Историческая информатика
Правильная ссылка на статью:
Котов А.С. Дообучение модели на основе архитектуры Transformer для нормализации корпуса средневековых текстов на немецком языке XIV-XV вв. из орденской Пруссии // Историческая информатика. 2025. № 4. DOI: 10.7256/2585-7797.2025.4.75275 EDN: XOHQXO URL: https://nbpublish.com/library_read_article.php?id=75275

Дообучение модели на основе архитектуры Transformer для нормализации корпуса средневековых текстов на немецком языке XIV-XV вв. из орденской Пруссии

Котов Антон Сергеевич

ORCID: 0000-0003-3036-5222

кандидат исторических наук

доцент; кафедра истории древнего мира, средних веков и методологии истории; Национальный исследовательский Томский государственный университет

634057, Россия, Томская обл., г. Томск, ул. Говорова, д. 48, кв. 342

waidelot@yandex.ru

Статья из рубрики "Искусственный интеллект и наука о данных"

DOI:

10.7256/2585-7797.2025.4.75275

EDN:

XOHQXO

Дата направления статьи в редакцию:

23-07-2025

Аннотация: Статья посвящена методам автоматической нормализации текстов на средневерхненемецком и раннем нововерхненемецком языка для применения NLP в исследованиях по средневековой истории. В статьей представлен обзор имеющихся подходов автоматической нормализации исторических текстов на немецком языке. Обозначены проблемы нормализации средневековых немецких текстов: особенности применения словарей подстановок, правил замены. Описаны ограничения применения таких подходов и необходимость учёта целей нормализации. Нейроязыковые модели определены как наиболее перспективные для автоматической нормализации. В исследовании проведено сравнение эффективности применения имеющихся нейроязыковых моделей (NMT) в отношении текстов на средневерхненемецком и раннем нововерхненемецком. Показана низкая эффективность использования NMT, обученных на выборке текстов Нового и Новейшего времени. С учетом представленных в

литературе обзоров утверждается необходимость подготовки NMT в зависимости от целей и корпуса. Для нормализации текстов XIV–XV вв., созданных в орденской Пруссии, дообучена нейроязыковая модель на основе архитектуры Transformer (BART) и представлена ее эффективность в сравнении других моделей. Модель обучена на собственной выборке пары слов: оригинальное- нормализованное, список составляет 6570 пар слов. Условия дообучения модели: Epoch = 28; Batch = 50. Для нормализации корпуса текстов на трех исторических формах немецкого языка выбрана модель DTAEC Type Normalizer. Проведено сравнение эффективности нормализации дообученной модели с уже имеющимися моделями, обученными на немецких текстах Нового и Новейшего времени по метрикам Accurancy, Accurancy OOV, CER и Levenshtein distance. Дообученная модель показывает значительную эффективность по сравнению с другими моделями. Предложено к ознакомлению одно нормализованное при помощи модели предложение и проведено сравнение с эталоном. Выявлены факты "галлюцинаций" дообученной модели. При Accurancy и Accurancy OOV равном 89,6 признано перспективным использование этого метода. Однако, выявленные недостатки при нормализации текста указывают на необходимость использовать дополнительные методы нормализации, такие как лемматизиция.

Ключевые слова:

нормализация, И И , transformer, B A R T , средневерхненемецкий, ранний нововерхненемецкий, Немецкий орден, П р у с с и я , Средние века, цифровая гуманитаристика

Введение

В последние годы расширяется цифровая инфраструктура гуманитарных исследований[1]. Цифровизация архивов и публикация текстов источников в цифровом виде увеличивает источниковую базу доступную для полнотекстового поиска и применения методов и инструментов обработки естественного языка (NLP). Рост числа примеров применения text mining для исследования исторических текстов по истории Нового и Новейшего времени поднимает проблему применимости этих подходов для анализа текстов, созданных до появления унифицированной стандартизированной орфографии. Отсутствие унифицированного написания слов становится камнем преткновения для предобработки текста, применения поиска и, следовательно, использования text mining. Ряд инструментов NLP создан для работы с современными языками, например, словари со стоп-словами, которые представлены в современной орфографии, или методы выявления сущностей и пр. Таким образом, исследователь, занимающийся древней и средневековой историей, сталкивается с проблемой приведения корпуса текстов к некоторой норме, т.е. нормализации орфографии.

Предлагаемая работа появилась как решение проблемы при применении инструментов text mining в отношении корпуса исторических источников орденской Пруссии XIV-XV вв. (ASP)[2] — это уникальное собрание актового материала на средневерхненемецком языке с элементами средненижненемецкого и раннего нововерхненемецкого: протоколы заседаний представителей сословий, постановления, донесения и пр. существенной части орденского и городских архивов. К сожалению, во время Второй Мировой войны часть источников была утеряна и, таким образом, многие документы представлены исключительно в издании XIX века. Благодаря сборнику можно проследить социально-

политические изменения в орденских землях, тем более что автор-составитель старался точно передать тексты и графику источников. Вместе с тем именно эта особенность издания не позволяет создавать «мешки слов», поскольку одно и тоже слово может иметь несколько графических форм даже внутри одно документа, например, «einunge», «eynunge», «ainigung», «eynegunge», - всё это совр. «Einigung (объединение)».

Проблема «нормализации» и ранее беспокоила археографов. При издании рукописей нормализация проводилась вручную самим исследователем или издателем. Уже на этом этапе возникал вопрос о том, как следует проводить нормализацию концептуально и технически, например, при сравнении разных списков одного и того же исторического источника, что оставлять в качестве основы. В XIX в. в Германии существовало два конкурирующих подхода т.н. методы Лахманна и Гримм и, не вдаваясь в особенности каждого из них, следует отметить их существенную роль в формировании принципов публикации источников на средневерхненемецком языке[3][4]. В этой связи уже опубликованные источники содержат в себе некоторый слой нормализации.

Современные исследования, производимые на основе изданий источников XIX века, продемонстрировали негативные последствия нормализации: приведение оригинальной формы слова к современной или же ее транскрипцию, т.е изменение темпорального регистра, или сведение графики только к значимым элементам, например, weket-wecket или же bevolhen - befohlen. Как графическое, так фонологическое вмешательство в текст может изменить репрезентативность текста, стереть способы адаптации графики к фонетике, скрыть традицию писцов и регионально-культурные особенности[5]. Дискуссия о принципах издания исторических источников привела к осмыслению проблем определения методов нормализации. Например, сейчас концептуальные и технические вопросы ручной нормализации сказываются на применении методов цифровой истории для средневековых источников [6].

Предложенные два примера указывают на необходимость учитывать не только технические аспекты нормализации, но и ставят под вопрос ее универсальный характер.

В европейском позднем Средневековье, где грамотность предполагала изучение, прежде всего, латыни, тексты на местных языках записывались, следуя разным традициям. Кроме того, по всей Европе сохранялась диалектная специфика. В канцеляриях и скрипториях закладывали и поддерживали определенные принципы написания, что способствовало разнообразию вариантов записи. Из-за этого, несмотря на существования в XIX в. двух подходов нормализации текстов на средневерхненемецком, иногда для издания источников специалистами готовились самостоятельные методы нормализации для конкретной группы текстов [7].

Конечно, существовали факторы, способствовавшие унификации графики, как например, канцелярия императора Священной Римской империи, поддерживая контакты с разными землями, с одной стороны, учитывала специфику региона, с другой стороны, распространяла свою рукописную традицию, однако, регионализмы были сильны [8]. Это также сказывается на выработке общих принципов нормализации – всегда остаются исключения.

В Средневековье в германских землях при записи писари в основном следовали фонетическому принципу, что позволяет относительно легко идентифицировать слова как современному специалисту, так и адресату того времени. Вместе с тем, существовали уникальные формы, обладающие региональной и диахронической спецификой. Сложные

случаи собраны в специализированных словарях (см. Mittelhochdeutsches Handwörterbuch von Matthias Lexer, digitalisierte Fassung im Wörterbuchnetz des Trier Center for Digital Humanities, Version 01/25. URL: https://www.woerterbuchnetz.de/Lexer (дата обращения: 02.08.2025); Frühneuhochdeutschen Wörterbuches (FWB). URL: https://fwb-online.de/ (дата обращения: 02.08.2025); Deutsches Rechtswörterbuch. URL: https://drw.hadw-bw.de/drw-cgi/zeige (дата обращения: 02.08.2025)), однако, словари также не охватывают все возможные варианты написания. Кроме того, словарные формы являются результатом развития принципов нормализации XIX века и соответствуют нормам академической орфографии средневерхненемецкого (1050-1350 гг.) и раннего нововерхненемецкого (1350-1650 гг.) для отражения фонетических особенностей. Для одного и того же термина предлагается разная орфография в зависимости от исторической эпохи, что, в свою очередь, накладывает ограничение на применение словарей для нормализации текстов, используемых в историческом исследовании, которое включает в себя оба периода развития письменности, определенных лингвистикой.

Благодаря цифровым технологиям стала доступна автоматическая нормализация текстов, написанных задолго до появления унифицированной орфографии. Попытки создания способов автоматической нормализации исторических текстов на разных европейских языках берут свое начало с 1980-х гг.[9]. Эти попытки можно варьировать по шкале от создания методов нормализации для определенного корпуса текстов до создания языковых нейросетевых моделей, претендующих на универсальный характер для определенного языка и времени. Последние несколько обзоров известных методов и моделей придерживаются позиции необходимости оценки эффективности методов в каждом конкретном случаи [10][11]. Кроме того, в большинстве случаев эффективные методы нормализации ориентированы на тексты начиная с XVI века. Именно книгопечатание оказалось тем институтом распространения и поддержания унифицированного написания, задолго до появления прескриптивных изданий по орфографии, поэтому универсальные модели нормализации лучше применимы к текстам с началом Нового времени.

Предпринятые попытки автоматической нормализации средневековых текстов также показывают низкую эффективность в сравнении с проектами, ориентированными на тексты Нового и Новейшего времени. Так, оценка трех методов нормализации средневековых немецких текстов показала Acc≈76,2 при Acc≈85,76 для текстов с XVI в. [12][13].

Цель и задачи, предмет и объект исследования

Рост значения цифрового инструментария для исторических исследований и некоторые попытки автоматической нормализации средневековых текстов, указывающие на преимущества и ограничения цифровых методов нормализации, определяют цель работы – выработать метод автоматической нормализации текстов, созданных в орденской Пруссии XIV-XV вв. на средневерхненемецком, как этап предварительной подготовки корпуса для применения инструментов обработки естественного языка. Достижение поставленной цели возможно при выполнении ряда задач. Прежде всего, это выявление наиболее эффективного метода автоматической нормализации с применением методов оценки эффективности нормализации. Кроме того, необходимо выработать качественные критерии нормализации, принимая во внимание то, что к нормализованным текстам далее планируется применение NLP. Наконец, последней задачей является представление цифрового метода нормализации.

Объектом исследования являются методы обработки исторических текстов, в частости, подготовительный этап, обладающий существенным значением для текстов на реликтовых языках и языках с нестандартизированной орфографией, а предметом исследования, таким образом, — это модель нейросетевого обучения для нормализации орфографии и распознавания средневековых текстов на немецком языке XIV-XV вв. из орденской Пруссии.

В исследовании впервые выработан метод нормализации, ориентированный, прежде всего, на определенный, узкий, текстовый корпус, в отличие от предшествующих попыток обучить нейросетевую модель для всех текстов на определенном языке определенного исторического периода. Учитывая доминирование локальных «школ» писарей в средневековой Европе, такое решение, кажется, более перспективным для увеличения точности нормализации и предсказания. Кроме того, в предложенном исследовании предлагается увеличить объем обучающей выборки, что ранее не предпринималось. На данный момент опубликовано только одно исследование нейросетевой модели для нормализации текстов на средневерхненемецком, которая была обучена на 2 500 парах слов: оригинальное-нормализованное [14].

Материалы и методы

Вместе с тем накопленный опыт автоматической нормализации текстов остается полезным для разработки соответствующего подхода и выработки стратегии нормализации для определенных задач. В литературе выделяются 6 подходов: словари подстановки (substitution list), метод замен на основе правил (Rule-based Method), метод, основанный на измерении расстояния Левенштейна (Distance-based Method), статистические модели (Statistical Models) и два типа нейросетевых языковых моделей (neural machine translation, NMT) [15].

Первый подход предполагает использование словарей подстановки [16]. Из-за регионализмов и разнообразия словоформ словари подстановки ожидаемо дают сравнительно низкую результативность. Этот подход требует выгрузки оригинальных слов (токенов) и формирование словаря с парой слов: оригинального и нормализованного, после чего проведение автоматической замены. Причем сопоставление оригинальной с нормализованной формой может носить ручной характер, поскольку привлечение специализированных академических словарей ограничено из-за отсутствия машиночитаемой версии. Относительно периодов развития немецкого языка реализованы два проекта по созданию машиночитаемых словарей и параллельных корпусов для средневерхненемецкого и раннего нововерхненемецкого (Referenzkorpus Mittelhochdeutsch (1050–1350). URL: https://www.linguistics.rub.de/rem/ (дата обращения: 02.08.2025); Referenzkorpus Frühneuhochdeutsch (1350–1650). URL: https://www.linguistics.rub.de/ref/ (дата обращения: 02.08.2025)). Однако, применение каждого словаря возможно только для соответствующего периода истории развития языка и вносит серьезные искажения при использовании для корпуса источников, охватывающего оба периода.

Фонетический принцип записи позволяет обратиться к следующему подходу нормализации текстов, который предполагает выработку некоторых правил изменения словоформы[17][18][19], например, czu – zu, ew – eu, v-f и т.д. Однако эти правила не всегда работают. Например, v может использоваться для передачи звуков [u], [f], [v] – bevohlen, Brunav, convente.

Деловая коммуникация предполагала взаимопонятность, но внутри социального круга, в котором курсируют сообщения, поэтому для выявления и применения таких правил необходимо учитывать место, время и круг участников коммуникации. Если в корпусе присутствуют тексты, созданные в разных традициях, то правила могу либо не затрагивать часть случаев, либо приводить к конкуренции правил между собой, что, в свою очередь, потребует выбора верного варианта или аннотирования со стороны специалиста. Тем не менее выявление некоторой группы правил является задачей в нейронном машинном переводе с применением длительной коротковременной памятью (LSTM)[20], а также в машинном переводе с помощью нейросетевых моделей [21][22].

Как показали сравнительные исследования, нейросетевые языковые модели (NMT) с машинным обучением являются наиболее эффективными для нормализации немецких текстов по отношению к уже перечисленными подходами: словарей подстановки, применение правил и статистических моделей, поскольку они так или иначе включают все варианты [23]. Для обучения могут быть использованы три подхода: 1) выявление последовательности букв в словах (Type-Bases Method), 2) выявление форм слов в предложении (Sentence-Based Method) – контекстуальное обучение и 3) гибридный, совмещающий оба подхода. Причем, последний подход на всё той же выборке дает более точную нормализацию [24][25][26].

Таким образом, для нормализации текстов наиболее эффективны нейросетвые модели (NMT). Для немецкоязычных источников создано три языковые модели (NMT), в основу которых положены как первый, так и третий принципы: Transnomer (Transnomer. A lexical normalizer for historical spelling variants using a transformer architecture. URL: https://github.com/ybracke/transnormer (дата обращения: 02.08.2025)), hybrid_textnorm (hybrid_textnorm Text normalization with hybrid model architecture. URL: https://github.com/aehrm/hybrid_textnorm (дата обращения: 02.08.2025); Ehrmanntraut A. Historical German Text Normalization Using Type-and Token-Based Language Modeling // arXiv:2409.02841v2 [cs.CL]. 25 Feb 2025. P. 1-27. URL: https://arxiv.org/abs/2409.02841 (дата обращения: 02.08.2025)), DTAEC Type Normalizer (aehrm/dtaec-type-normalizer. URL: https://huggingface.co/aehrm/dtaec-type-normalizer (дата обращения: 02.08.2025)). Все перечисленные языковые модели разработаны на базе BART - трансформерной модели с архитектурой энкодер-декодер (seq2seq), широко используемой для генерации текста и задач нормализации. Благодаря способности к генерации модель после обучения может предсказывать корректные формы слов. Transnomer и DTAEC Type Normalizer обучены по принципу выявления последовательности букв в словах (Type-Bases Method), а в hybrid_textnorm применяются обе обучающие модели. Каждая модель была применена для нормализации выборки слов из ASP и показала сравнительно низкую эффективность как по точности, так и по предсказанию (см. таб. 1, рис. 1, 2, 3). Такой результат ожидаем, поскольку указанные нейросетевые модели были обучены на корпусе параллельных немецких текстов, изданных в период с 1780 по 1901 гг. - German Text Archive (Deutsches Textarchiv (DTA). URL: https://www.deutschestextarchiv.de/ (дата обращения: 02.08.2025)) [27].

Название модели	WordAcc		WordAcc OOV		Levenshtein distance	CER
	Pretrained dataset	Finetuned model dataset	Pretrained dataset	Finetuned model dataset		
Transnormer	98.979	22,57	91.653	13,76	1.5131	0.2017
Pretrained	95.46	22,29	90.96	13,42	1.5299	0.2039
Hybrid	99.194	25.31	91.701	16.67	1.5110	0.2014

Hybrid	99.194	25,31	91.701	16,07	1.5110	0.2014
Finetuned	X	89,60	X	89,65	0.1464	0.0195

Таблица 1. Сравнительная таблица эффективности нейроязыковых моделей. Pretrained dataset – результаты, полученные на основе датасета разработчиков; Pretrained dataset – результаты, полученные на основе собственного датасета, собранного на основе ASP.

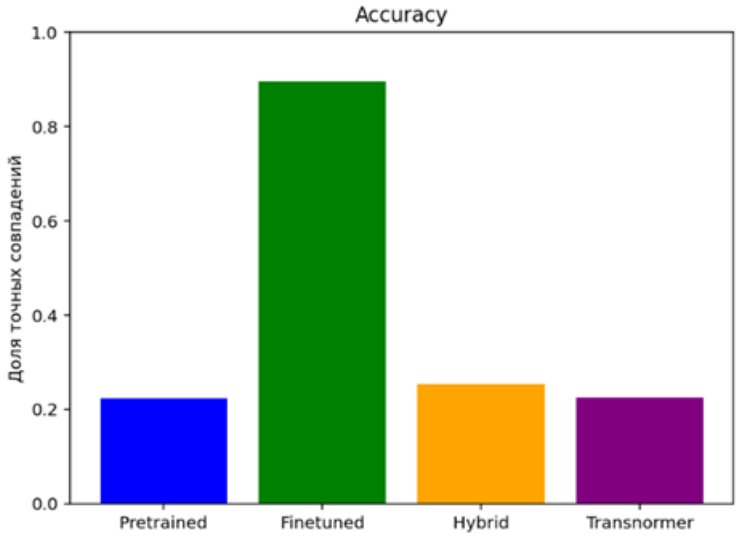

Рис. 1. Сравнение NMT моделей для немецкоязычных текстов по точности предсказания в соответствии формой из обучающей выборки (Acc.)

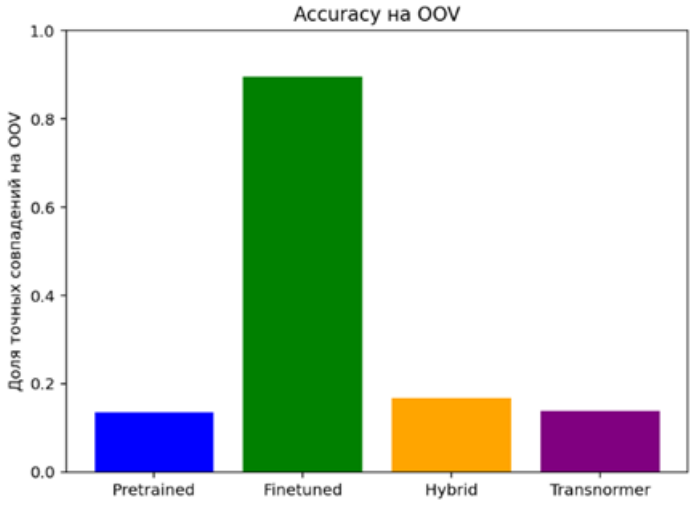

Рис. 2. Сравнение NMT моделей для немецкоязычных текстов по точности в предсказании формы, отсутствующей в обучающей выборке (Acc. OOV)

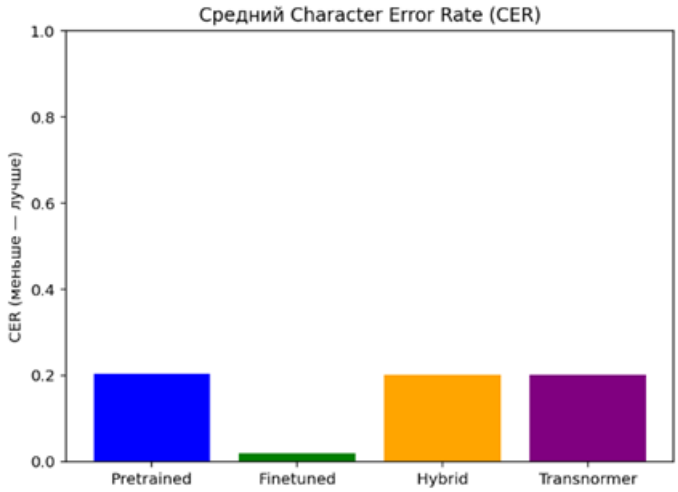

Рис. 3. Сравнение Сравнение NMT моделей для немецкоязычных текстов по CER

В связи с отсутствием доступного эффективного метода нормализации была дообучена модель DTAEC Type Normalizer.

Как показывает обзор нормализации средневековых текстов [28][29], независимо от метода нормализации первостепенным значением обладает цель и задачи исследования, в котором используется нормализованный текст. Именно это определяет содержание понятия и методов нормализации. Одной из задач нашего исследования является определение корреляция понятий (ключевых слов), поэтому для обработки текста важны одинаковые формы слов во всем корпусе исторических источников, чтобы можно было применять «мешок слов». По этой причине нормализация в данном случае означает сокращение разнообразия графических форм до одного с учетом словарной формы, при этом формы служебных слов: предлоги, артикли, наречия и пр., - не учитываются, так как не несут смысловой нагрузки для анализа. Последняя оговорка существенна, потому что, например, в немецком языке омонимичны слова «в» (+ Masc. Dat. Sg.) и «ему»: im/ihm, «в» и «его, им»: in/ihn, при этом в актовом материале может использоваться in и его значение определяется контекстом. Помимо этого, не имеет значение регистр и пунктуация. Принимая во внимания влияние регионального и диахронических факторов, целью нормализации было приведение форм к близки современному немецкому языку, за исключением тех случаев, когда слово более не используется. Однако, диакритические символы не использовались.Таким образом, была выбрана только модель, созданная для нормализации последовательности букв (Type-Based). Применение других методов обучения ограничено в связи с отсутствием параллельных корпусов.

Для дообучения выбранной модели вручную был создан аннотированный список, состоящий из 3921 пары слов: оригинальное-нормализованное в формате таблицы .xlsx. Слова были выгружены из ASP, затем для повышения устойчивости модели список пар был дополнен за счет дублирования некоторых из них до 6570. Условия дообучения модели: Epoch = 28; Batch = 50.

Результаты и их обсуждение

В результате была получена дообученная модель на основе DTAEC Type Normalizer – finetuned (URL: https://huggingface.co/Antnis/text-normalization-for-german-order-acts/tree/main (дата обращения: 02.08.2025)). Эффективность модели была проверена

по четырем метрикам: WordAcc, WordAcc OOV, Levenshtein distance, CER. Низкий уровень точности характерен для всех 4 моделей. Это связано с графическим разнообразием словоформ, что приводит к погрешностям при определении принципов записи. Однако, дообученная модель, как следует из метрик, справляется с задачей лучше.

Модель finetuned была применена для нормализации нескольких текстов из ASP. В качестве примера представлено одно предложение из выборки. Как следует из таб. 2. расстояние Левенштейна между нормализованный текстом и эталоном 9, CER – 0,0563, при этом расстояние Левенштейна между оригиналом и эталоном составляет 27, а CER – 0,1667, что примерно в 3 раза улучшило текст.

Вместе с тем, языковые модели, построенные на архитектуре BART, имеют склонность с «галлюцинациям», что хорошо видно по примеру одного из нормализованных предложений. Несмотря на наличие в обучающей выборки пар wart-ward, при нормализации была выбрана форма глагола sein в Präteritum; форма rethe при наличии ее в обучающей выборке также при нормализации перешла в ложную форму rate исходная форма. Как в обучающей выборке, так и в корпусе текстов встречается значительное разнообразие форм, что, в свою очередь, ожидаемо снижает вероятность воспроизведения одинаковых форм при нормализации.

Оригинал Hiruff mir geantwert wart durch des keysers rethe und ouch durch unsern doctorem, is mochte nicht gesein uff diese czeit die weyle wir in hengendem rechte sein, sundir dornoch findet man wol rot.
Эталон Hierauf mir geantwortet ward durch des keisers rat und auch durch unsern doktor, es mochte nicht gesein auf diese zeit die weile wir in hangendem rechte sein, sonder danach findet man wohl rat.
Нормализованный текст Hierauf mir *geantwort (2)* war(1) durch des keisers *rate (1)* und auch durch unsern daktorem (3), es mochte nicht gesein auf diese zeit die weile wir in hengendem (1) rechte sein, sonder danach findet man wohl ***rot** (1)***.

Таблица 2. Пример нормализации текста при помощи модели finetuned. Предложение взято из ASP [30].

Дообучение нейросетевой модели для нормализации корпуса текстов на средневерхненемецком языке дает неоднозначные результаты. С одной стороны, дообученная модель нормализации исторических текстов демонстрирует улучшение точности (WordAcc по сравнению с исходными моделями) и дают надежду на перспективность применения этого подхода за счет увеличения аннотированного списка пар и увеличение эпох обучения. Это подтверждает перспективность подхода и позволяет рассматривать переход к контекстуальным методам обучения (Sentence-Based Method). Нормализация текстов на средневерхненемецком при помощи улучшенной модели предоставляет корпус текстов для последующего обучения с параллельным корпусом с минимальным участием человека для редактирования.

С другой стороны, даже дообученная модель требует постобработки: из-за низкого WordAcc OOV остается необходимость в проверке экспертом. Улучшение дообученной

модели и требует дополнительных метод обработки, например, лемматизации или комбинирования BART с другими подходами.

Библиография

1. Burch Th. Infrastrukturprojekte zur digitalen Lexikographie. Vorgestellt am Beispiel des Zentrums für Historische Lexikographie // Digitale Mediävistik. Perspektiven der Digital Humanities für die Altgermanistik / Hrsg. Gabriel Lienert, Elisabeth Hamm, Joachim Hausmann, Albrecht Viehhauser. Oldenburg, 2022. (BmE Themenheft 12). S. 97-108.
2. Acten der Ständetage Preußens unter der Herrschaft des Deutschen Ordens / Hrsg. von M. Toeppen. Bd. I-V. Leipzig, 1878–1886.
3. Ehrismann O., Rmange H. Mittelhochdeutsch: Eine Einführung in das Studium der deutschen Sprachgeschichte. Tübingen, 1976. S. 28-29.
4. Primavesi O., Bleuler A.K. Einleitung: Lachmanns Programm einer historischen Textkritik und seine Wirkung // Lachmanns Erbe. Editionsmethoden in klassischer Philologie und germanistischer Mediävistik. Berlin, 2022. S. 11-107.
5. Kragl F. Normalmittelhochdeutsch. Theorieentwurf einer gelebten praxis // Zeitschrift für Deutsches Altertum und Deutsche Literatur. 2015. T. 144, № 1. S. 1-27.
6. Atzenhofer-Baumgartner F., Kovacs T. Is text normalization relevant for classifying medieval charters? // Antonacopoulos A., et al. Linking Theory and Practice of Digital Libraries. TPDL 2024. Lecture Notes in Computer Science. V. 15178. Springer, Cham, 2024. P. 126-127.
7. Kragl F. Normalmittelhochdeutsch. Theorieentwurf einer gelebten praxis // Zeitschrift für Deutsches Altertum und Deutsche Literatur. 2015. T. 144, № 1. S. 26.
8. Ehrismann O., Rmange H. Mittelhochdeutsch: Eine Einführung in das Studium der deutschen Sprachgeschichte. Tübingen, 1976. S. 45-48.
9. Fix H. Automatische Normalisierung – Vorarbeit zur Lemmatisierung eines diplomatischen altisländischen Textes // Maschinelle Verarbeitung altdeutscher Texte. Beiträge zum dritten Symposion, Tübingen, 17–19. Februar 1977. Ed. by Paul Sappler, Erich Straßner. Tübingen, 1980. S. 92-100.
10. Bollmann M. A Large-Scale Comparison of Historical Text Normalization Systems // Proceedings of NAACL-HLT. Minneapolis, 2019. P. 3885.
11. Bawden R., Poinhos J., Kogkitsidou E., Gambette Ph., Sagot B., Gabay S. Automatic Normalisation of Early Modern French // Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022). Marseille, 2022. P. 3354.
12. Korchagina N. Normalizing Medieval German Texts: from rules to deep learning // Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language. Gothenburg, 2017. P. 16.
13. Bollmann M. A Large-Scale Comparison of Historical Text Normalization Systems // Proceedings of NAACL-HLT. Minneapolis, 2019. P. 3893.
14. Korchagina N. Normalizing Medieval German Texts: from rules to deep learning // Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language. Gothenburg, 2017. P. 15.
15. Bollmann M. A Large-Scale Comparison of Historical Text Normalization Systems // Proceedings of NAACL-HLT. Minneapolis, 2019. P. 3886-3887.
16. Fix H. Automatische Normalisierung – Vorarbeit zur Lemmatisierung eines diplomatischen altisländischen Textes // Maschinelle Verarbeitung altdeutscher Texte. Beiträge zum dritten Symposion, Tübingen, 17–19. Februar 1977. Ed. by Paul Sappler, Erich Straßner. Tübingen, 1980. S. 92-100.
17. Rayson P., Archer D., Smith N. VARD versus Word: A comparison of the UCREL variant detector and modern spell checkers on English historical corpora // Proceedings of the

Corpus Linguistics Conference CL2005. Birmingham, 2005. URL: https://eprints.lancs.ac.uk/id/eprint/12686/ (дата обращения: 02.08.2025).
18. Gotscharek A., Reffle U., Ringsltetter C., Schulz K.U., Neumann A. Towards information retrieval on historical document collections: The role of matching procedures and special lexica // International Journal on Document Analysis and Recognition. 2011. T. 14, № 2. P. 159-171. DOI: 10.1007/s10032-010-0132-6 EDN: GWJMQK.
19. Korchagina N. Normalizing Medieval German Texts: from rules to deep learning // Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language. Gothenburg, 2017. P. 12-17.
20. Bollmann M., Bingel J., Søgaard A. Learning attention for historical text normalization by learning to pronounce // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. V. 1. Vancouver, 2017. P. 332-344.
21. Tang G., Cap F., Pettersson E., Nivre J. An Evaluation of Neural Machine Translation Models on Historical Spelling Normalization // Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, 2018. P. 1320-1331.
22. Wu L., Cheng S., Wang M., Li L. Language Tags Matter for Zero-Shot Neural Machine Translation // Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. C. 3001-3007. URL: https://aclanthology.org/2021.findings-acl.264.pdf
23. Bollmann M. A Large-Scale Comparison of Historical Text Normalization Systems // Proceedings of NAACL-HLT. Minneapolis, 2019. P. 3889.
24. Bollmann M. A Large-Scale Comparison of Historical Text Normalization Systems // Proceedings of NAACL-HLT. Minneapolis, 2019. P. 3887.
25. Bawden R., Poinhos J., Kogkitsidou E., Gambette Ph., Sagot B., Gabay S. Automatic Normalisation of Early Modern French // Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022). Marseille, 2022. P. 3355-3356.
26. Ehrmanntraut A. Historical German Text Normalization Using Type-and Token-Based Language Modeling // arXiv:2409.02841v2 [cs.CL]. 25 Feb 2025. P. 11-27. URL: https://arxiv.org/abs/2409.02841 (дата обращения: 02.08.2025).
27. Jurish B., Ast H. Using an Alignment-based Lexicon for Canonicalization of Historical Text // Historical Corpora: Challenges and Perspectives. V. 5. Tübingen, 2015. P. 197-208.
28. Atzenhofer-Baumgartner F., Kovacs T. Is text normalization relevant for classifying medieval charters? // Antonacopoulos A., et al. Linking Theory and Practice of Digital Libraries. TPDL 2024. Lecture Notes in Computer Science. V. 15178. Springer, Cham, 2024. P. 130-131.
29. Fix H. Automatische Normalisierung – Vorarbeit zur Lemmatisierung eines diplomatischen altisländischen Textes // Maschinelle Verarbeitung altdeutscher Texte. Beiträge zum dritten Symposion, Tübingen, 17.-19. Februar 1977. Ed. by Paul Sappler, Erich Straßner. Tübingen, 1980. S. 92-100.
30. Acten der Ständetage Preußens unter der Herrschaft des Deutschen Ordens / Hrsg. von M. Toeppen. Bd. III. Leipzig, 1882. S. 635.

Результаты процедуры рецензирования статьи

В связи с политикой двойного слепого рецензирования личность рецензента не раскрывается.
Со списком рецензентов издательства можно ознакомиться здесь.

Предметом исследования в рецензируемом исследовании выступает модель нейросетевого обучения для нормализации орфографии и распознавания средневековых текстов на немецком языке XIV-XV вв. из орденской Пруссии, в публикации основное внимание сфокусировано на процессе дообучения модели для

анализа текстов, созданных до появления унифицированной стандартизированной орфографии.

Методология исследования базируется на сравнительном анализе существующих методов нормализации (словари подстановки, rule-based подходы, статистические модели, нейросетевые методы), обосновании выбора дообучения модели DTAEC Type Normalizer путем создания датасета из пар словоформ (оригинал–норма) и дообучения модели с учетом специфики средневековых текстов.

Актуальность работы авторы связывают с тем, что цифровизация архивов и публикация текстов источников в цифровом виде увеличивает источниковую базу, доступную для полнотекстового поиска и применения методов и инструментов обработки естественного языка, но отсутствие унифицированного написания слов препятствует предобработке текста, применению поиска и использованию технологии интеллектуального анализа текста.

Научная новизна работы, к сожалению, авторами четко не сформулирована.

В публикации отмечено, что е европейском позднем Средневековье грамотность предполагала изучение, прежде всего, латыни, и тексты на местных языках записывались по разным традициям, сохранялась диалектная специфика, а в канцеляриях и скрипториях закладывались и поддерживались определенные принципы написания, что способствовало разнообразию вариантов записи. Благодаря цифровым технологиям начиная с начало с 1980-х гг. стала доступна автоматическая нормализация текстов, написанных задолго до появления унифицированной орфографии, но попытки автоматической нормализации средневековых текстов показывают низкую эффективность в сравнении с проектами, ориентированными на тексты Нового и Новейшего времени. Авторы отмечают, что существующие модели, обученные на текстах Нового времени, плохо справляются со средневековыми источниками и рассматривают возможность адаптации существующих моделей для распознавания средневерхненемецких и ранненововерхненемецких текстов.

Библиографический список включает 40 источников – научные публикации зарубежных авторов по рассматриваемой теме на иностанных языках языках, а также интернет-ресурсы. В тексте публикации имеются адресные отсылки к списку литературы, подтверждающие наличие апелляции к оппонентам.

Из недостатков и резервов улучшения статьи стоит отметить следующие. Во-первых, текст статьи не структурирован надлежащим образом, в нем не выделены такие общепринятые в современных научных публикациях разделы как Введение, Материал и методы, Результаты и их обсуждение, Выводы или Заключение. Во-вторых, в статье не отражены такие важные элементы методологического аппарата любого научного исследования, как цель и задачи, предмет и объект, не сформулирована рабочая гипотеза, отсутствуют четкие формулировки элементов приращения научного знания, а также значимости полученных результатов для практики в современных условиях. В-третьих, в разделе «Библиография» следует соблюсти рекомендации издательства по оформлению списка литературы: «В списке литературы не указываются … Интернет-источники, включая информацию с сайтов… Все вышеперечисленные источники упоминаются в тексте статьи в скобках, наряду с прочими комментариями и примечаниями авторов» – это касается источников под номерами 9-11, 19-20, 332, 34, 36, 39. Кроме этого название таблицы 1 надо указать не после, а перед таблицей.

Тема статьи актуальна, материал безусловно соответствует тематике журнала «Историческая информатика», может вызвать интерес у читателей, но нуждается в доработке в соответствии с высказанными замечаниями.

Результаты процедуры повторного рецензирования статьи

В связи с политикой двойного слепого рецензирования личность рецензента не раскрывается.
Со списком рецензентов издательства можно ознакомиться здесь.

Представленная статья на тему «Дообучение модели на основе архитектуры Transformer для нормализации корпуса средневековых текстов на немецком языке XIV-XV вв. из орденской Пруссии» посвящена актуальному вопросу выбора методов автоматической нормализации текстов, созданных в орденской Пруссии XIV-XV вв., как этап предварительной подготовки корпуса для применения инструментов обработки естественного языка.
В статье представлен широкий анализ литературных зарубежных источников по теме исследования. Список литературы содержит 30 источников, на каждый источник в тексте имеется ссылка.
Авторами в статье четко сформулированы цель, задачи, предмет и объект исследования. Также по тексту прослеживается новизна исследования, которая заключается в том, что впервые выработан метод нормализации, ориентированный на определенный текстовый корпус, в отличие от предшествующих попыток обучить нейросетевую модель для всех текстов на определенном языке определенного исторического периода.
Стиль и язык изложения материала является научным и доступным для широкого круга читателей. Статья по объему соответствует рекомендуемому объему от 12 000 знаков.
Статья достаточно структурирована – в наличии введение, внутреннее членение основной части (цель и задачи, предмет и объект исследования; материалы и методы; результаты и их обсуждение).
Проанализировав литературные источники, авторы выделяют 6 подходов: словари подстановки, метод замен на основе правил, метод, основанный на измерении расстояния Левенштейна, статистические модели и два типа нейросетевых языковых моделей. Результаты сравнения эффективности нейроязыковых моделей авторами представлены в виде таблицы. Также в статье имеется графический материал в виде диаграмм, на которых изображены результаты сравнения NMT моделей для немецкоязычных текстов по точности предсказания в соответствии формой из обучающей выборки; сравнение NMT моделей для немецкоязычных текстов по точности в предсказании формы, отсутствующей в обучающей выборке; сравнение NMT моделей для немецкоязычных текстов по CER.
Рассмотренные авторами методы обработки исторических текстов, а также модель нейросетевого обучения для нормализации орфографии и распознавания средневековых текстов на немецком языке XIV-XV вв. из орденской Пруссии легли в основу их исследования. В заключительной части авторы указывают, что разработанный метод дообучения нейросетевой модели для нормализации корпуса текстов на средневерхненемецком языке дает неоднозначные результаты. Дообученная модель нормализации исторических текстов демонстрирует улучшение точности и является перспективной для применения этого подхода за счет увеличения аннотированного списка пар и увеличение эпох обучения. Также дообученная модель требует постобработки, включая необходимость в проверке экспертом.
Статья «Дообучение модели на основе архитектуры Transformer для нормализации корпуса средневековых текстов на немецком языке XIV-XV вв. из орденской Пруссии» может быть рекомендована к публикации в журнале «Историческая информатика».