НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
июль–август 2023 Том 23 № 4 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
July–August 2023 Vol. 23 No 4 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2023-23-4-743-749
УДК 004.89

Метод аугментации текстовых данных с сохранением стиля речи и лексики персоны

Анастасия Андреевна Матвеева[1]✉, Олеся Владимировна Махныткина[2]

[1,2] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] anastasiamatveevaitmo@gmail.com✉, https://orcid.org/0000-0002-2935-991X
[2] makhnytkina@itmo.ru, https://orcid.org/0000-0002-8992-9654

Аннотация

Введение. В настоящее время часто для различных задач обработки естественного языка требуются большие наборы данных. Однако для многих задач сбор большого набора данных является трудоемким, дорогостоящим процессом и требует привлечения экспертов. Увеличение объема информации возможно достичь с использованием методов аугментации данных. Вместе с тем использование классических подходов может привести к включению в корпус данных фраз, которые отличаются по стилю речи и лексикону целевой персоны. Это сопровождается изменением целевого класса и появлением реплик с неестественным использованием лексики и отсутствием смысла. Предложен новый метод аугментации тестовых данных, учитывающий стиль и лексикон персоны. **Метод.** В работе разработан новый метод аугментации текстовых данных, сохраняющий индивидуальные речевые характеристики и словарный запас. Основная идея метода заключается в формировании индивидуальных шаблонов для каждого человека на основе анализа синтаксических деревьев высказываний и последующего создания новых реплик по сформированным шаблонам. **Основные результаты.** Метод апробирован на задаче оценки эмоционального состояния пользователя в диалоге. Исследования проведены для датасетов на английском и русском языках. Предложенный метод позволил повысить качество решения этих задач как для английского, так и для русского языков. Получено увеличений значений метрик accuracy и weighted F1 для разных моделей. **Обсуждение.** Результаты работы могут быть применены для повышения метрик accuracy и weighted F1 моделей, предназначенных для решения различных задач для английского и русского языков.

Ключевые слова

аугментация текстовых данных, распознавание эмоций, оценка валентности высказываний

Благодарности

Исследование выполнено за счет гранта Российского научного фонда (№ 22-11-00128, https://www.rscf.ru/project/22-11-00128/).

Ссылка для цитирования: Матвеева А.А., Махныткина О.В. Метод аугментации текстовых данных с сохранением стиля речи и лексики персоны // Научно-технический вестник информационных технологий, механики и оптики. 2023. Т. 23, № 4. С. 743–749. doi: 10.17586/2226-1494-2023-23-4-743-749

Text augmentation preserving persona speech style and vocabulary

Anastasia A. Matveeva[1]✉, Olesia V. Makhnytkina[2]

[1,2] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] anastasiamatveevaitmo@gmail.com✉, https://orcid.org/0000-0002-2935-991X
[2] makhnytkina@itmo.ru, https://orcid.org/0000-0002-8992-9654

Abstract

Currently, various natural language processing tasks often require large data sets. However, for many tasks, collecting large datasets is quite tedious and expensive, and requires the involvement of experts. An increase in the amount of data can be achieved using methods of data augmentation, however, the use of classical approaches can lead to the inclusion of phrases in the data corpus that differ in the speech style and vocabulary of the target person, which can lead to both

a change in the target class as well as the appearance of replicas with unnatural vocabulary use and lack of meaning. In this context, a new method for test data enrichment is proposed that takes into account the person's style and vocabulary. In this article, a new method for expanding text data that preserves individual language features and vocabulary is proposed. The core of the method is to create individual templates for each person based on the analysis of syntactic trees of propositions and then to create new replicas according to the generated templates. The method was tested on the task of assessing the user's emotional state in a dialogue. The search was carried out for data sets in English and Russian. The proposed method made it possible to improve the quality of solving these problems for both the English and Russian languages. Up to a 2 % increase in accuracy and weighted F1 metrics has been noted for various models. The results of the work can be applied to improve the accuracy and weighted F1 metrics of models designed to solve various problems for the English and Russian languages.

Keywords
text data augmentation, emotion recognition, statement valence evaluation

Acknowledgements
This research was supported by a grant from the Russian Science Foundation (22-11-00128 https://www.rscf.ru/project/22-11-00128/).

For citation: Matveeva A.A., Makhnytkina O.V. Text augmentation preserving persona speech style and vocabulary. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 4, pp. 743–749 (in Russian). doi: 10.17586/2226-1494-2023-23-4-743-749

Введение

В настоящее время активно развивается направление обработки естественного языка. В основном повышение качества решения задач происходит за счет появления более сложных и глубоких архитектур нейронных сетей, требующих больших датасетов для обучения. Использование предобученных моделей и их дообучение на целевых датасетах зависит от объема целевых датасетов. Для получения качественных моделей необходимо наличие больших датасетов. Увеличение набора данных возможно с использованием различных методов аугментации текста.

Аугментация текстовых данных может осуществляться на двух уровнях: текста и векторных представлений слов и реплик.

Аугментация на уровне векторных представлений слов и реплик подразумевает модификацию не текстовых сообщений, а векторов слов, полученных с использованием различных языковых моделей. В работах [1, 2] рассмотрены методы аугментации на уровне векторных представлений слов: интерполяции, экстраполяции, добавления случайного шума. Различные варианты добавления шума в векторные представления слов, полученные с использованием модели Word2Vec, исследованы в работе [3]: гауссовский шум, шум Бернулли, враждебный шум (Adversarial Noise) и т. д. Исследования добавления шума в векторные представления слов, приведенные в [1–3], в целом показали неплохие результаты, однако использование таких методов аугментации подразумевает использование эмбеддингов не существующих слов, что потенциально может привести к рассогласованию с метками классов сообщений, нарушению семантической целостности, изменению лексики и стиля.

Аугментацию текстовых данных на уровне слов и реплик можно условно подразделить на методы, способные к сохранению стиля и словарного состава текста, и методы неспособные к их сохранению. Методы аугментации данных на уровне текста, неспособные к сохранению стиля и словарного состава, могут включать в себя простые методы редактирования текста. Например, набор простых техник аугментации представлен в алгоритме EDA [4], который состоит из четырех операций: замена синонимами; случайные вставка, перестановка и удаление. При аугментации текстовых данных важно сохранение смысла текста, в связи с этим часто используются: замена слов на синонимы; различные словари (например, WordNet [1, 4]) или предобученные языковые модели (BERT [5, 6], GPT2 [6], Word2Vec [2], Glove [1]) и другие.

Основываясь на гипотезе о том, что предложения являются естественными, даже когда слова в предложениях заменяются другими словами с парадигматическими отношениями, в работе [7] предложен подход контекстного дополнения. Еще одним подходом может быть использование обратного перевода вместо создания парафраз.

В [8] рассмотрена аугментация данных посредством перевода сообщений с английского на французский и с французского на английский языки для датасета Stanford Question Answering Dataset. В работе [9] использовано несколько этапов машинного перевода с помощью Google Translate[1]. При этом максимальная длина цикла перевода задавалась до трех языков, например, английский → немецкий → датский → английский. Исследование глубоких нейронных сетей на предмет эффективности обратного перевода привело к положительным результатам в работе [10]. Нетривиальная техника аугментации представлена в [11], которая выполнена путем перемешивания текста с использованием нейронной сети. Однако ни один из вышеперечисленных методов не сохраняет исходный стиль сообщения и лексику.

При изменении текстовых данных преобразования могут вносить искажения в текст, делая его грамматически или семантически неверным, или стилистически отличным от исходного текста. По этой причине требуются приемы, которые могут аугментировать текст, при этом обеспечив сохранение стиля, лексики и синтаксической целостности.

[1] Google Translate [Электронный ресурс]. Режим доступа: https://cloud.google.com/translate (дата обращения: 14.09.2022).

Существуют также методы аугментации, способные сохранять стиль и словарный состав текста, например, генератор парафраз, основанный на преобразовании синтаксических деревьев [12]. В этом методе текст модифицируется путем преобразования синтаксического дерева на основе общеупотребимых синтаксических грамматик. Аугментация текста с помощью синтаксических деревьев с генерацией новых данных на основе синтаксических шаблонов рассмотрена в [13, 14]. Данные методы действительно позволяют сохранить словарный состав, обеспечить сохранение семантической целостности, но не позволяют сохранить синтаксические особенности речи, в связи с чем стиль речи может быть значительно искажен.

Описание метода аугментации

В настоящей работе предложен новый метод аугментации, сохраняющий уникальные стиль речи и лексику (рисунок). Под стилем речи подразумевается совокупность отличительных языковых признаков речи персоны, например, синтаксические признаки (сохранение синтаксических структур, наиболее часто используемых персоной) и морфологические признаки (сохранение частотности использования определенных частей речи). Данный метод основан на схеме аугментации данных, предложенной в работе [12]. Важно, что этап преобразования и генерации перефразированных данных в предлагаемом методе — единый процесс. В качестве синтаксических шаблонов использованы синтаксические структуры, которые характерны для речи рассматриваемой персоны, что позволяет сохранить синтаксические особенности речи этой персоны и, таким образом, обеспечить высокую вероятность сохранения стиля речи. Отметим, что метод адаптирован для работы с данными на русском языке.

Пусть дано n персон, и каждая персона имеет m (для каждой персоны свое число) реплик $A_0 \ldots A_m$. Для каждой реплики извлекается синтаксическое дерево S_i с помощью парсера Stanford Core NLP для английского и русского языков [15]. Получим синтаксическое дерево предложения «This is a test»:

```
(ROOT
  (S
    (NP (DT This))                    (1)
      (VP (VBZ is)
        (NP (DT a) (NN test))
      )))
```

На основе синтаксического дерева (1) с помощью изъятия всех слов из исходного предложения и сохранения только синтаксических структурных единиц, получим синтаксический шаблон T_j. Синтаксический шаблон для предложения «This is a test» имеет вид:

$$(ROOT(S(NP(DT))(VP(VBZ)(NP(DT)(NN))))), \quad (2)$$

где ROOT — корень предложения; DT — определитель; VP — глагольная фраза; VBZ — глагол настоящего времени 3-го лица единственного числа; NP — именная группа; NN — существительное, нарицательное, единственное число.

Для каждой персоны $P_i \ldots P_n$ и множества принадлежащих ей синтаксических шаблонов (2), составим частотность использования, т. е. создадим пару наборов шаблонов T_i и f (количество использований данного шаблона и похожих на него более чем на 96 % шаблонов, при этом похожие шаблоны удаляются из множества). Из оставшегося множества выберем r наиболее встречающихся шаблонов.

К извлеченным наборам шаблонов, полученных для каждой персоны, применим предобработку для получения формата, принимаемого OpenAttacker Syntactically Controlled Paraphrase Network (SCPN) — в конец строки добавим тег окончания строк (End-Of-String, EOS).

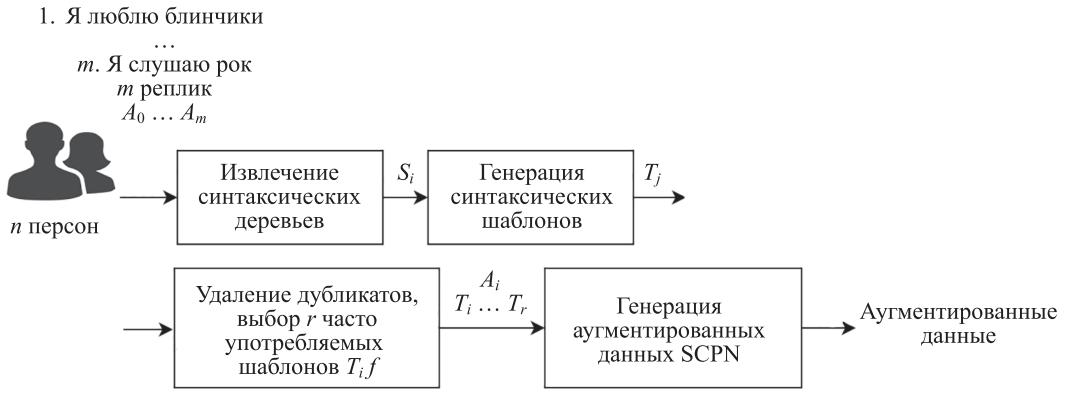

Рисунок. Блок-схема метода аугментации.

A_0 — первая реплика персоны; S_i — синтаксическое дерево; A_i — одна из реплик персоны; T_j — синтаксический шаблон для реплики A_i, где $j = i$; f — частота использования синтаксического шаблона; r — настраиваемый параметр количества вариантов аугментации одной реплики; T_i — один из синтаксических шаблонов после очистки дубликатов

Figure. Augmentation method flowchart.

A_0 — person's first replica; S_i — syntax tree; A_i — one of the person's replicas; T_j — syntactic template for the replica A_i, where $j = i$; f — frequency of using a syntactic template; r — configurable parameter for the number of augmentation options for one replica; T_i — one of the syntax patterns after clearing duplicates

Вместе с исходной репликой человека теги отправляются в SCPN — модель кодера-декодера для синтаксически контролируемой генерации парафраз из фреймворка OpenAttaker для генерации аугментированных данных. Таким образом, каждая реплика персоны может быть преобразована r различными способами. В настоящей работе выбрано $r = 5$ и проведены эксперименты с учетом данного значения.

Поскольку для аугментации данных использованы только характерные для человека синтаксические конструкции, то сохранены синтаксические особенности речи человека. Лексика и стиль частично сохранены, в связи с тем, что аугментация основана на преобразовании синтаксического дерева, а при данном преобразовании в первую очередь используется лексика исходного предложения в случае, если какая-либо из частей речи отсутствует в исходной реплике, SCPN добавляет необходимые части речи (союзы, предлоги, частицы) для сохранения синтаксической согласованности дополненных реплик. Если остальные части речи отсутствуют, они добавляются с помощью Long short-term memory, генерирующей слова. Например, фраза «Oh, unbelievable, we had the best time» с синтаксическим шаблоном

(ROOT(S(INTJ(UH))(NP(PRP))VP(VBD)×
×(NP(NP(DT)(JJS)(NN)(SBAR(IN)(PRP)×
(VP(VBD)))))(ADJP(JJ)))))

преобразуется во фразу «Oh, it was the best time that we had, unbelievable». Для работы с русскоязычными данными выполнена дополнительная подготовка. Для SCPN она заключалась в дообучении SCPN-модели на русскоязычных данных. Для этого был сформирован h5 файл, в котором содержались примеры синтаксических деревьев, для каждой из синтаксических структур были сформированы входные данные из одной фразы и выходные данные из похожей фразы с ее синтаксической структурой. Данные использовались из датасета Тайга корпуса Наташа. Помимо этого, для датасета Russian Tweet Corpus (RuTweetCorp) проведена дополнительная очистка данных — из сообщений удалены обращения и хэштеги.

Эксперименты и результаты

Эксперименты проведены на следующих наборах данных:

— Multimodal EmotionLines Dataset (MELD) — набор данных, содержащий более 1400 диалогов и 13 000 высказываний из сериала «Друзья». В диалогах участвовали несколько дикторов. Каждое высказывание в диалоге было отмечено любой из семи эмоций: Гнев, Отвращение, Печаль, Радость, Нейтральность, Удивление и Страх;

— MELDRU — набор данных MELD, переведенный на русский язык с помощью модели машинного перевода MarianMT;

— RuTweetCorp — русскоязычный корпус коротких текстов на основе русскоязычных постов из микроблога Twitter, состоящий из 17 639 674 записей. Корпус содержит более 139 000 различных дикторов. Данные размечены на два класса по эмотивной валентности: позитивные и негативные.

Пример аугментированных данных с помощью синтаксического перефразирования представлен в табл. 1.

В связи с тем, что в рамках работы рассмотрены датасеты, содержащие неформальную разговорную речь (реплики диалогов и твиты), в данных присутствовало небольшое количество высказываний, обогащенных сложными синтаксическими конструкциями.

Для аугментированных данных оценивалась сохранность языковых характеристик человека с помощью Mean Opinion Score (MOS) оценки. Данная оценка часто используется для предоставления численного значения о качестве аугментированной аудио информации. Для оценки было опрошено 10 носителей рус-

Таблица 1. Пример аугментированных данных с помощью синтаксического перефразирования

Table 1. Examples of data augmented with syntactic paraphrasing

Исходный текст	Датасет	Результат
Oh hey, don't thank me, thank yourself. You're the one who faced her fears and ultimately overcame them.	MELD	['You're the one who faced her fears and eventually stopped them. Thank yourself.', 'Don't thank me, the one who faced her fears and surpassed them is you.', 'Oh, thank yourself, you 're the one who faced her fears , and you missed them.', 'Of course, you 're the one who faced her fears and eventually stopped them', «Oh, you don't thank me , you 're the one who faced her fears and eventually stopped them .»]
Так ты говоришь мне, что между тобой и Чендлером ничего нет.	MELDRU	['Ты говоришь мне, что ничего не происходит между тобой и Чендлером.', 'Так что скажи мне, что это не происходит между тобой и Чендлером.', ''Так ты говоришь мне, что это не происходит между тобой и Чендлером.', 'Так ты говоришь мне, что ничего не происходит между тобой и Чендлером.', 'Так что скажи мне, что с тобой и Чендлером ничего нет.']
Самое тяжкое, что долго не живут животные. Привыкаешь.	RuTweetCorp	['Самое трудное — животные живут недолго, а ты привыкаешь', 'Самое трудное в том, что животные не живут долго. Привязываешься', 'Трудная вещь, когда долго не живут питомцы. Привязываешься.', 'Тяжкое, животные не живут долго']

ского языка по трем анкетам на основе наборов данных MELD, MELDRU и RuTweetCorp. В связи с тем, что все персоны, принявшие участие в опросе, являются носителями русского языка, выполнена оценка только русскоязычных наборов данных, так как результаты оценки англоязычного набора данных не были бы объективными из-за отсутствия соответствующего культурного отпечатка и менталитета. В анкетах опросов содержались примеры фраз, характерные для персон. Данные примеры представляли из себя список из пяти фраз, соотнесенных с их персоной. Для примеров из используемых датасетов выбирались наиболее отражающие стиль речи персоны фразы (фразы, в которых присутствовала наиболее часто употребляемая персоной лексика или синтаксическая структура предложения). Далее респонденты выбирали из реплик, аугментированных с помощью различных методик (eda, mt5, предложенный метод), наиболее подходящую реплику для указанной персоны. Для обеспечения чистоты эксперимента каждым методом аугментации было сгенерировано равное количество аугментированных предложений. Выбранная реплика оценивалась значениями от 1 до 5 по шкале осмысленности и логичности, где 5 — «фраза полностью логична», «фраза полностью осмыслена». Оценки представлены в табл. 2.

В результате отметим, что респонденты чаще выбирали реплики, сгенерированные с помощью разработанного авторами данной работы метода, как более подходящие персоне. Заметим, что предложенный метод генерирует реплики с достаточно высоким уровнем логичности и осмысленности. Так как предлагаемые респондентам реплики в анкетах никак не были помечены алгоритмами их сгенерировавшими, то они выбирали подходящие персоне реплики без знания об алгоритмах. Иначе говоря, метод сохраняет отличительные характеристики речи лучше, чем аналоги, с которыми проведено сравнение.

Для оценки влияния аугментированных данных на точность моделей, данные были использованы для распознавания эмоций и анализа эмотивной валентности текста. В связи с тем, что анализ эмотивной валентности в данной работе используется лишь для оценки влияния аугментированных данных на метрики моделей, была рассмотрена эмотивная валентность самого высказывания в диалоговом контексте, при этом особенности характера персоны не учитывались.

Распознавание эмоций в полилогах (Emotion recognition in conversation, ERC) с использованием современных архитектур нейронных сетей достигло заметных успехов. Распознавание эмоций может осуществляться для разных модальностей, текстовая модальность при одномодальных исследованиях показывает лучшие результаты [16, 17]. В последних работах в этой области использованы различные виды трансформеров для кодирования текстовых контекстов [18–20] или, например, графовых сетей [21]. Отсутствие доступных и подходящих наборов данных является основным препятствием для решения этих вопросов для русского языка. Однако существует множество таких наборов данных для английского языка. Исходя из этого, в качестве одного из подходов к решению этой проблемы часто предлагается использовать автоматический перевод англоязычных наборов данных на русский язык. В данном случае использование такого подхода дает результаты достаточного качества. Для исследования был выбран EmoBERTa для решения проблемы ERC. Наряду с хорошей точностью (топ-2 для набора данных MELD), он достаточно прост и удобен в использовании:

— модель RoBERTa предварительно обучается как часть задачи MLM (маскированного языкового моделирования) на целевом языке;
— модель дообучается для задачи классификации текста.

Данные для обучения классификации представляют собой конкатенированные диалоговые контексты. Каждый диалоговый контекст содержит целевую реплику и четыре другие реплики, задействованные в том же диалоге, что и целевая. Включение дополнительных реплик позволяет при оценке эмотивной валентности учитывать контекст целевой реплики за счет предыдущих фраз дикторов и реплик собеседников. Средняя длина одной реплики составляет 8 слов (токенов) для датасета MELD, 6 — для MELDRU и 10 — для RuTweetCorp. Для аугментированных данных диалоговый контекст собран следующим образом: для каждой реплики из контекста, в том числе и для целевой, случайным образом с заданной вероятностью берется один из пяти аугментированных вариантов и присоединяется к общему контексту.

Для анализа эмотивной валентности текста сообщений на русском языке была выбрана предобученная Multilingual BERT модель. Сравнение влияния ауг-

Таблица 2. MOS оценка на данных датасетов MELDRU и RuTweetCorp
Table 2. MOS assessment based on MELDRU and RuTweetCorp datasets

Параметры	Датасет					
	MELDRU			RuTweetCorp		
	EDA (Easy Data Augmentation)	MT5	Предложенный метод	EDA	MT5	Предложенный метод
Средняя оценка осмысленности	4,2	4,4	4,4	4,3	3,8	4,2
Средняя оценка логичности	4,3	4,4	**4,5**	4,6	3,2	3,8
Процент анкет с преобладанием данного метода	—	12	**88**	—	30	**70**
Процент доли выбора метода для всех вопросов	17	18	**65**	18	32	**50**

Таблица 3. Результаты точности моделей с аугментированными данными
Table 3. Performance of the tested augmentation approaches

Модель	Датасет	Доля аугментированных данных	Предложенный метод		EDA	
			Accuracy	Weighted F1	Accuracy	Weighted F1
RoBERTa-base	MELDRU	0,0	56,02	50,01	56,02	50,01
RoBERTa-base	MELDRU	0,5	55,86	**53,17**	53,25	51,14
RoBERTa-base	MELDRU	1,0	52,43	50,51	51,76	48,34
Distill RoBERTa-base	MELD	0,0	60,30	58,19	60,30	58,19
Distill RoBERTa-base	MELD	0,5	**61,87**	**60,40**	53,39	48,01
Distill RoBERTa-base	MELD	1,0	53,26	48,77	51,92	47,20
MBERT-base	RuTweetCorp	0,0	68,30	67,27	68,30	67,27
MBERT-base	RuTweetCorp	0,5	**70,60**	**69,40**	62,76	57,74
MBERT-base	RuTweetCorp	1,0	67,45	66,29	64,24	61,36

ментированных вариантов предложенным методом проведено с классическим методом аугментации EDA, в котором аугментация текста производится за счет замены синонимами, с помощью случайных вставки, перестановки и удаления.

Эксперимент по распознаванию эмоций и анализу эмотивной валентности выполнен следующим образом. Было опробовано три варианта использования аугментированных данных. Вариант aug-prob=0,0 соответствует обучению без аугментаций. В варианте aug-prob=1,0 каждое высказывание заменяется аугментацией. Вариант augprob=0,5 — точка баланса, при котором половина высказываний заменяется аугментацией, другая половина остается неизменной. Результаты экспериментов приведены в табл. 3.

В результате оценки отметим, что предложенный метод не только позволил достичь повышения точности для разных моделей до 2 %, но и показал результат по влиянию на точность лучше, чем метод аугментации EDA.

Заключение

Предложен метод аугментации текстовых данных, который сохраняет стиль речи и словарный запас. В работе использованы наборы данных MELD, MELDRU и RuTweetCorp для экспериментов по оценке эмоционального состояния пользователя. Замечено, что аугментация данных с помощью представленного метода приводит к повышению точности при доле аугментированных данных к исходным менее 1,0, потому что в этом случае есть шанс, что исходное сообщение останется неизменным. Исходное сообщение, каким бы хорошим ни было дополнение, содержит данные, которые максимально подходят к домену, из которого оно было взято. С другой стороны, полезно увеличить разнообразие данных с помощью аугментации, повышающей приобретенную способность сохранять стиль речи за счет лучшего обобщения. Наблюдалось увеличение точности для разных моделей до 2 %.

Литература

1. Giridhara P.K., Mishra C., Venkataramana R.K., Bukhari S.S., Dengel A.R. A study of various text augmentation techniques for relation classification in free text // Proc. of the 8th International Conference on Pattern Recognition Applications and Methods. 2019. P. 360–367 https://doi.org/10.5220/0007311003600367
2. Papadaki M. Data Augmentation Techniques for Legal Text Analytics: A thesis submitted to Athens University of Economics and Business in fulfillment of the requirements for the degree of Master in Data Science. 2017. 33 p.
3. Zhang Z., Zweigenbaum P. GNEG: Graph-based negative sampling for word2vec // Proc. of the 56th Annual Meeting of the Association for Computational Linguistics. V. 2. 2018. P. 566–571. https://doi.org/10.18653/v1/P18-2090
4. Wei J., Zou K. EDA: Easy data augmentation techniques for boosting performance on text classification tasks // Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2018. P. 6382–6388. https://doi.org/10.18653/v1/D19-1670
5. Wu X., Xia Y., Zhu J., Wu L., Xie S., Fan Y., Qin T. mixSeq: A simple data augmentation method for neural machine translation // Proc. of the 18th International Conference on Spoken Language Translation (IWSLT 2021). 2021. P. 192–197. https://doi.org/10.18653/v1/2021.iwslt-1.23

References

1. Giridhara P.K., Mishra C., Venkataramana R.K., Bukhari S.S., Dengel A.R. A study of various text augmentation techniques for relation classification in free text. *Proc. of the 8th International Conference on Pattern Recognition Applications and Methods*, 2019, pp. 360–367 https://doi.org/10.5220/0007311003600367
2. Papadaki M. *Data Augmentation Techniques for Legal Text Analytics*. A thesis submitted to Athens University of Economics and Business in fulfillment of the requirements for the degree of Master in Data Science, 2017, 33 p.
3. Zhang Z., Zweigenbaum P. GNEG: Graph-based negative sampling for word2vec. *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics. V. 2*, 2018, pp. 566–571. https://doi.org/10.18653/v1/P18-2090
4. Wei J., Zou K. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. *Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2018, pp. 6382–6388. https://doi.org/10.18653/v1/D19-1670
5. Wu X., Xia Y., Zhu J., Wu L., Xie S., Fan Y., Qin T. mixSeq: A simple data augmentation method for neural machine translation. *Proc. of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, 2021, pp. 192–197. https://doi.org/10.18653/v1/2021.iwslt-1.23

6. Kumar V., Choudhary A., Cho E. Data augmentation using pre-trained transformer models // Proc. of the 2nd Workshop on Life-long Learning for Spoken Language Systems. 2020. P. 18–26.
7. Kobayashi S. Contextual augmentation: Data augmentation by words with paradigmatic relations // Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). 2018. P. 452–457. https://doi.org/10.18653/v1/N18-2072
8. Yu A., Dohan D., Luong M., Zhao R., Chen K., Norouzi M., Le Q. QANet: Combining local convolution with global self-attention for reading comprehension // Proc. of the ICLR Conference. 2018.
9. Mehdi R., Meyer M., Goutal S. Text Data Augmentation: Towards better detection of spear-phishing emails // arXiv. 2020. arXiv:2007.02033. https://doi.org/10.48550/arXiv.2007.02033
10. Edunov S., Ott M., Auli M., Grangier D. Understanding back-translation at scale // Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018. P. 489–500. https://doi.org/10.18653/v1/D18-1045
11. Guo H., Mao Y., Zhang R. Augmenting data with mixup for sentence classification: An empirical study // arXiv. 2019. arXiv:1905.08941. https://doi.org/10.48550/arXiv.1905.08941
12. Coulombe C. Text data augmentation made simple by leveraging NLP cloud APIs // arXiv. 2018. arXiv:1812.04718. https://doi.org/10.48550/arXiv.1812.04718
13. Shen T., Lei T., Barzilay R., Jaakkola T. Style transfer from non-parallel text by cross-alignment // Advances in Neural Information Processing Systems. 2017. V. 30.
14. Yang S., Huang X., Lau J.H., Erfani S. Robust task-oriented dialogue generation with contrastive pre-training and adversarial filtering // Findings of the Association for Computational Linguistics (EMNLP 2022). 2022. P. 1220–1234.
15. Kovriguina L., Shilin I., Shipilo A., Putintseva A. Russian tagging and dependency parsing models for stanford CoreNLP natural language toolkit // Communications in Computer and Information Science. 2017. V. 786. P. 101–111. https://doi.org/10.1007/978-3-319-69548-8_8
16. Matveev Y., Matveev A., Frolova O., Lyakso E., Ruban N. Automatic speech emotion recognition of younger school age children // Mathematics. 2022. V. 10. N 14. P. 2373. https://doi.org/10.3390/math10142373
17. Lyakso E., Frolova O., Matveev A., Matveev Y., Grigorev A., Makhnytkina O., Ruban N. Recognition of the emotional state of children with down syndrome by video, audio and text modalities: human and automatic // Lecture Notes in Computer Science. 2022. V. 13721. P. 438–450. https://doi.org/10.1007/978-3-031-20980-2_38
18. Kim T., Vossen P. EmoBERTa: Speaker-Aware Emotion Recognition in Conversation with RoBERTa // arXiv. 2021. arXiv:2108.12009. https://doi.org/10.48550/arXiv.2108.12009
19. Song X., Zang L., Zhang R., Hu S., Huang L. Emotionflow: Capture the dialogue level emotion transitions // Proc. of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2022. P. 8542–8546. https://doi.org/10.1109/ICASSP43922.2022.9746464
20. Shen W., Chen J., Quan X., Xie Z. DialogXL: All-in-One XLNet for multi-party conversation emotion recognition // Proceedings of the AAAI Conference on Artificial Intelligence. 2021. V. 35. N 15. P. 13789–13797 https://doi.org/10.1609/aaai.v35i15.17625
21. Shen W., Wu S., Yang Y., Quan X. Directed acyclic graph network for conversational emotion recognition // Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021. P. 1551–1560.

6. Kumar V., Choudhary A., Cho E. Data augmentation using pre-trained transformer models. *Proc. of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, 2020, pp. 18–26.
7. Kobayashi S. Contextual augmentation: Data augmentation by words with paradigmatic relations. *Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2018, pp. 452–457. https://doi.org/10.18653/v1/N18-2072
8. Yu A., Dohan D., Luong M., Zhao R., Chen K., Norouzi M., Le Q. QANet: Combining local convolution with global self-attention for reading comprehension. *Proc. of the ICLR Conference*, 2018.
9. Mehdi R., Meyer M., Goutal S. Text Data Augmentation: Towards better detection of spear-phishing emails. *arXiv*, 2020, arXiv:2007.02033. https://doi.org/10.48550/arXiv.2007.02033
10. Edunov S., Ott M., Auli M., Grangier D. Understanding back-translation at scale. *Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 489–500. https://doi.org/10.18653/v1/D18-1045
11. Guo H., Mao Y., Zhang R. Augmenting data with mixup for sentence classification: An empirical study. *arXiv*, 2019, arXiv:1905.08941. https://doi.org/10.48550/arXiv.1905.08941
12. Coulombe C. Text data augmentation made simple by leveraging NLP cloud APIs. *arXiv*, 2018, arXiv:1812.04718. https://doi.org/10.48550/arXiv.1812.04718
13. Shen T., Lei T., Barzilay R., Jaakkola T. Style transfer from non-parallel text by cross-alignment. *Advances in Neural Information Processing Systems*, 2017, vol. 30.
14. Yang S., Huang X., Lau J.H., Erfani S. Robust task-oriented dialogue generation with contrastive pre-training and adversarial filtering. *Findings of the Association for Computational Linguistics (EMNLP 2022)*, 2022, pp. 1220–1234.
15. Kovriguina L., Shilin I., Shipilo A., Putintseva A. Russian tagging and dependency parsing models for stanford CoreNLP natural language toolkit. *Communications in Computer and Information Science*, 2017, vol. 786, pp. 101–111. https://doi.org/10.1007/978-3-319-69548-8_8
16. Matveev Y., Matveev A., Frolova O., Lyakso E., Ruban N. Automatic speech emotion recognition of younger school age children. *Mathematics*, 2022, vol. 10, no. 14, pp. 2373. https://doi.org/10.3390/math10142373
17. Lyakso E., Frolova O., Matveev A., Matveev Y., Grigorev A., Makhnytkina O., Ruban N. Recognition of the emotional state of children with down syndrome by video, audio and text modalities: human and automatic. *Lecture Notes in Computer Science*, 2022, vol. 13721, pp. 438–450. https://doi.org/10.1007/978-3-031-20980-2_38
18. Kim T., Vossen P. EmoBERTa: Speaker-Aware Emotion Recognition in Conversation with RoBERTa. *arXiv*, 2021, arXiv:2108.12009. https://doi.org/10.48550/arXiv.2108.12009
19. Song X., Zang L., Zhang R., Hu S., Huang L. Emotionflow: Capture the dialogue level emotion transitions. *Proc. of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8542–8546. https://doi.org/10.1109/ICASSP43922.2022.9746464
20. Shen W., Chen J., Quan X., Xie Z. DialogXL: All-in-One XLNet for multi-party conversation emotion recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, no. 15, pp. 13789–13797 https://doi.org/10.1609/aaai.v35i15.17625
21. Shen W., Wu S., Yang Y., Quan X. Directed acyclic graph network for conversational emotion recognition. *Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 1551–1560.

Авторы

Матвеева Анастасия Андреевна — инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 57204215042, https://orcid.org/0000-0002-2935-991X, anastasiamatveevaitmo@gmail.com

Махныткина Олеся Владимировна — кандидат технических наук, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 57208002090, https://orcid.org/0000-0002-8992-9654, makhnytkina@itmo.ru

Authors

Anastasia A. Matveeva — Engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 57204215042, https://orcid.org/0000-0002-2935-991X, anastasiamatveevaitmo@gmail.com

Olesia V. Makhnytkina — PhD, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 57208002090, https://orcid.org/0000-0002-8992-9654, makhnytkina@itmo.ru

Статья поступила в редакцию 21.05.2023
Одобрена после рецензирования 20.06.2023
Принята к печати 28.07.2023

Received 21.05.2023
Approved after reviewing 20.06.2023
Accepted 28.07.2023