Нормализация русскоязычных текстов на примере корпуса социальных сетей

Г. О. Феоктистов, Д. А. Морозов

***Аннотация*—Разметка текста — это процесс обогащения текста металингвистической информацией, например, леммами, морфологическими пометами, синтаксическими связями. Разметка текста является фундаментальной задачей компьютерной лингвистики и играет ключевую роль как в фундаментальных, так и в прикладных задачах обработки естественного языка. Масштабы современных текстовых корпусов с объёмом больше миллиарда словоупотреблений привели к необходимости использования автоматизированных инструментов разметки. Большинство подобных решений разработаны и протестированы на текстах со стандартной орфографией. Как следствие, качество их работы может значительно отличаться при применении к текстам из социальных сетей, содержащим нестандартные употребления. Одним из способов преодоления этой проблемы является предварительная нормализация текста. Эта задача схожа с автоматическим исправлением орфографии, но не эквивалентна ей, и остается значительно менее изученной. Нормализация текста требует не только исправления опечаток и орфографических ошибок, но и восстановления сокращений, особенно распространённых в онлайн-общении. В этой статье мы представляем корпус русскоязычных предложений из социальных сетей и их нормализованных вариантов, доступный по адресу https://huggingface.co/datasets/ruscorpora/normalization. На основе этого корпуса мы составили список типичных искажений речи и сравнили эффективность различных методов нормализации текста.**

***Ключевые слова*—корпусная лингвистика, нормализация текста, русский язык, социальные сети.**

I. Введение

Свобода от нормативных ограничений в языке социальных сетей представляет значительный интерес для исследования новых языковых явлений. В отличие от письменной речи в новостях, художественных и научно-популярных текстах, дискурс социальных сетей постоянно меняется из-за культурного обмена и неконтролируемого влияния иностранных языков. Языковые явления, такие как сленг, жаргон, нестандартные сокращения и грамматические конструкции, распространённые опечатки, делают социальные сети бесценным источником материала для изучения языковых изменений в реальном времени.

Для систематического анализа этих явлений исследователи должны полагаться на аннотированные корпусы — структурированные коллекции текстов, обогащенные лингвистической разметкой. Наиболее распространенным типом разметки является морфологическая, включающая часть речи и грамматические категории. Также востребованы и прочие виды разметки, например, разметка лемм, синтаксических связей, семантических категорий.

Для аннотирования больших по объёму корпусов единственным возможным вариантом является применение автоматических инструментов разметки. Однако такие модели обычно обучают на данных, содержащих крайне мало примеров нестандартных употреблений. Это приводит к резкому снижению качества разметки при аннотировании корпусов с нестандартной речью, например, корпусов социальных сетей.

Эффективность разметки в такой ситуации можно повысить двумя способами: адаптировать обучающие данные и, как следствие, модели разметки или нормализовать, то есть привести к стандартной письменной речи, тексты из социальных сетей. В первом случае, чтобы дообучить модель разметки, требуется ручное аннотирование большого количества текстов из социальных сетей, к которому необходимо привлечь профессиональных лингвистов для установления корректных частей речи, грамматических категорий, лемм и так далее. Во втором случае, напротив, будет достаточно грамотных носителей языка, чтобы исправить ошибки, опечатки и другие искажения письменной речи и, тем самым, подготовить данные для обучения модели нормализации. Обучив подобную модель и нормализовав целевой текст, можно получить высококачественную разметку при помощи стандартной модели разметки и перенести её на исходный текст.

В то же время область автоматической нормализации текста исследована достаточно слабо. Не в последнюю очередь это связано с нехваткой подготовленных наборов данных. Важно отметить, что хотя эта задача на первый взгляд очень похожа на задачу исправления орфографии, для которой существуют подготовленные наборы данных, мы считаем, что разница между этими задачами фундаментальна: задача нормализации текста намного шире и включает не только исправление орфографических ошибок, но и, например, восстановление исходных словоформ из сленговых вариантов (например, *разраб → разработчик, заблочили → заблокировали*) и расшифровку сокращений (например, *тп → тому подобное, шт → штук*).

В целях устранения этого пробела мы подготовили

Статья получена 10 ноября 2025.

Феоктистов Григорий Олегович, Новосибирский государственный университет (e-mail: g.feoktistov@g.nsu.ru).

Дмитрий Алексеевич Морозов, Новосибирский государственный университет (e-mail: morozowdm@gmail.com).

набор данных, состоящий из 2000 русскоязычных предложений с приписанными им нормализованными вариантами. Для аннотирования данных мы разработали протокол, формализующий концепцию «нормализации текста», и вручную подготовили нормализованные варианты для предложений, отобранных из корпуса социальных сетей Национального корпуса русского языка [1]. Предложения для набора отбирались при помощи набора эвристических правил таким образом, чтобы достичь, как ожидается, повышенной доли искаженных написаний. На материале собранного набора данных мы также протестировали несколько базовых подходов к нормализации русскоязычных текстов.

II. Обзор предметной области

Задаче нормализации текста уделяется значительно меньше внимания, чем задаче исправления орфографии. Большинство актуальных на сегодняшний день подходов основаны на многоэтапных конвейерах обработки, которые сочетают эвристические правила с техниками машинного обучения [2,3,4]. Например, алгоритм MoNoise [2] генерирует кандидатов на замену для каждого слова, используя множество источников: словарь Aspell, векторную word2vec-модель, эвристики для восстановления пропущенных пробелов и т.д. Затем каждый кандидат оценивается с использованием набора признаков (например, измеряется расстояние Левенштейна от исходного слова и косинусное сходство между соответствующими каждому из слов word2vec-эмбеддингами), а окончательное решение о замене принимается обученной моделью случайного леса. Авторы [3] исследовали нормализацию с помощью предсказания замаскированных токенов, схожего с протоколом обучения моделей типа BERT. Представленный подход представляет собой гибридную пошаговую систему: выявление ненормализованных токенов, расшифровка сокращений и аббревиатур на базе словаря, распознавание именованных сущностей и маскирование оставшихся ненормализованных токенов, далее восстанавливаемых при помощи BERT-подобной модели. Также исследовалась нормализация на основе seq2seq-генерации с использованием предобученной модели BART [4]. При этом провести репрезентативное сравнение перечисленных подходов на материале русского языка невозможно из-за отсутствия подходящего набора данных.

Применительно к русскому языку следует упомянуть соревнование SpellRuEval, посвящённое разработке алгоритмов исправления орфографии [5]. Анализ наиболее эффективных решений, представленных в рамках соревнования, позволяет выявить несколько важных особенностей задачи:

1) Выбор среди кандидатов оказывается более сложной задачей, чем их генерация. Основная трудность заключается не в генерации потенциальных исправлений, а в точном выборе правильного кандидата.

2) Наибольшую сложность представляют ошибки, совпадающие с реальными словами. Опечатки и ошибки, которые приводят к существующим словарным словам (вместе с неправильно расставленными дефисами и пробелами), хуже всего обрабатываются алгоритмами.

3) Единообразная методология лучших подходов. Три ведущие команды использовали схожие подходы, сочетая учёт расстояния Левенштейна с триграммными языковыми моделями.

Усовершенствованный метод, который превзошел лучшие результаты SpellRuEval, был представлен в работе [6]. В особенности прогресса удалось достичь в обработке ошибок с дефисами и пробелами. Это исследование ещё раз подтвердило, что ключевой проблемой являются ошибочные написания слов, совпадающие с другими существующими словами. Около 30% ошибок модели связаны с обработкой подобных случаев. На основании этого авторы приходят к выводу о необходимости применения предобученных языковых моделей, способных учитывать семантику контекста словоупотребления.

Эффективность применения больших языковых моделей (LLM) в задаче коррекции орфографии была исследована в недавнем комплексном исследовании [7]. В рамках этой работы авторы сравнили несколько LLM, включая M2M100, FredT5, модели OpenAI и решения с открытым исходным кодом, на разнообразных тестовых наборах (RUSpellRU, MultidomainGold, MedSpellChecker и GitHubTypeCorpusRu) с использованием zero-shot подхода и дообучения на целевом домене. Наибольшую эффективность продемонстрировали дообученные модели, причём достигнутый уровень позволяет предположить, что в будущем лучшие результаты и в нормализации текста могут быть достигнуты с использованием LLM.

III. Подготовка данных

Наш набор данных был сформирован на основе предложений из корпуса социальных сетей Национального корпуса русского языка. Сначала мы собрали десять тысяч случайно выбранных предложений из публичных каналов и чатов Telegram и обсуждений из подкорпуса социальной сети VK (vk.com). Далее мы отфильтровали эти предложения, чтобы выбрать те, которые с большей вероятностью содержат примеры ненормализованного написания. Для этой фильтрации мы использовали словарь Aspell для русского языка, исключая из рассмотрения слова, содержащие некириллические символы (цифры, латиницу и т.д.). В итоговый набор данных вошло две тысячи ненормализованных предложений.

Перед нормализацией все предложения были токенизированы: вся пунктуация была удалена, а токены разделены пробелами. Нормализация набора данных была разделена на 3 этапа. Сначала мы описали черновую версию протокола для первой нормализации. Затем, используя этот протокол, мы нормализовали весь набор данных, обсуждая неоднозначные случаи и корректируя протокол в соответствии с примерами из набора данных. Наконец, мы сформулировали финальную версию протокола и в соответствии с ней обновили аннотацию данных. В процессе нормализации

мы стремились сохранить исходное написание в случае неоднозначности и отсутствия контекста. Итоговый протокол содержит следующие положения:

1. **Исправление**. В эту категорию мы включили случаи, в рамках которых искажение исправлялось. В рамках этой категории искажения имеют единственный вариант исправления:

a) Сокращения и модификации. В эту группу включены искажённые варианты слов, которые образованы удалением части слова (сокращения) и модификации: *уши → наушники*, *прошка → про-версия*, *50к → 50 тысяч*.

b) Разговорное написание. Например, *чё → что*, *гамно → говно*, *тыща → тысяча*, *бачи → баксы*. Стоит отметить, что в эту группу не входят разговорные варианты, образованные по правилам русского языка, например, диминутивы: *ясненько* (*ясно*), *скучновато* (*скучно*).

c) Раздельное, слитное и дефисное написание. В рамках нормализации ошибок в правописании слитного и дефисного написания используется знак "|" для сохранения соответствий слов между исходным и нормализованным предложением. Примеры: *коллцентры → кол-центры*, *по больше → побольше*, *шьюбки → шью|бки*.

d) Согласование и предлоги. Ошибки в окончаниях и использование неправильных предлогов: *(спасибо чудесной) моделе → (спасибо чудесной) модели, на голову приходит → в голову приходит, с старой → со старой*.

e) Орфографические ошибки и опечатки. Примеры: *типо → типа, пачему → почему, учпеть → успеть*.

f) Повторяющиеся слова и слоги. Например, *ноно там ещё → но там ещё*.

g) Отсутствие буквы «ё». Например, *причем → причём*.

i) Буквенные наращения. Например, *16тый → 16-й*.

j) Комбинация предыдущих случаев. Например, *обращаи тесь → обращайтесь*, где встречается как ошибочное раздельное написание, так и опечатка.

2. Сохранение. В эту категорию включены случаи, которые не требуют нормализации.

a) **Исходное написание** абсценной лексики ввиду отсутствия регуляций в русском языке. Тем не менее, опечатки в данных словах и их сокращения исправляются.

b) **Игра слов**. Например, вымираты (вымирать + эмираты).

c) **Окказионализмы**. Например, вкдилдодрон.

d) **Неологизмы**. Например, дистрибутив, фейк.

e)**Иноязычные слова и варваризмы**. Например, диджтл (англ. digital), комьюнити (англ. community).

f) **Жаргон**. Например, бабки (деньги), жопожник (скряга).

g) **Синлексемы**. Например, голосовое (голосовое сообщение).

3. Замена слов на теги. В некоторых случаях мы заменяли слова на теги.

a) **Слова, написанные на других языках**, были заменены на тег <foreign>. Если предложение написано целиком на иностранном языке, то оно полностью заменяется на набор тегов <foreign>.

b) **В случае исправлении написания на слитное**, например, обращаи тесь, то первое слово заменяется на корректное, в данном случае на обращайтесь, а второе слово на тег <conjoint>, чтобы сохранить соответствие слов в предложениях. Таким образом, обращаи тесь → обращайтесь <conjoint>.

c) **В случае исправления написания на дефисное**, например, что то, то первое слово заменяется на корректное, в данном случае на что-то, а второе слово на тег <hyphen>, чтобы сохранить соответствие слов в предложениях. Таким образом, что то → что-то <hyphen>.

d) **Междометия** заменяются на тег <interjection>. Например, Вауууу → <interjection>.

4. Сложные случаи. В эту категорию попали те случаи, которые имеют разные варианты исправлений и требовали обсуждений во время нормализации.

Орфографические ошибки и опечатки в словах, не представленных в словарях. Мы стремились к исправлению на наиболее часто используемый вариант написания. Например, павербанк → повербанк, коммьюнити → комьюнити, скил → скилл.

a) **Ошибки капитализации**. Мы исправляли написание слов в начале предложений, имён собственных и слов, целиком написанных прописными буквами. Однако, из-за удаления пунктуационных знаков во время предобработки предложений исправление прописного и строчного написания остаётся непоследовательным.

b) **Аббревиатуры**. Мы сохраняли написание стран, организаций и так далее, например, ОАЭ (Объединённые Арабские Эмираты), ТТ (англ. TikTok).

c) Также мы сохраняли **сокращения в терминологии**, например, ОЗУ (оперативное запоминающее устройство), для аббревиатур иностранного происхождения, например, ВПН (англ. virtual private network, VPN) и для аббревиатур, которые обычно остаются зашифрованными, например, ФИО (фамилия, имя, отчество). Однако, мы расшифровывали типичные для интернет-дискурса аббревиатуры, например, мб → может быть, лс → личные сообщения, НГ → Новый год и те, которые были неочевидны, например, КЗ → Казахстан.

Несколько примеров из подготовленного набора данных приведены в Табл. 1. Итоговый набор данных доступен для скачивания по адресу: https://huggingface.co/datasets/ruscorpora/normalization.

Таблица 1. Примеры предложений из подготовленного набора данных

Исходное предложение	Нормализованное предложение
Еще есть твинк с 5к	Ещё есть твинк с 5|тысячами
чо там как приложение для apple tv не подвезли юзаем easy tv пока	Что там как приложение для <foreign> <foreign> не подвезли юзаем <foreign> <foreign> пока
авчом прикол то	А|в|чём прикол-то <hyphen>
На голову приходит только	В голову приходит только

Samsung galaxy S9 или iPhone 8	<foreign> <foreign> <foreign> или <foreign> 8
ну хз мне норм	Ну хрен|знает мне нормально
у кавото робит у кавото нет	У кого-то работает у кого-то нет
Вауууу 20 забагованных кусков говна с бесконечными данжонами крутаааа	<interjection> 20 забагованных кусков говна с бесконечными данжонами круто
вприципи класна но вообще та гавно	В|принципе классно но вообще-то <hyphen> говно

IV. Тестирование моделей нормализации

Для оценки сложности нормализации созданного набора данных мы провели эксперименты с несколькими простыми базовыми подходами.

A. *Алгоритмы нормализации*

Мы рассмотрели несколько подходов, адаптируя их для русского языка, где это было необходимо:

1. **MoNoise** [2]**.** Этот алгоритм состоит из двух этапов: генерации кандидатов и их ранжирования с последующим выбором наиболее вероятного. В рамках адаптации для русского языка мы использовали словарь Aspell для русского языка и две обученные нами векторные word2vec-модели: модель, обученную на ненормализованных данных из корпуса социальных сетей, и модель, обученную на материалах русскоязычного домена Википедии.

2. **Гибридная модель на основе BERT** [3]. Мы адаптировали подход, состоящий из следующих этапов: 1) первичная нормализация с помощью эвристик; 2) замена оставшихся искаженных слов специальным токеном [MASK]; 3) ранжирование предложенных моделью BERT (в нашем случае ruBert-base [8]) кандидатов на место токенов [MASK] на основе фонетической и графической схожести с исходным словом. Наши изменения по сравнению с оригинальным алгоритмом включали расширение списка эвристик для нормализации специфичных для русского языка искажений.

Стоит отметить, что мы рассмотрели 2 стратегии взаимодействия с BERT: 1) замена всех токенов, которые не были обработаны эвристиками, на токен [MASK] (модели с пометкой ALL); 2) замена только тех токенов, которые не были распознаны словарём (в нашем случае Aspell для русского языка) (модели с пометкой OOV). Кроме того, мы рассмотрели возможность дообучения используемой BERT-подобной модели. Дообучение проводилось в течение трёх эпох со стандартными параметрами.

3. **seq2seq-модели** [4]**.** Мы дообучили предварительно обученные генеративные seq2seq-модели на нашем наборе данных. Мы рассмотрели как генерацию всего предложения целиком, так последовательную генерацию отдельных токенов. Во втором случае входные данные содержали токен и контекст из двух слов с каждой стороны, отделённых при помощи специального токена _SEP_. В эксперименте с генерацией предложений мы использовали модели ruT5-base [8] и mbart-large-50-many-to-many-mmt [9] (далее MBart); в эксперименте с потокенной генерацией — только MBart, так как модель ruT5-base показала плохой результат в ходе предварительных экспериментов. Модели дообучались со стандартными параметрами в течение трех эпох.

B. *Метрики*

Качество генерации оценивалось с использованием следующих метрик:

1. ROUGE-1 для полных предложений.
2. F1-мера для токенов, требующих только исправления.
3. Среднее расстояние Левенштейна между целевыми и предсказанными предложениями (как в сыром виде, так и нормализованное по количеству слов).
4. Среднее расстояние Левенштейна между целевыми и предсказанными токенами (как в сыром виде, так и нормализованное по количеству символов).

C. *Результаты тестирования*

Перед перечислением полученных нами результатов крайне важно отметить, что они являются предварительными и не претендуют на всестороннее исследование подходов к нормализации текста. Нашей задачей было протестировать простые, интуитивно понятные решения на материале подготовленного набора данных и получить оценку базового уровня. По нашему мнению, модели, предварительно обученные на задаче исправления орфографии, или большие языковые модели (англ. Large Language Models, LLMs), смогут справиться с этой задачей гораздо лучше, и подобное исследование должно быть проведено в дальнейшем на материале созданного нами набора данных.

Результаты экспериментов представлены в Табл. 2-3. Табл. 2 содержит метрики, вычисленные для предложений целиком, Табл. 3 содержит метрики, нормированные на длину предложений. Постфиксы -t и -s в названиях моделей (Mbart-s) означают вариацию модели для генерации отдельных токенов и предложений, соответственно. FT означает то, что модель дообучалась на целевой задаче. Для подсчёта метрик мы сравнивали сгенерированные предложения с эталонной нормализацией. Строка «Предложения без норм.» соответствует сравнению исходных (ненормализованных) предложений с эталоном. Жирным в таблицах выделены лучшие достигнутые результаты.

Таблица 2. Результаты метрик по нормализации предложений

		ROUGE-1	Среднее расстояние Левенштейна	Норм. среднее расстояние Левенштейна
Предложения без норм.		0.11	4.3	0.10
MoNoise		0.19	3.0	0.09
RuBERT	OOV	0.15	5.0	0.15
	ALL	0.09	28.7	0.77
	FT OOV	0.15	5.1	0.14
	FT ALL	0.14	26.0	0.67

Mbart-t	**0.20**	**2.5**	**0.06**
Mbart-s	0.18	2.6	**0.06**
ruT5	0.11	10.9	0.21

Таблица 3. Результаты метрик по нормализации отдельных слов (токенов)

		F1	Среднее расстояние Левенштейна	Норм. среднее расстояние Левенштейна
Предложения без норм.		0.00	**0.28**	0.05
MoNoise		0.18	0.45	0.13
RuBERT	OOV	0.40	0.77	0.12
	ALL	0.17	4.55	0.84
	FT OOV	0.40	0.77	0.12
	FT ALL	0.18	4.09	0.76
Mbart-t		**0.71**	0.36	**0.04**
Mbart-s		0.66	0.32	**0.04**
ruT5		0.69	10.94	0.21

Из результатов очевидно, что рассмотренные базовые подходы справляются с задачей достаточно плохо. Для лучшего понимания особенностей работы моделей мы проанализировали вручную несколько показательных предложений из тестовой выборки, которые содержали следующие искажения письменной речи: аббревиатуры; разговорные и искажённые варианты; неверное раздельное, слитное и дефисное написание; орфографические ошибки; ошибки в предлогах и отсутствие буквы «ё». Мы выявили, что в общем случае модели имеют тенденцию сохранять исходное написание, что может быть следствием недостатка определённых примеров исправлений в обучающей выборке. Однако существуют и другие причины ошибок. Для иллюстрации мы приведём 4 предложения и варианты их исправления каждой моделью (Табл. 4), а также укажем на возможные причины ошибок.

Таблица 4. Исходное и целевое написание показательных предложений из тестовой выборки

Исходное написание	**Целевое написание**
На голову приходит только Samsung galaxy S9 или iPhone 8	В голову приходит только <foreign> <foreign> <foreign> или <foreign> 8
ну хз мне норм	Ну хрен|знает мне нормально
у кавото робит у кавото нет	У кого-то работает у кого-то нет
впринципи класна но вообще та гавно	В|принципе классно но вообще-то <hyphen> говно

1. **MoNoise.** Несмотря на то, что некоторые сокращения встречались достаточно часто, например *норм*, модель всё равно сохраняет исходное написание (Табл. 5). Мы связываем это с недостаточным объёмом данных для обучения модели случайного леса: используемый в оригинальной работе [2] набор данных значительно больше.

Таблица 5. Целевое и предсказанное адаптированной моделью MoNoise написание показательных предложений из тестовой выборки

Целевое написание	**MoNoise**
В голову приходит только <foreign> <foreign> <foreign> или <foreign> 8	на голову приходит только <foreign> <foreign> <foreign> Или <foreign> 8
Ну хрен|знает мне нормально	ну хз Мне норм
У кого-то работает у кого-то нет	у кавото работает у кавото Нет
В|принципе классно но вообще-то <hyphen> говно	впринципи класна но вообще та гавно

2**. Гибридная модель на основе BERT.** Большинство ошибок гибридной модели на основе BERT можно связать с недостатком контекста. Предложения в представленном наборе данных достаточно короткие, что приводит к тому, что после замены искажённых слов специальным токеном [MASK], оставшегося контекста не хватает для предсказания корректных вариантов исправления. В Табл. 6 можно найти этому подтверждение: модель склонна вставлять слова, которых не было в изначальном тексте.

Таблица 6. Целевое и предсказанное гибридной моделью на основе BERT написание показательных предложений из тестовой выборки

Целевое написание	**Гибридная модель на основе BERT**
В голову приходит только <foreign> <foreign> <foreign> или <foreign> 8	На голову приходит только <foreign> <foreign> <foreign> или <foreign> 8
Ну хрен|знает мне нормально	Ну хрен знает мне норм
У кого-то работает у кого-то нет	у тебя работает у гугла нет
В|принципе классно но вообще-то <hyphen> говно	Короче кака но вообще та кака

3. **seq2seq-модели.** Для анализа ошибок мы выбрали модель с лучшим результатом — Mbart-t. Из Табл. 7 видно, что модель справилась значительно лучше остальных подходов, в частности, она исправляет сокращения. Тем не менее, при значительной искажённости изначального текста, модель не способна к нормализации, как показано в последних двух предложениях. Мы считаем, что это вызвано сложностью данных примеров (в одном слове сразу несколько случаев нормализации) и недостатком обучающих данных, из-за чего даже довольно частые случаи (например, сокращение *норм*) модель предпочитает не исправлять.

Таблица 7. Целевое и предсказанное гибридной моделью на основе BERT написание показательных предложений из тестовой выборки

Целевое написание	**Гибридная модель на основе BERT**
В голову приходит только <foreign> <foreign> <foreign> или <foreign> 8	На голову приходит только <foreign> <foreign> или <foreign> 8
Ну хрен|знает мне нормально	Ну хрен|знает мне норм

У кого-то работает у кого-то нет	У кавото робит у кавото нет
В|принципе классно но вообще-то <hyphen> говно	Впринципи класна но вообще та говно

V. Заключение

Актуальность задачи нормализации текстов социальных сетей связана с ценностью этого языкового материала для лингвистов, исследующих развитие современного русского языка. В настоящей статье мы представляем открытый набор данных, состоящий из 2 тысяч пар русскоязычных предложений из социальных сетей: с оригинальным и нормализованным написанием. В ходе создания этого набора мы разработали протокол аннотирования, учитывающий разнообразные случаи искажений речи для русского языка. Публикация этого протокола позволит в дальнейшем расширить набор данных.

Подготовленный набор может быть использован для тестирования автоматизированных подходов к нормализации текста. В частности, мы проверили на его материале три наивных подхода к нормализации. Полученные нами результаты показывают, что задача нормализации текста остаётся нерешённой для русского языка и требует дальнейшего исследования.

Библиография

[1] Савчук С. О., Архангельский Т. А., Бонч-Осмоловская А. А., Донина О. В., Кузнецова Ю. Н., Ляшевская О. Н., Орехов Б. В., Подрядчикова М. В. Национальный корпус русского языка 2.0: новые возможности и перспективы развития // Вопросы языкознания. 2024. № 2. С. 7–34.

[2] van der Goot R., van Noord G. Monoise: Modeling noise using a modular normalization system. 2017. URL: https://arxiv.org/abs/1710.03476 (дата обращения: 10.11.2025).

[3] Doshi F., Gandhi J., Gosalia D., Bagul S. Normalizing text using language modelling based on phonetics and string similarity. 2020. URL: https://arxiv.org/abs/2006.14116 (дата обращения: 10.11.2025).

[4] Nguyen D.H., Nguyen A.T.H., Van Nguyen K. A weakly supervised data labeling framework for machine lexical normalization in vietnamese social media // Cognitive Computation. 2025. Vol. 17. Article 57. DOI: 10.1007/s12559-024-10356-3.

[5] Sorokin A., Shavrina T., Baytin A., Galinskaya I., Rykunova E. SpellRuEval: The first competition on automatic spelling correction for Russian // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (Москва, 1–4 июля 2016 г.). Вып. 15 (22). М.: Изд-во РГГУ, 2016. С. 660–673.

[6] Sorokin A. Spelling correction for morphologically rich language: a case study of Russian // Erjavec T., Piskorski J., Pivovarova L., Snajder J., Steinberger J., Yangarber R. (eds.) / Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing. Association for Computational Linguistics, Valencia, Spain (Apr 2017). 2017. P. 45–53. DOI: 10.18653/v1/W17-1408.

[7] Martynov N., Baushenko M., Kozlova A., Kolomeytseva K., Abramov A., Fenogenova A. A methodology for generative spelling correction via natural spelling errors emulation across multiple domains and languages // Graham Y., Purver M. (eds.) / Findings of the Association for Computational Linguistics: EACL 2024. Association for Computational Linguistics, St. Julian's, Malta (Mar 2024). 2024. P. 138–155. URL: https://aclanthology.org/2024.findings-eacl.10/ (дата обращения: 10.11.2025).

[8] Zmitrovich D., Abramov A., Kalmykov A., Kadulin V., Tikhonova M., Taktasheva E., Astafurov D., Baushenko M., Snegirev A., Shavrina T., Markov S.S., Mikhailov V., Fenogenova A. A family of pretrained transformer language models for Russian // Calzolari N., Kan M.Y., Hoste V., Lenci A., Sakti S., Xue N. (eds.) / Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). ELRA and ICCL, Torino, Italia (May 2024). 2024. P. 507–524. URL: https://aclanthology.org/2024.lrec-main.45/ (дата обращения: 10.11.2025).

[9] Tang Y., Tran C., Li X., Chen P.J., Goyal N., Chaudhary V., Gu J., Fan A. Multilin-gual translation with extensible multilingual pretraining and finetuning. 2020. URL: https://arxiv.org/abs/2008.00401 (дата обращения: 10.11.2025).

Феоктистов Григорий Олегович, Новосибирскийгосударственный университет (e-mail: g.feoktistov@g.nsu.ru).

Дмитрий Алексеевич Морозов, Новосибирский государственный университет (e-mail: morozowdm@gmail.com).

Text Normalization for Social Media Corpus

Grigory Feoktistov, Dmitry Morozov

***Abstract*—Text markup is the process of enriching text with metalinguistic information, such as lemmata, morphological tags, and syntactic relations. Text markup is a fundamental task in computational linguistics and plays a key role in both fundamental and applied natural language processing. The scale of modern text corpora, with over a billion word tokens, has necessitated the use of automated tagging tools. Most such solutions have been developed and tested on texts with standard orthography. Consequently, their performance can vary significantly when applied to texts from social media, which contain non-standard usages. One way to overcome this problem is through pre-normalization of the text. This task is similar to, but not equivalent to, automatic spell correction and remains significantly less studied. Text normalization requires not only the correction of typos and spelling errors but also the restoration of abbreviations, especially those common in online communication. In this article, we present a corpus of Russian-language sentences from social media and their normalized variants, available at https://huggingface.co/datasets/ruscorpora/normalization. Using this corpus, we compiled a list of typical speech distortions and compared the effectiveness of different text normalization methods.**

***Keywords*—corpus linguistics, text normalization, Russian language, social networks.**

REFERENCES

[1] S. O. Savchuk, T. Arkhangelskiy, A. A. Bonch-Osmolovskaya, O. V. Donina, Yu. N. Kuznetsova, O. N. Lyashevskaya, B. V. Orekhov and M. V. Podryadchikova, "Russian National Corpus 2.0: New opportunities and development prospect,". *Voprosy Jazykoznanija*, No. 2, pp. 7–34, 2024.

[2] R. van der Goot and G. van Noord, "Monoise: Modeling noise using a modular normalization system," 2017. Available: https://arxiv.org/abs/1710.03476.

[3] F. Doshi, J. Gandhi, D. Gosalia, S. Bagul, "Normalizing text using language modelling based on phonetics and string similarity," 2020. Available: https://arxiv.org/abs/2006.14116.

[4] D.H. Nguyen, A.T.H. Nguyen, K. Van Nguyen, "A weakly supervised data labeling framework for machine lexical normalization in vietnamese social media," *Cognitive Computation*, Vol. 17, article 57, 2025, doi: 10.1007/s12559-024-10356-3.

[5] A. Sorokin, T. Shavrina, A. Baytin,, I. Galinskaya and E. Rykunova, "SpellRuEval: The first competition on automatic spelling correction for Russian," In *Computational Linguistics and Intellectual Technologies Proceedings of the Annual International Conference "Dialogue" (2016)*, Issue 15, pp. 660–673, 2016.

[6] A. Sorokin, "Spelling correction for morphologically rich language: a case study of Russian," In T. Erjavec, J. Piskorski, L. Pivovarova, J. Snajder, J. Steinberger and R. Yangarber (eds.), *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pp. 45–53, Association for Computational Linguistics, Valencia, Spain (Apr 2017), 2017, doi: 10.18653/v1/W17-1408.

[7] N. Martynov, M. Baushenko, A. Kozlova, K. Kolomeytseva, A. Abramov and A. Fenogenova, "A methodology for generative spelling correction via natural spelling errors emulation across multiple domains and languages," In Y. Graham and M. Purver (eds.), *Findings of the Association for Computational Linguistics: EACL 2024*. pp. 138–155. Association for Computational Linguistics, St. Julian's, Malta (Mar 2024), 2024. Available: https://aclanthology.org/2024.findings-eacl.10/

[8] D. Zmitrovich, A. Abramov, A. Kalmykov, V. Kadulin, M. Tikhonova, Taktasheva, E., Astafurov, D., Baushenko, M., Snegirev, A., T. Shavrina, S. S. Markov, V. Mikhailov, and A. Fenogenova, "A family of pretrained transformer language models for Russian," In N. Calzolari, M.Y. Kan, V. Hoste, A. Lenci, S. Sakti and N. Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 507–524. ELRA and ICCL, Torino, Italia (May 2024), 2024. Available: https://aclanthology.org/2024.lrec-main.45/.

[9] Y. Tang, C. Tran, X. Li, P.J. Chen, N. Goyal, V. Chaudhary, J. Gu and A. Fan, "Multilin-gual translation with extensible multilingual pretraining and finetuning," 2020. Available: https://arxiv.org/abs/2008.00401.

Feoktistov Grigory Olegovich, Novosibirsk State University (e-mail: g.feoktistov@g.nsu.ru).

Dmitry Alekseevich Morozov, Novosibirsk State University (e-mail: morozowdm@gmail.com).