Digital archive of the literary magazine with the pre-reform spelling «Otechestvennye Zapiski» (1839-1884)

Zakovorotnaia E. M.
Higher School of Economics /
Moscow, Russia
haylin65@yandex.ru

Voloshina E. Y.
Higher School of Economics /
Moscow, Russia
vokat@mail.ru

Klyshinsky E. S.
Higher School of Economics /
Moscow, Russia
klyshinsky@mail.ru

Kim J. L.
Higher School of Economics /
Moscow, Russia
juliakimjk@live.com

Kudryavtseva P. S.
Higher School of Economics /
Moscow, Russia
pkspacewalker@gmail.com

Abstract

The paper describes an initial version of the digital archive of the literary magazine with the pre-reform orthography «Otechestvennye Zapiski». Today, the corpus contains 10 XML-volumes of the literary magazine (~ 2 mil. words). The web-application of the digital archive allows users to search for words and lemmas in corpus and to edit magazine's texts online. The future optimization of the digital archive includes expanding the composition of corpus and switching modes between pre-reform and modern orthography.

Keywords: Otechestvennye Zapiski; digital archive; spellchecker; corpus
DOI: 10.28995/2075-7182-2021-20-1239-1244

Цифровой архив литературного журнала с дореформенной орфографией «Отечественные Записки» (1839-1884)

Заковоротная Е.М.
НИУ-ВШЭ / Москва, Россия
haylin65@yandex.ru

Волошина Е. Ю.
НИУ-ВШЭ / Москва, Россия
vokat@mail.ru

Клышинский Э.С.
НИУ-ВШЭ / Москва, Россия
klyshinsky@mail.ru

Ким Ю.Л.
НИУ-ВШЭ / Москва, Россия
juliakimjk@live.com

Кудрявцева П. С.
НИУ-ВШЭ / Москва, Россия
pkspacewalker@gmail.com

Аннотация

В данной статье представлена начальная версия цифрового архива литературного журнала с дореформенной орфографией «Отечественные Записки». Корпус содержит десять томов, размеченных в формате XML,

и насчитывает более 2 млн слов. Для доступа к архиву разработан веб-интерфейс, с помощью которого пользователи смогут проводить поиск по корпусу, скачивать тома «Отечественных Записок» в машиночитаемом формате и редактировать выпуски журнала в режиме онлайн. В будущем планируется расширить цифровой архив и добавить возможность переключать режимы орфографии, с дореформенной на современную.

Ключевые слова: Отечественные записки, цифровой архив, спеллчекер, корпус, XML

1 Введение

Журнал «Отечественные Записки» играет важную роль в изучении социальной, культурной и духовной жизни России XIX века. С 1818 по 1884 гг. в нем публиковали свои произведения В. А. Жуковский, М. Ю. Лермонтов, А. С. Пушкин, В. Г. Белинский, М. А. Бакунин, А. И. Герцен, Н. А. Некрасов, М. Ф. Салтыков-Щедрин, А. С. Островский, Ф. М. Достоевский, Л. Н. Толстой. Кроме художественных произведений, в данном журнале также печатались статьи по гуманитарным и естественнонаучным направлениям, рецензии на работы русских и зарубежных авторов, очерки о путешествиях. Истории издания, опубликованным материалам, цензурированию, политическому и историческому контексту публикации журнала посвящены работы исследователей В. Боград, П. Усова, Н. Емельянова, А. Г. Дементьева, Л. П. Громовой и др. [3-8]. Актуальность данной работы заключается в том, что несмотря на высокую научную значимость материалов, опубликованных в журнале, его выпуски нельзя найти онлайн в удобной для обработки машиночитаемой форме. В сети доступны либо нераспознанные изображения, либо разрозненные PDF-документы с текстовым слоем неприемлемого качества: доля ошибочно распознанных символов там достигает 50 %, а символы дореформенной орфографии, действовавшей в России до 1918, утрачены. Из-за этого невозможны ни адекватный полнотекстовый поиск, ни автоматическая обработка текстов.

Таким образом, для проведения исторических, лингвистических, филологических, социальных исследований с использованием статей «Отечественных записок» требуется разработка открытого, общедоступного цифрового корпуса, отвечающего современным требованиям работы с данными.

2 Состав корпуса

Изначально в «Отечественных записках» публиковались путевые заметки и исследования редактора журнала П.П. Свиньина, а также материалы по истории и географии. С 1839 г., когда владельцем стал А.А. Краевский, «Отечественные записки» приобрели более четко выраженную структуру в виде следующих разделов: «Современная хроника России», «Науки», «Словесность», «Художества», «Домоводство, сельское хозяйство и промышленность вообще», «Критика», «Современная библиографическая хроника», «Смесь», посвященных науке, политике, быту и культуре [6]. Когда в издании «Отечественных записок» начинает принимать участие литературовед и критик В. Г. Белинский с 1839 г., в журнале появляются отрывки из произведений современных авторов, а также рецензии на прозу и поэзию. Небольшим, но существенным изменением является преобразование системы нумерации томов — с 1839 г. отсчет издаваемых выпусков стал вновь начинаться с единицы.

Таким образом, разнообразие тематик, публикация литературных произведений, а также четко выраженная структура журнала повлияли на решение начать сбор материалов именно с 1839 г. На данный момент, корпус содержит 10 томов, выборочно со 2 по 49 номер. В каждом из выпусков насчитывается от 500 до 1100 страниц. Общий объем начальной версии корпуса составляет 2 100 000 слов. Последующее улучшение цифрового архива подразумевает расширение корпуса томами не только второго, но и первого периода существования журнала. Необходимо отметить, что тома в начальной версии корпуса не будут содержать таблиц, поскольку для этого требуется создание дополнительных инструментов разметки.

Цифровой архив состоит из двух основных элементов: базы данных, где хранятся выпуски журнала, а также веб-сайта. Пользователи смогут проводить поиск по словоформам в корпусе и скачивать машиночитаемые выпуски с XML-разметкой. Кроме двух основных функций, веб-ресурс позволит исправлять опечатки в томах журнала в режиме онлайн и сохранять результаты в базу данных.

3 Описание ресурса

Первый этап работы над цифровым архивом заключался в сборе материала, подготовке машиночитаемых данных и составлении корпуса. В ходе поисков были обнаружены два источника выпусков журнала: Google Books и сайт Российской национальной библиотеки (РНБ). Так как у найденных томов в формате PDF не везде присутствовал распознаваемый машиночитаемый слой, первым этапом обработки выпусков журнала стало оптическое распознавание документов (optical character recognition, OCR). Процесс электронного перевода изображений PDF-документов проводился с помощью программы ABBYY FineReader. Перед запуском был настроен формат изображения (A5), а также языковой параметр. Кроме русского в дореформенной орфографии, были выбраны несколько европейских языков, поскольку предварительное изучение материала показало, что в «Отечественных Записках» встречаются отрывки на английском, французском, немецком, испанском, итальянском, греческом и латинском языках. Результатом OCR-обработки стали файлы в формате DOCX.

Важно отметить, что из-за качества изображений страниц журнала в обработанных файлах содержалось множество опечаток, лишних отступов и символов. Кроме того, формат DOCX не удобен для автоматической обработки. Как следствие, следующий этап заключался в удалении опечаток в текстах с помощью волонтеров и привлекаемых платных корректоров. Параллельно, для ускорения процесса редактирования текста разрабатывалась программа автоматической проверки орфографии, или спеллчекер. Более подробное описание алгоритма приводится ниже, в разделе «Разработка спеллчекера».

Сейчас все тексты корпуса вычитаны исключительно вручную, без предварительной или последующей обработки спеллчекером. Как только алгоритм автоматической проверки орфографии будет интегрирован в рабочий процесс, будет проведена повторная обработка вычитанных вручную томов, так как есть вероятность наличия малого количества опечаток, которые могли пропустить волонтеры. Более того, при работе с греческими отрывками возникли трудности из-за сложности распознавания символов. Средний показатель word-error rate для всех томов был рассчитан с помощью библиотеки python-Levenshtein и составил ~0.1975.

По завершению этапа вычитки, DOCX-документы переводились в машиночитаемый формат XML с помощью полуавтоматического алгоритма разметки. Определенные части текста, например, страницы, заголовки, текст и содержание заключались в специальные теги. Таким образом, XML-разметка делит текст тома на две основных части: метаинформация, где указываются название, тип, номер, эпиграф, год издания журнала, типография, фамилии ответственных редакторов, и основную, где публикуются сами тексты и названия разделов. Полный список тегов можно найти в описании проекта на его странице в Github (https://github.com/dhhse/Otechestvennie_zapiski/blob/master/list_of_tags.md). В некоторых выпусках, кроме вышеупомянутой информации, в начале выпуска встречается примечание редактора и оглавление, которые также выделяются тегами. Алгоритм разметки не является полностью автоматическим, так как для каждого тома надо отдельно указывать те страницы в коде, на которых встречается информация о типе, годе публикации, редакторе, номере тома.

Для работы с отредактированными томами журнала был написан скрипт на языке Python. Сначала документы заново переводились в формат PDF. Процедура переконвертирования была обусловлена тем, что DOCX-формат не позволяет автоматическими методами определить номер страницы, не прибегая к ручной работе с файлом через текстовый редактор. Поэтому было принято решение перевести полученный DOCX-файл в PDF-формат для сохранения корректной пагинации. Кроме того, библиотека fitz (для работы с PDF форматом) обладает большим функционалом и удобнее в использовании, чем библиотека python-docx. С помощью методов Python-библиотек fitz и xml.dom извлекалась информация из документа, различным частям которой присваивались теги. Код алгоритма разметки доступен в репозитории проекта по адресу https://github.com/dhhse/Otechestvennie_zapiski/blob/master/making_tei.py. Корпус, состоящий из 10 выпусков журнала в формате XML, доступен для скачивания в репозитории проекта по ссылке https://github.com/dhhse/Otechestvennie_zapiski/tree/master/corpus_Otechestvennye_zapiski.

После разметки XML-файлы были переформатированы в JSON для дальнейшего импорта в NoSQL базу данных MongoDB.

Для доступа пользователя к машиночитаемым выпускам журнала, проведения поиска по корпусу и онлайн-редактирования не вычитанных после OCR-распознавания томов был разработан веб-интерфейс. Он состоит из набора функций для извлечения необходимой информации из базы данных и графического оформления. Пользователю предоставляется информация о содержании корпуса через поисковый запрос. На данный момент, поиск по корпусу осуществляется по словоформам и леммам, написанным в дореформенной орфографии. В ответ на запрос программа выдает страницы из разных выпусков журнала с указанием метаинформации: количество результатов, название и номер тома, год издания, номер журнальной страницы, а также название раздела. В будущем планируется добавить поиск по фразам от 2 до 4 слов в современной орфографии, а также возможность переключения формата выдачи с дореформенной на современную орфографию. Для возможности онлайн-редактирования и исправления опечаток был подключен текстовый редактор Editor.js на языке Javascript (https://editorjs.io/). Отредактированный текст страницы сохраняется в базу данных. Опция онлайн-редактирования будет доступна только для зарегистрированных пользователей (авторизация через логин и пароль).

Примеры сайта можно найти в разделе Приложение. Back-end и front-end веб-интерфейса цифрового архива доступен по этому адресу https://github.com/zijane/web-dev-Otechestvennie-zapiski. В дальнейшем планируется изменение графического оформления, интеграция спеллчекера в back-end, добавление переключения формата выдачи запрошенной информации с дореформенной на современную орфографию.

4 Разработка спеллчекера

Изначально было принято решение обучать спеллчекер на вычитанных томах журнала для контроля качества материала. Тем не менее существуют примеры других программ для исправления опечаток в текстах, которые поддерживают дореформенную орфографию. Например, алгоритм на базе словаря oldrus-ispell [2], для составления которого был использован словарь современного русского правописания А. И. Лебедева. Другой спеллчекер для дореформенной орфографии был построен на основе Google N-grams для русского языка [1]. Материалы для обучения модели были собраны с сайтов wikisource.org, arhivarij.narod.ru, russportal.ru. Так как показатель метрики Precision составляет 81 % [1], было принято решение использовать данную модель в качестве основы.

Еще одной важной составляющей спеллчекера является словарь. Он представляет собой совокупность всех уникальных словоформ, найденных в выпусках журнала «Отечественные Записки», которые были очищены от опечаток и исправлены вручную. В результате модель была обучена на ~ 160 000 уникальных словоформах.

Спеллчекер представляет собой гибридную систему, где используется LSTM-модель, расстояние Левенштейна и набор правил. Для каждого тома составляется набор уникальных словоформ, затем каждый токен проходит через спеллчекер. Важно отметить, что слова короче 5 символов не исправляются. Сначала программа проверяет, есть ли слово в словаре, затем с помощью LSTM-модели рассчитывает вероятность слова, и, если вероятность слова выше нижней границы и ниже верхней, то спеллчекер подбирает кандидатов по расстоянию Левенштейна. В общем случае расстояние Левенштейна считается для трех возможных операций: вставка, удаление или замена. В расстояние Дамерау-Левенштейна также включена операция перестановки, однако в нашей версии спеллчекера используются три базовых операции по причине отсутствия в текстах ошибок, требующих перестановки символов для исправления. В данной версии спеллчекера исправлялись кандидаты на расстоянии 1 или 2 (т. е. отличающиеся на 1 или 2 ошибки от исходного слова). На настоящем этапе разработки спеллчекера модель работает только на уровне слов, поэтому она не исправляет случаи, если одно слово разбито на два при распознавании или если два слова “склеились” в одно. Для слов, чья вероятность не принадлежит заданному отрезку, применялись правила, отделяющие слова в дореформенной орфографии от слов, написанных латиницей.

LSTM-модель, использованная в спеллчекере, предсказывает вероятность последующего символа в слове по формуле:

$$P(word) = \prod_{i=1}^{|word|} \frac{p(char_i|word[:i])}{|word|}$$

Таким образом, входные токены исправлялись в зависимости от показателей верхней и нижней границы вероятностей: если вероятность этого слова высока, то исправлять его не нужно, если же вероятность слишком низкая, то это может быть неправильно распознанное слово из другого языка.

Модель тренировалась на словаре, описанном ранее. В качестве функции потерь для LSTM-модели использовались Cross-Entropy Loss и Perplexity Loss, а в качестве метода оптимизации — Adam. Сама модель представляет из себя однонаправленную LSTM-модель с двумя скрытыми слоями. Входной слой модели принимает эмбеддинги размерности 128, размерность скрытого слоя модели равна 64. Модель показывает среднее качество 0.46 (Cross-Entropy Loss), 0.22 (perplexity) на валидационной выборке.

Качество модели алгоритма проверки орфографии проверялось на двух тестовых выборках из 1000 слов, в одной из которых были только слова из словаря модели, а во второй - только слова, которые не включены в словарь. Модель спеллчекера показывает качество 0.973 (accuracy) для слов, которые есть в словаре. Однако, если словоформы в словаре отсутствуют, то исправление опечаток в тексте осуществляется с помощью LSTM-модели, и тогда accuracy равно 0.508.

На данный момент спеллчекер используется следующим образом: он обрабатывает машиночитаемые тома с разметкой в виде тэгов в формате XML, выдавая файл с меньшим количеством опечаток. Однако алгоритм используется обособленно от остальных процессов обработки текстов. Поэтому, в дальнейшем планируется оптимизировать процесс корректировки текста, добавив спеллчекер в back-end сайта. Описание модели и код доступен по этой ссылке https://github.com/EkaterinaVoloshina/spellchecker_for_pre_1918_russian.

5 Заключение

В статье был описан процесс создания цифрового архива литературного журнала XIX века «Отечественные записки». Данный ресурс позволит исследователям проводить поиск по словоформам по материалу корпуса и исправлять опечатки в не вычитанных томах в режиме онлайн. В будущем планируется расширить состав корпуса, интегрировать спеллчекер в back-end сайта и добавить функцию переключения режимов орфографии, с дореформенной на современную, при выдаче результатов поиска. Доступ к цифровому архиву журнала пользователи смогут получить в конце лета — начале осени 2021 года.

Библиография

[1] Mitrofanova M. (2019), Experiments with Automatic Spelling Correction of Russian Google ngrams, available at: https://github.com/kak-to-tak/Google_rusngram_spellcheck.
[2] Winitzki S. (2013), Old Russian Spellchecker, available at: http://oldrus-ispell.sourceforge.net/.
[3] Боград В. Журнал «Отечественные записки». 1868 – 1884. Указатель содержания. М., 1971.
[4] Боград В. Журнал «Отечественные записки». 1839 – 1848. Указатель содержания. М., 1985.
[5] Громова Л.П. А.А. Краевский – редактор и издатель: учеб. пособие. СПб., 2001.
[6] Дементьев А.Г. Очерки по истории русской журналистики 1840 – 1850-х гг. // М.; Л., 1951.
[7] Демченко А.А. «Отечественные записки» и цензура 1840 – 1880-х гг // Изв. ВГПУ. – 2012. – № 4. – С. 119–122.
[8] Емельянов Н.П. «Отечественные записки» Н.А. Некрасова и М.Е. Салтыкова-Щедрина (1868–1884). Л.: Художественная литература, 1986. 336 с.

Приложение

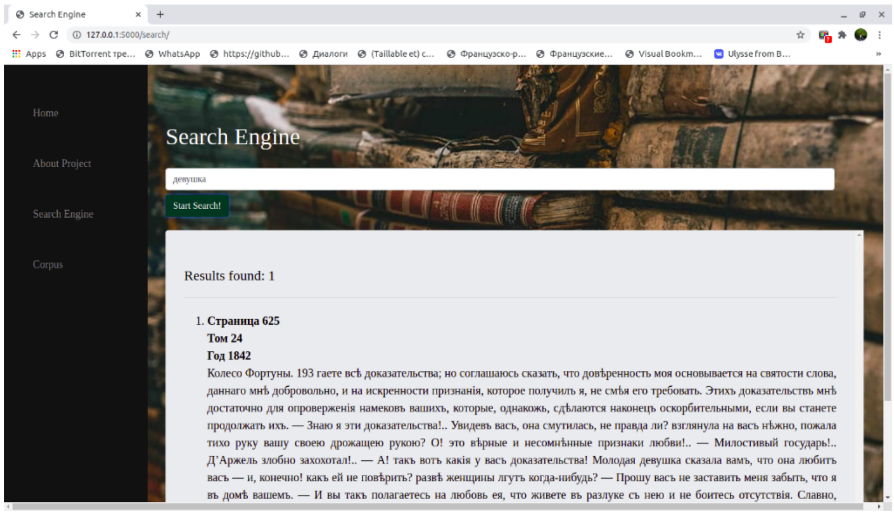

Рис. 1. Пример поисковой выдачи

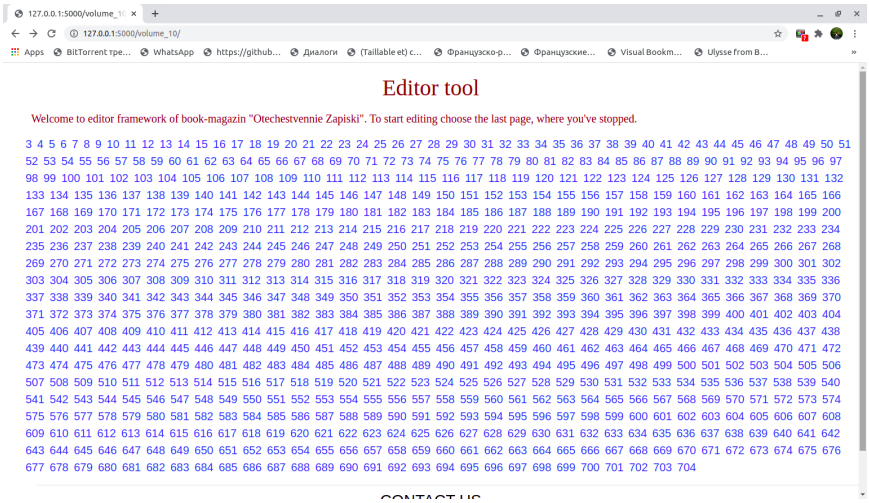

Рис. 2. Пример оформления редакторского фреймворка.

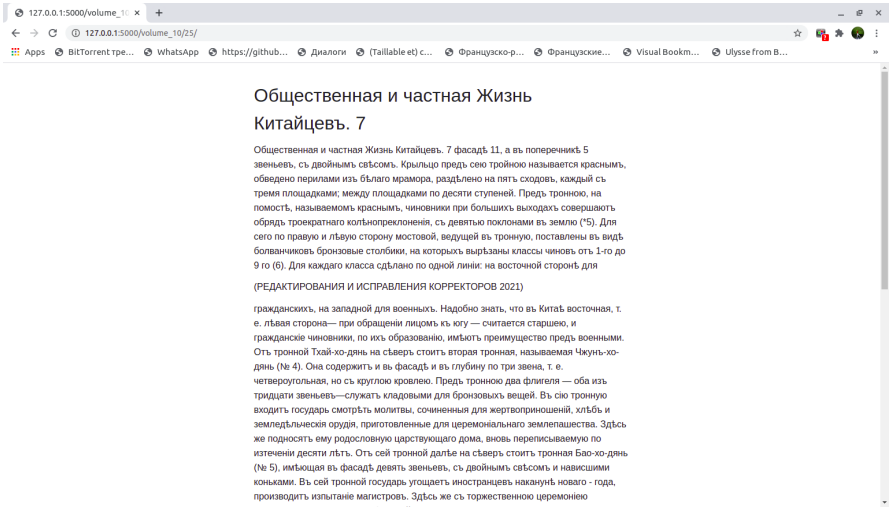

Рис. 4. Пример отредактированной страницы