DOI:

ПЕРСПЕКТИВЫ ИСПОЛЬЗОВАНИЯ АЛГОРИТМОВ МАШИННОГО САМООБУЧЕНИЯ В ЗАДАЧАХ ОБРАБОТКИ ТЕКСТОВ НА ЕСТЕСТВЕННЫХ ЯЗЫКАХ

Коротеев М.В.
Финансовый университет при Правительстве Российской Федерации, Россия, г. Москва
Ленинградский пр., д.49
mvkoroteev@fa.ru

Аннотация: Данная работа посвящена анализу применимости методов обработки текстов на естественных языках к задаче оценке схожести текстовых последовательностей произвольной длины на русском языкке. Данная задача является частной проблемой, находящей свое применение во многих областях компьютерной лингвистики, в первую очередь — в задачах анализа эффективности компьютерной генерации текстов. При анализе было выявлено преимущество использования метода BERTScore, основанном на семантической схожести текстовых вложений, особенно в части надежности и последовательности обнаружения смыслового сходства предложений по сравнению с традиционными текстовыми метриками, основанными на посимвольном анализе текстов.

Ключевые слова: компьютерная лингвистика, анализ естественных языков, семантический анализ, текстовые вложения, машинное обучение.

Введение

Центральным событием 2019 года в области обработки текстов на естественных языках стало представление новой предобученной модели текстовых вложений BERT, позволяющей достигать невиданных результатов точности во многих задачах автоматической обработки текстов. Эта модель, похоже, заменит по распространенности широко известную модель word2vec, став, фактически, стандартом индустрии. На протяжении всего 2019 года практически все научные статьи, посвященные проблеме обработки текстов на естественных языках, так или иначе являлись реакцией на выход этой новой модели, авторы которой стали одними из самых цитируемых исследователей в области машинного обучения.

Задачи, связанные с обработкой текста на естественных языках включают в себя широкий спектр приложений от разговорных ботов и машинного перевода до голосовых помощников и он-лайн перевода речи. За последние несколько лет эта отрасль пережила бурный рост как количественный, в объеме рыночных приложений и продуктов, так и качественный, в эффективности последних моделей и приближенности к человеческому уровню понимания языка.

1 Теоретические основы используемых методов

1.1 Оценка схожести текстов, основанная на BERT

Задача оценивания качества генерации текстов возникает в процессе решения многих задач, например, машинного перевода. Она может быть сведена к сравнению набора предложений-кандидатов с предложением-образцом в данном текстовом контексте. Однако, наиболее употребимые метрики сходства предложений, например, описанный в 2002 BLEU (bilingual evaluation understudy), концентрируются только на поверхностном сходстве. Упомянутая метрика BLEU, наиболее распространенная при разработке систем машинного перевода, полагается на сравнение пересечения n-грамм текста. При всей простоте, такие метрики упускают лексическое и семантическое разнообразие естественных языков.

Наглядный пример этой проблемы, описанный в 2005 Б. Сатанеевым в [4]: ряд популярных текстовых метрик при данном предложении-референсе “Люди любят заграничные автомобили” отдает предпочтение предложению “Люди любят ездить за границу” по сравнению с очевидно более семантически близким оригиналу “Клиенты предпочитают иномарки”. Как следствие, системы машинного перевода, использующие такие метрики для оценки качества перевода будут предпочитать синтаксически и лексически сходные конструкции, что является субоптимальным в условиях широкого лингвистического разнообразия.

Представление системы BERT [5] позволяет использовать ее как основу для измерения сходства предложений на естественных языках, используя метрику расстояния между текстовыми вложениями сравниваемых предложений.

В общем случае, текстовая метрика, или метрика качества генерации текста - это функция $f(x, \hat{x}) \in \mathbb{R}$, где $x \rightarrow < x_1, x_2, ..., x_k >$ - векторизованное представление предложения-образца, а $\hat{x} \rightarrow < \hat{x}_1, \hat{x}_2, ..., \hat{x}_l >$ - векторизованное представление предложения-кандидата. Хорошая метрика должна как можно ближе отражать суждения человека, то есть показывать высокую корреляцию с результатами оценивания человеком. Все существующие метрики можно условно отнести к одной из четырех категорий: совпадение n-грамм, расстояние правки, сравнение вложений и обученные функции.

Хорошая метрика должна как можно ближе отражать суждения человека, то есть показывать высокую корреляцию с результатами оценивания человеком. Все существующие метрики можно условно отнести к одной из четырех категорий: совпадение n-грамм, расстояние правки, сравнение вложений и обученные функции.

Наиболее распространены текстовые метрики, основанные на подсчете количества n-грамм, встречаемых в обоих предложениях. Чем больше размерность n в n-грамме, тем больше метрика способна уловить сходство целых слов и их порядок, но в то же время тем более эта метрика ограниченная и привязанная к конкретной формулировке и формам слов. К этой категории метрик относятся уже упомянутые BLUE и MeTEOR.

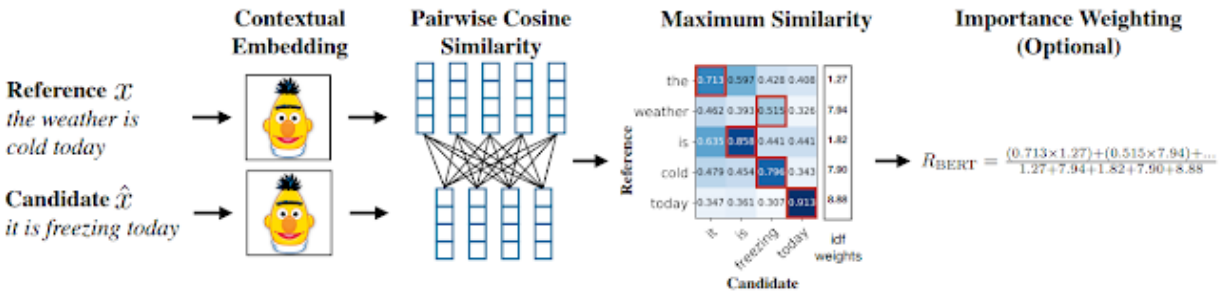
Существуют методы, подсчитывающие близость предложений по количеству правок, переводящих кандидата в референс. Такие методы как TER и ITER учитывают семантическую близость слов и нормализацию грамматических форм. Сюда же можно отнести методы типа PER и CDER, учитывающие перестановку текстовых блоков. Некоторые более современные методы (CharacTER, EED) неожиданно более хороший результат.

В последние годы начали появляться метрики, основанные на использовании словарных вложений, то есть обученных плотных векторизаций слов (MEAN, YISI-1). Преимуществом использования BERT как основы для построения таких метрик является то, что он учитывает контекст слова в рамках его окружения, что позволяет использовать вложения не на уровне отдельных слов, а на уровне всего предложения.

Так как критерием качества текстовой метрики является ее корреляция с человеческими суждениями, неудивительно, что используется машинное обучение для построения обученных метрик, где целевой функцией как раз является такое совпадение. Например, BLEND использует регрессионную модель для взвешивания 29 известных метрик. Недостатком данного подхода можно назвать зависимость от наличия корпуса предразмеченных данных для обучающей выборки, а также риск переобучения на определенной предметной области и, как следствие, утрата обобщающей способности и универсальности.

В феврале 2020 года исследователи из университета Корнуэлла предложили специальный механизм оценки эффективности текстовых моделей, основанных на BERT - BERTScore [2]. BERTScore используется для автоматической оценки качества генерации текстов на естественных языках. Авторы предложения утверждают, что BERTScore лучше коррелирует с человеческими суждениями о качестве текста и, как следствие, может лечь в основу более эффективного и результативного процесса выбора моделей. На сегодняшний день это самая прогрессивная метрика оценки качества текстовой генерации.

В основе BERTScore лежит простой алгоритм подсчета косинусного расстояния между векторизованными представлениями каждого слова в предложении референсе и предложении-кандидате. Расстояния вычисляются попарно, а затем для каждого слова в референсе отбираются наиболее близкие слова в кандидате, эти расстояния усредняются и составляют оценку $R_{BERT} = \frac{1}{|x|} \sum_{x_j \in x} \max_{\hat{x}_j \in \hat{x}} x_i^T \hat{x}_j$ - оценка отзыва. Оценка точности считается аналогично, но инвертировано относительно кандидата и референса - $P_{BERT} = \frac{1}{|x|} \sum_{\hat{x}_i \in \hat{x}} \max_{x_j \in x} x_i^T \hat{x}_j$. В дополнение к этому вычисляется оценка F1 - $F_{BERT} = 2\frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}}$. $R_{BERT} = \frac{1}{|x|} \sum_{x_j \in x} \max_{\hat{x}_j \in \hat{x}} x_i^T \hat{x}_j$ - оценка отзыва. Оценка точности считается аналогично, но инвертировано относительно кандидата и референса - $P_{BERT} = \frac{1}{|x|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_j \in x} x_i^T \hat{x}_j$. В дополнение к этому вычисляется оценка F1 - $F_{BERT} = 2\frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}}$.



Источник: [2]

Кроме базового алгоритма авторы BERTScore используют метрику IDF для определения более редких слов, так как предыдущие исследования текстовых метрик показывают, что редкие слова могут быть более показательными для оценки сходства двух предложений. Метрика IDF, оцененная на референсе и сглаженная (+1) для учета новых слов, используется как вес соответствующей косинусной меры при усреднении.

При сравнительном анализе метрик схожести текста метрики, основанные на BERT показывают стабильно более высокие результаты, нежели классические текстовые метрики. Это значит, что они статистически значимо гораздо ближе к оценкам, данным человеком.

Кроме того, оригинальная статья, представляющая BERTScore, уделяет внимание вопросу производительности. В этом плане BERTScore, конечно, зачастую медленнее, нежели классические модели. Авторами дана оценка сравнения с популярной реализацией SacreBLEU, в котором BERTScore работает примерно в три раза медленнее. Так как для оценки схожести используется предобученная модель BERT, рост точности оценки дается за счет не такого падения скорости, который можно было бы ожидать от настолько более сложной модели. Учитывая типичный размер датасетов обработки естественных текстов, рост времени вычисления во время валидации модели не должен оказывать значительный эффект на производительность системы машинного обучения в целом.

Уже начинают появляться публикации [3], улучшающие оригинальный алгоритм BERTScore за счет распараллеливания вычислительных операций. Представлены вариации классических метрик, использующие BERT, и показывающие улучшение корреляции с человеческими оценками.

Несомненно, направлением будущего развития текстовых метрик будет более широкое применение BERT как основания для оценивания, своеобразного семантического движка. Также перспективным представляется разработка специфичных моделей, учитывающих особенности конкретных предметных областей и повышающих базовый уровень точности за счет специализации. Что касается BERTScore, еще одним его преимуществом является дифференцируемость, что позволит в будущем встроить его в методику обучения текстовых моделей, что обещает дальнейший рост производительности и качества моделей машинного обучения обработки естественных текстов.

1.2 Использование метрик сходства для генерации текстов

Модели машинного обучения очень часто могут быть уязвимы ко входным данным, для человека неотличимыми от реального примера, тем не менне дающих неправильный выходной сигнал. Такие специально подобранные (состязательные) примеры корректно классифицируются экспертом-человеком, но дают неправильный результат в модели, что ставит под сомнение безопасность и надежность применяемых моделей машинного обучения. При этом, было показано, что робастность и обобщаемость моделей может быть существенно повышена путем генерации качественных состязательных примеров и включении их в обучающую выборку.

В настоящее время достигнуты определенные успехи в создании методики построения состязательных примеров в области распознавания изображений и обработки аудио. Однако, область обработки текстов на естественных языках остается недостаточно исследованной из-за специфической природы входных данных. Помимо собственно состязательности, такие текстовые примеры должны обладать дополнительными условиями:

- последовательность человеческой интерпретации - состязательный пример должен восприниматься человеком так же, как и исходный;
- семантическое сходство - состязательный пример должен нести тот же семантический смысл для человека, что и исходный;
- языковая корректность - состязательный пример должен восприниматься человеком как синтаксически корректный.

Существующие методы создания текстовых состязательных примеров включают неправильное написание слов, удаление индивидуальных слов из текста, удаление или вставка целых фраз, и все они генерируют примеры, явно искусственные.

Постановка метода генерации состязательных примеров для задач классификации текстов может быть формализована следующим образом [1]: имея набор предложений $X = x_1, x_2, ..., x_N$ и соответствующий набор меток $Y = y_1, y_2, ..., y_N$, дана предобученная модель $F : X \to Y$, которая ставит в соответствие предложению X метку Y. В таком случае, для произвольного предложения $x \in X$ состязательный пример x_{ADV} должен удовлетворять следующим условиям: $F(x_{ADV}) \neq F(x), Sim(x, x_{ADV}) \geq \epsilon$, где $Sim(\cdot)$ - функция синтаксического и семантического подобия, а ϵ - минимальный уровень сходства между оригинальным и состязательным примером.

При соблюдении условия черного ящика методика генерации состязательных примеров не должна располагать информацией о архитектуре, параметрах и структуре обучающей выборки целевой модели машинного обучения. Она может лишь использовать модель на данном входном примере для получения результата классификации и уровня уверенности.

Дж.Говард, Дж. Жиждинг, Дж Тианиджоу и П. Золовитс из MIT и университета Гонгконга [1] предложили алгоритм построения состязательных текстовых примеров TEXTFOOLER для исследования моделей черного ящика, включая BERT, состоящий из следующих шагов:

- ранжирование важности слов. Практика применения современных текстовых моделей показывает, что лишь некоторые слова служат индикативными факторами при работе моделей, что согласуется с исследованиями BERT Т. Нивена и Х. Као [6], которые показывают важность механизма внимания. Обратим внимание, что этот шаг является тривиальным при условии модели белого ящика и сводится к подсчету градиента выхода модели при изменении индивидуальных слов.
- трансформация слов. На этом шаге для слов с выявленной высокой важностью для модели они заменяются. Замена должна удовлетворять трем условиям: иметь сходное семантическое значение, синтаксически подходить к исходному контексту и заставить модель дать неверное предсказание. На этом этапе можно выделить несколько подэтапов:
  1. выделение синонимов. Выбирается набор синонимов на основе меры косинусного расстояния между исходным и всеми остальными словами в словаре. Здесь могут применяться текстовые вложения на уровне слов.
  2. фильтрация по части речи. Для обеспечения синтаксической связности необходимо оставить только те синонимы-кандидаты, которые совпадают по грамматической роли с исходным словом.
  3. проверка семантического сходства. Для каждого кандидата проводится проверка сходства текстового представления исходного предложения и предложения, в котором исходное слово заменено на него (кандидата в состязательные примеры). Для этого используется модель USE, представляющая универсальное текстовое представление на уровне предложений и косинусная метрика сходства. Необходимо добиться сходства, превышающего заданный порог ϵ.
  4. проверка предсказания модели. Для каждого кандидата в состязательные примеры проводится оценка предсказания модели и, если находятся примеры, изменяющие предсказание модели относительно исходного предложения, они добавляются в результирующий набор. Если нет, то выбирается слово со следующей в порядке убывания важностью, оцененной на первом шаге и процесс повторяется.

Аппробация данного алгоритмы провводилась на пяти общедоступных датасетах для такстовой классификации - AG's news, Fake news detection, MR, IMDB и Yelp. Были обучены три современные модели текстовых представлений: WordCNN, WordLSTM и BERT. Результаты моделирования приведены в таблице:

	WordCNN					WordLSTM					BERT				
	MR	**IMDB**	**Yelp**	**AG**	**Fake**	**MR**	**IMDB**	**Yelp**	**AG**	**Fake**	**MR**	**IMDB**	**Yelp**	**AG**	**Fake**
Original Accuracy	78.0	89.2	93.8	91.5	96.7	80.7	89.8	96.0	91.3	94.0	86.0	90.9	97.0	94.2	97.8
After-Attack Accuracy	2.8	0.0	1.1	1.5	15.9	3.1	0.3	2.1	3.8	16.4	11.5	13.6	6.6	12.5	19.3
% Perturbed Words	14.3	3.5	8.3	15.2	11.0	14.9	5.1	10.6	18.6	10.1	16.7	6.1	13.9	22.0	11.7
Semantic Similarity	0.68	0.89	0.82	0.76	0.82	0.67	0.87	0.79	0.63	0.80	0.65	0.86	0.74	0.57	0.76
Query Number	123	524	487	228	3367	126	666	629	273	3343	166	1134	827	357	4403
Average Text Length	20	215	152	43	885	20	215	152	43	885	20	215	152	43	885

Источник: [1]

При анализе результатов оказалось, что набор данных, состоящий из сгенерированных состязательных примеров может снизить эффективность текстовой классификации с 80-97% до 0-20%. Это говорит об успешности проведенной атаки на модель машинного обучения.

Добавление сгенерированного состязательного набора данных и дообучение модели на нем значительно повышает эффективность моделей классификации текстов, которые показывают на 2-7 процентных пункта более высокую эффективность на тестовом состязательном наборе, нежели без подобного дообучения.

	MR	IMDB	SNLI	MNLI(m)
After-Attack Accu.	11.5/6.2	13.6/11.2	4.0/3.6	9.6/7.9
% Perturbed Words	16.7/14.8	6.1/4.0	18.5/18.3	15.2/14.5
Query Number	166/131	1134/884	60/57	78/70
Semantic Similarity	0.65/0.58	0.86/0.82	0.45/0.44	0.57/0.56

Источник: [1]

Несомненно, исследования устойчивости моделей машинного обучения являются непременным условием широкого внедрения таких интеллектуальных систем в процесс принятия решений, что составляет актуальность данного направления исследований. Как уже было сказано ранее, разработка и изучение различного типа автоматизированных атак на модели машинного обучения могут дать нам направление дальнейшего совершенствования процесса разработки и тестирования интеллектуальных систем в целом, не ограничиваясь моделями классификации.

2 Исследование применимости методики BERTScore для анализа естественных текстов на русском языке

Как известно, анализ естественных текстов имеет сильную языковую специфику. В рамках проводимого авторами и коллегами исследования применимости современных методов машинного самообучения для задач управляемой генерации естественных текстов необходимо решить задачу оценки эффективности генерации текстов. Эта задача предполагает использование метрики близости текстовых последовательностей для оценки эффективности систем генерации.

Подобная система, учитывая последние достижения, описанные в предыдущих главах данной работы, должна строиться на оценке в первую очередь семантического сходства, а не механическом анализе на уровне символов или слов.

Начальный этап исследования предполагает оценку применимости современных методов, в частности, BERTScore, на наборе данных русскоязычных текстовых последовательностей. Для этого мы выбрали корпус текстов NLPDatasets, а именно набор данных перестановочных перефразировок. Эти данные позволят определить применимость метрик семантического расстояния текстовых последовательностей для определения схожести предложений по смыслу. Данный набор данных удобен тем, что он разделен на группы предложений, каждое из которых несет абсолютно идентичный смысл, но грамматически и порядково они имеют существенное различие.

Нами была использована реализация метода BERTScore от авторов методики в виде пакета для языка программирования Python под названием bert_score.

Методика определения интегрального показателя качества метрики семантического сходства следующая:

- Из исходного датасета выбирается одно предложение и сравнивается со всеми остальными. При этом метрика, учитывающая именно семантическое сходство должна показать высокую степень близости референсного предложения с его парным из исходного набора и низкую - со всеми остальными.
- Для данного предложения строится метрика качества по следующей формуле: $Q_i = F_i - \hat{F}$, где F_i - оценка метрики F1 сходства референсного предложения с его парным, $\hat{F}$ - среднее арифметическое метрик F1 сходства данного предложения со всеми остальными. Данный показатель будет приближаться к 1, в случае идеальной классификации и к нулю при отсутствии дискриминации референсного предложения относительно всей выборки.
- Данная процедура повторяется для всех референсных предложений из начального набора данных. Полученные показатели усредняются для оценки средней эффективности данного метода определения семантического сходства текстовых последовательностей.

- Вышеописанная оценка качества производится для трех метрик расстояния между текстовыми последовательностями - BERTScore, нормализованное расстояние Левенштейна и выраженную в долях единицы метрику SequenceMatcher. Последние две метрики являются стандартными способами программно определить схожесть двух текстовых последовательностей произвольной длины.

Для сокращения времени выполнения программы, что наиболее актуально для метрики BERTScore, которая использует глубокую нейронную сеть BERT для представления текстовых эмбеддингов, мы использовали в каждом тестовом примере 20 предложений из начального датасета для сравнения с каждым референсным предложением так, чтобы эти 20 примеров обязательно включали в себя парное к референсному. Таким образом было проведено 500 сравнений различных референсных предложений для модели BERTScore и 11000 для символьных моделей. Результаты сравнения можно увидеть в таблице 1.

	BERT	Levenstein	SM
	1	2	3
Среднее	0,24995	0,01146	0,37585
Дисперсия	0,00082	0,00007	0,01320
Минимальное	0,18985	-0,00858	0,10730
Максимальное	0,33478	0,03738	0,68553
Время выполнения для 100 примеров, сек	0,56182	0,21429	

Таблица 1. Результаты сравнения эффективности различных методов оценки расстояния между текстовыми последовательностями.

Необходимо обратить внимание, что для анализа текстов на русском языке использовалась модель, основанная на предобученной нейронной сети BERT Multilingual. Дополнительного дообучения на корпусе русских текстов не производилось, так как целью исследования была в первую очередь оценка нативной производительности данной модели, без повышения ее производительности путем специализированного дообучения на корпусе текстов из целевой предметной области.

Наиболее показательным по сути является показатель среднего отклонения расстояния референсного предложения и его парафраза и среднего расстояния референсного предложения от 20 предложений из начального датасета. В данном случае сравнивается способность метрики расстояния определить парафраз предложения на фоне других, не имеющих с ним ничего общего как по смыслу, так и по форме. Чем больше этот показатель, тем большую предиктивную силу имеет метрика.

Также нами была оценена дисперсия этого показателя на массива оцененных референсных предложений. Данное значение оценивает надежность и последовательность работы метрики, ее независимость от формы и смысла предложения. Чем меньше этот показатель, тем больше надежность метрики.

Минимальное и максимальное значение отклонения характеризую разброс значений. Здесь интерес представляют те случаи, когда минимальное значение оказывалось меньше нуля. Это значит, что данная метрика хотя бы в одном случае полность провалила задачу определения парафраза предложения, то есть оценила его расстояние от референсного выше, чем от какого-то другого, не относящегося к нему.

Кроме того интерес с точки зрения практического применения данных методов представляет время выполнения сравнения, которое характеризует производительность и скорость метрики. Здесь нужно сказать, что сравнения по методу BERTScore производились на облачном сервисе Google Colab, обработка данных производилась на CPU. Расчет метрик 2 и 3, то есть основанных на посимвольном анализе последовательностей производилась на компьютере с процессором AMD Ryzen 5 1600, 16GB DDR4.

3 Выводы и благодарности

При анализе результатов численного эксперимента можно сделать ряд существенных выводов о применимости методов оценки расстояния между текстовыми последовательностями, использующими грамматический и семантический подходы. В первую очередь необходимо отметить значительно более

высокую вычислительную сложность семантического подхода. Как уже было сказано, любые модели и методы анализа естественных текстов, использующие модель BERT требуют существенно более серьезных вычислительных мощностей, желательно включающих возможность расчета на GPU или TPU.

Что касается непосредственной эффективности, можно видеть, что метрика, основанная на вычислении расстояния Левенштайна не дает должного уровня дискриминации парафразов, что видно по существенно более низним значениям среднего отклонения. Более того, в ряде случаев данная метрика вообще дает отрицательные отклонения, что свидетельствует о возможной выдаче более низких значений расстояния для предложений, схожих по смыслу и различных по форме, нежели для предложений, схожих по форме и различных по смыслу.

Что касается метода 3 (SequenceMatcher), то в среднем он дает более четкую дифференциацию парафраза референсного предложения, чем даже семантический метод BERTScore. Однако, в сравнении с ним, он показывает гораздо более высокую дисперсию результатов, что говорит о серьезной зависимости надежности этого метода от формы и грамматического сходства сравниваемых предложений. Напротив, метод BERTScore при более консервативном среднем результате имеет очень низкую дисперсию, за счет анализа именно текстовых вложений, а не символьной схожести текстовых последовательностей.

Сразу после своего появления, модель BERT получила интенсивную реакцию научного сообщества и на сегодняшний день применяется практически во всех задачах обработки текстов. Почти сразу появились рассмотренные в данной работе предложения по совершенствованию модели, приведшие к улучшению результатов ее применения во всех последующих задачах. С учетом всего сказанного, можно уверенно утверждать, что BERT явил собой качественный скачок в отрасли интеллектуальной обработки естественных языков и закрепил превосходство использования предобученных на огромных наборах данных моделей представления текста как универсальной основы для построения интеллектуальных алгоритмов решения конкретных задач.

Несомненно, мы увидим еще немало новых научных результатов, основанных на применении и адаптации модели BERT к различным задачам обработки текстов на естественных языках. Дальнейшее совершенствование архитектуры нейронной сети, вкупе с тонкой подгонкой процедуры и параметров обучения неизбежно приведет к существенному улучшению многих компьютерных алгоритмов NLP, от классификации и аннотирования текстов, до машинного перевода и вопросно-ответных систем.

Автор выражает признательность коллективу Департамента анализа данных и машинного обучения Финансового университета при Правительстве Российской Федерации за консультативную поддержку исследования.

Литература

1. *Di Jin, Zhijing Jin, Joey Tianyi Zhou, и Peter Szolovits.* 2019. Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment. Извлечено 20 март 2020 г. от http://arxiv.org/abs/1907.11932
2. *Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, и Yoav Artzi.* 2019. BERTScore: Evaluating Text Generation with BERT. Извлечено 20 март 2020 г. от http://arxiv.org/abs/1904.09675
3. *Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, и Steffen Eger.* 2019. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). DOI:https://doi.org/10.18653/v1/d19-1053
4. *Satanjeev Banerjee и Alon Lavie.* 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. В Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 65–72. Извлечено 20 март 2020 г. от https://www.aclweb.org/anthology/W05-0909.pdf
5. *Jacob Devlin, Ming-Wei Chang, Kenton Lee, и Kristina Toutanova.* 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Извлечено 20 март 2020 г. от http://arxiv.org/abs/1810.04805
6. *Timothy Niven и Hung-Yu Kao.* 2019. Probing Neural Network Comprehension of Natural Language Arguments. Извлечено 20 март 2020 г. от http://arxiv.org/abs/1907.07355