

- [CLS] – класса;
- [SEP] – границ последовательностей;
- [S/O] – участников отношений ($\underline{\text{E}}_{subj}$/$\underline{\text{E}}_{obj}$);
- Группы FRAMES и внимание к ним остальных токенов контекста;
- Группы SENTIMENT и внимание к ним остальных токенов контекста.

Рис. 4 иллюстрирует послойную оценку значений весов внимания к приведенным группам токенов. Средние значения по каждому слою указаны[16] в таблице 11. Следует отметить высокие показатели внимания к токену класса [CLS] на слоях 2-5 до 35-40%. Для SentRuBERT наблюдается повышение внимания на токенах [SEP] (слои 7-10) и [S/O] (на конечных слоях). Также наблюдается повышение внимания к токенам FRAMES и SENTIMENT от прочих токенов на конечных слоях до 7-10%. Применение опосредованного обучения (SentRuBERT-NLI$_P$) повысило внимание к [S/O] на конечных слоях: весовые значения увеличились вдвое при сравнении с SentRuBERT. Отмечается также дополнительное повышение внимания к токенам SENTIMENT и FRAMES от прочих токенов на средних и конечных слоях.

В целях наглядной иллюстрации влияния дообучения, на рис. 5 приведена визуализации весов головы №2 для каждой анализируемой модели BERT, по слоям (слева-направо) 2, 4, 8, 11 следующего примера: *«Ведя такую игру, $\underline{E}_{subj}$ окончательно лишилась доверия $\underline{E}_{obj}$ и стран $\underline{E}$. $\underline{E}_{subj}$ к $\underline{E}_{obj}$ в контексте «$\underline{E}_{subj}$ окончательно лишилась доверия $\underline{E}_{obj}$»*. В модели SentRuBERT-NLI$_P$, среди прочих, наиболее выражен фокус внимания ко вхождениям фреймов «окончательно» и «лишиться доверия» (слой 8).

Заключение

В данной работе предложен подход автоматического построения обучающей коллекции в задаче извлечения оценочных отношений из новостных текстов. Разметка основана на применении двух различных техник выделения оценочных отношений для взаимопроверки результатов. Первая подразумевает автоматическое порождение списка оценочных пар посредством предварительного анализа новостной коллекции. Вторая техника заключается в извлечении оценочных отношений из новостных заголовков на основе лексикона оценочных фреймов. В качестве дополнительного этапа предложен подход автоматической разметки нейтральных отношений. Задача извлечения оценочных отношений рассматривалась как двуклассовая (положительные и отрицательные отношения) и трехклассовая (с введением нейтральных отношений) задачи классификации.

[16] Для усредненных оценок к группам FRAMES и SENTIMENT учитываются только такие контексты, которые содержат хотя бы одно вхождение терма соответствующей группы. В результате 68% контекстов учитывалось в статистике «прочие→FRAMES», и 75% в статистике «прочие→SENTIMENT»

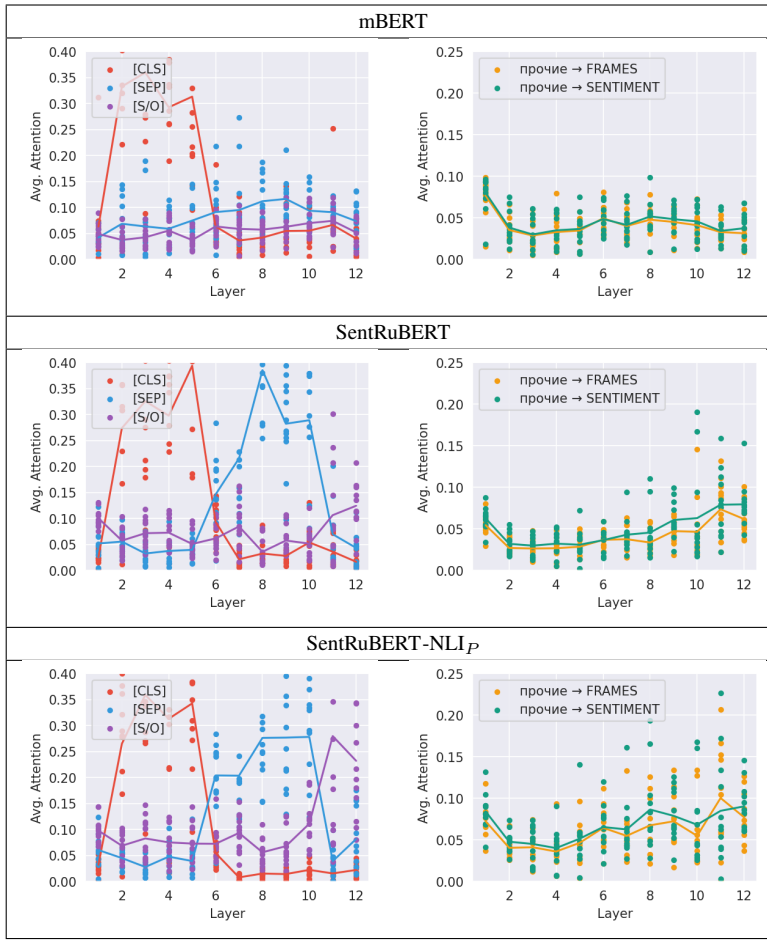

Рис. 4: Послойная оценка распределения внимания языковых моделей BERT к токенам [CLS], [SEP], объектам и субъектам отношения [S/O] (левая колонка) и фреймов и оценочных слов (правая колонка); линиями соединены средние значения весов каждого слоя модели [31]

Применение опосредованного обучения показало наибольший прирост качества в случая трехклассовой классификации. Прирост качества при обучении языковых моделей BERT составил 10-13% по метрике F1 при сравнении с подходом без использования такой коллекции в обучении, и на 25% при сравнении с аналогичными наилучшими результатами моделей сверточных и рекуррентных нейронных сетей.

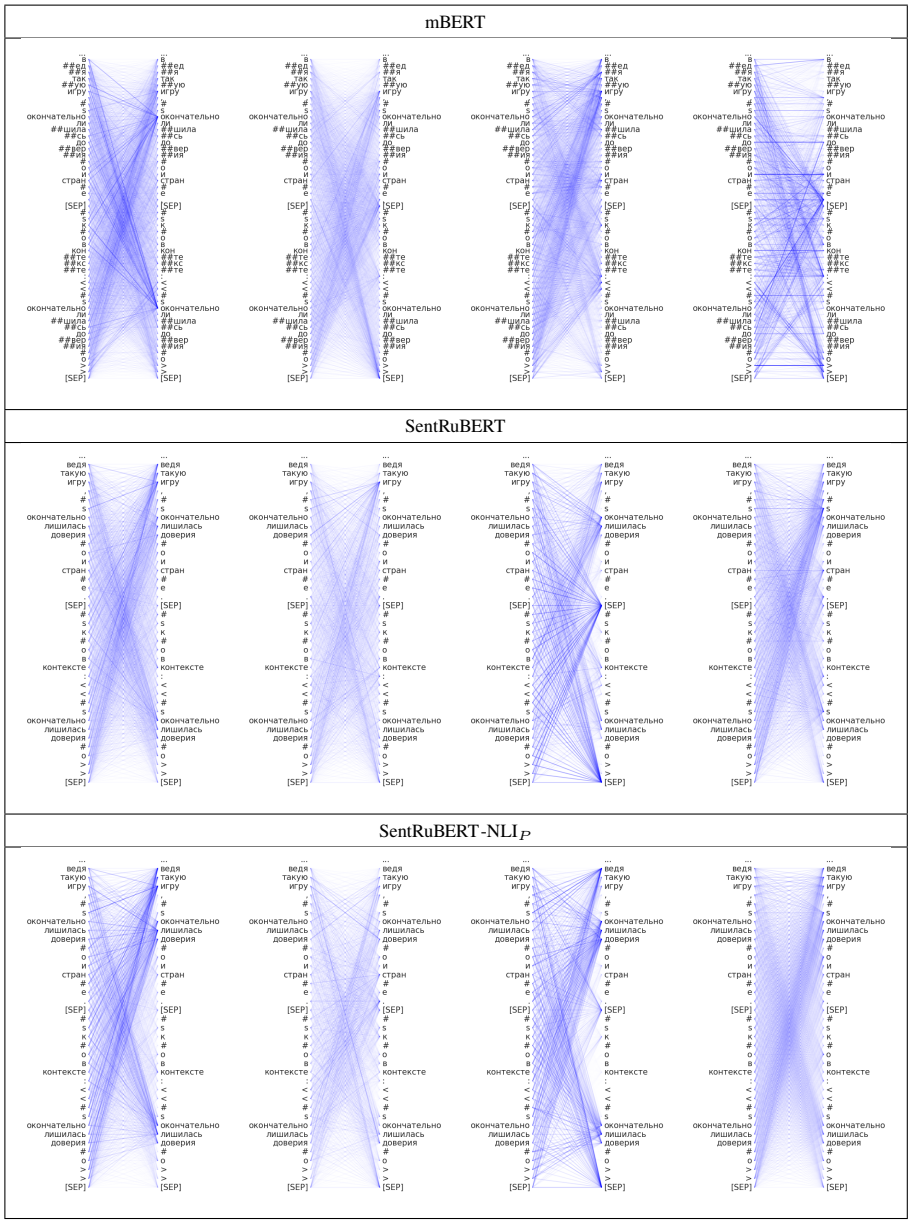

Рис. 5: Пример визуализации [31] весов головы №2 (слои слева-направо: 2,4,8,11) как эволюции внимания модели mBERT в процессе дообучения на примерах SentRuBERT и SentRuBERT-NLI$_P$

Список литературы / References

[1]. N. Loukachevitch и N. Rusnachenko. Extracting sentiment attitudes from analytical texts. *Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialogue-2018 (arXiv:1808.08932)*:459—468, 2018.

[2]. M. Mintz, S. Bills, R. Snow и D. Jurafsky. Distant supervision for relation extraction without labeled data. в *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, страницы 1003—1011. Association for Computational Linguistics, 2009.

[3]. R. Hoffmann, C. Zhang, X. Ling, L. Zettlemoyer и D. S. Weld. Knowledge-based weak supervision for information extraction of overlapping relations. в *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, страницы 541—550, 2011.

[4]. S. Vashishth, R. Joshi, S. S. Prayaga, C. Bhattacharyya и P. Talukdar. RESIDE: improving distantly-supervised neural relation extraction using side information:1257—1266, окт. 2018. url: http://aclweb.org/anthology/D18-1157.

[5]. N. Rusnachenko, N. Loukachevitch и E. Tutubalina. Distant supervision for sentiment attitude extraction. в *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, страницы 1022—1030, Varna, Bulgaria. INCOMA Ltd., сент. 2019. doi: 10.26615/978-954-452-056-4_118. url: https://www.aclweb.org/anthology/R19-1118.

[6]. N. Loukachevitch и N. Rusnachenko. Sentiment frames for attitude extraction in russian. *Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialogue-2020 (arXiv preprint arXiv:2006.10973)*, 2020.

[7]. J. Devlin, M.-W. Chang, K. Lee и K. Toutanova. Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[8]. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser и I. Polosukhin. Attention is all you need. в *Advances in neural information processing systems*, страницы 5998—6008, 2017.

[9]. A. Radford, K. Narasimhan, T. Salimans и I. Sutskever. Improving language understanding by generative pre-training, 2018.

[10]. T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell и др. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[11]. C. Alt, M. Hübner и L. Hennig. Improving relation extraction by pre-trained language representations. *arXiv preprint arXiv:1906.03088*, 2019.

[12]. C. Sun, L. Huang и X. Qiu. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. в *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, страницы 380—385, 2019.

[13]. Y. Kuratov и M. Arkhipov. Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*, 2019.

[14]. S. R. Bowman, G. Angeli, C. Potts и C. D. Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

[15]. A. Conneau, G. Lample, R. Rinott, A. Williams, S. R. Bowman, H. Schwenk и V. Stoyanov. Xnli: evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*, 2018.

[16]. K. Clark, M.-T. Luong, Q. V. Le и C. D. Manning. Electra: pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.

[17]. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer и V. Stoyanov. Roberta: a robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[18]. A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer и V. Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

[19]. M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer и O. Levy. Spanbert: improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64—77, 2020.

[20]. I. Hendrickx, S. N. Kim, Z. Kozareva, P. Nakov, D. Ó Séaghdha, S. Padó, M. Pennacchiotti, L. Romano и S. Szpakowicz. Semeval-2010 task 8: multi-way classification of semantic relations between pairs of nominals. в *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, страницы 94—99. Association for Computational Linguistics, 2009.

[21]. S. Wu и Y. He. Enriching pre-trained language model with entity information for relation classification. в *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, страницы 2361—2364, 2019.

[22]. D. Dowty. Thematic proto-roles and argument selection. *language*, 67(3):547—619, 1991.

[23]. R. Weischedel, M. Palmer, M. Marcus, E. Hovy, S. Pradhan, L. Ramshaw, N. Xue, A. Taylor, J. Kaufman, M. Franchini и др. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23, 2013.

[24]. N. Loukachevitch, G. Lashevich и B. Dobrov. Comparing two thesaurus representations for russian. в *Proceedings of Global WordNet Conference GWC*, страницы 35—44, 2018.

[25]. N. Rusnachenko и N. Loukachevitch. Neural network approach for extracting aggregated opinions from analytical articles. *International Conference on Data Analytics and Management in Data Intensive Domains*:167—179, 2018.

[26]. N. Rusnachenko и N. 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), June 30-July 3, 2020, Biarritz, France*, 2020. doi: 10.1145/3405962.3405985.

[27]. N. Rusnachenko и N. Loukachevitch. Studying attention models in sentiment attitude extraction task. в *Proceedings of the 25th International Conference on Natural Language and Information Systems*, 2020. doi: 10.1007/978-3-030-51310-8_15.

[28]. N. Rusnachenko и N. Loukachevitch. Using convolutional neural networks for sentiment attitude extraction from analytical texts. *EPiC Series in Language and Linguistics*, 4:1—10, 2019.

[29]. S. Hochreiter и J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735—1780, 1997.

[30]. N. Loukachevitch и A. Levchik. Creating a general russian sentiment lexicon. в *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, страницы 1171—1176, 2016.

[31]. K. Clark, U. Khandelwal, O. Levy и C. D. Manning. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019.

Информация об авторах / Information about authors

Николай Леонидович РУСНАЧЕНКО — аспирант кафедры «Теоретической информатики и компьютерных технологий» (ИУ-9) Московского государственного технического университета им. Н.Э Баумана. Окончил МГТУ им. Н.Э. Баумана в 2016 году (специалист). Область научных интересов: обработка естественного языка, анализ тональности сообщений, извлечение отношений.

Nicolay Leonidovich RUSNACHENKO — PhD student of «Theoretical Informatics and Computer Technologies» (IU-9), Bauman Moscow State Technical University (BMSTU) (Moscow, Russia). Graduated from BMSTU in 2016 (master degree). Scientific interests: computational linguistics, sentiment analysis, information retrieval.