УДК 004.5

Матвей Михайлович Правосудов, студент, e-mail: matvey@pravosudov.com,
Национальный исследовательский Томский политехнический университет,
Игорь Петрович Скирневский, ассистент, e-mail: skirnevskiy@tpu.ru

ИНФОРМАЦИОННАЯ СИСТЕМА ДЛЯ АНАЛИЗА СЛОВ РУССКОГО ЯЗЫКА «СЛОВОЛИТ»

Аннотация. В статье освещается сервис для автоматического морфологического и морфемного анализа. Дополнительными функциями системы являются подбор синонимов и поиск значений слов русского языка. На данный момент существует ограниченный набор систем, позволяющих выполнять подобный анализ, а большинство алгоритмов нет в открытом доступе. Эти алгоритмы востребованы в сфере школьного образования. В основе логики работы системы лежат оцифрованные словари разборов Тихонова и Ефремовой. Для морфемного анализа разработан новый алгоритм на основе сравнения морфов в слове с морфемами из словаря. Система разработана на веб-стеке технологий, что позволяет использовать ресурс на любом устройстве, включая мобильный телефон или настольный браузер. Основным ограничением системы является отсутствие полного покрытия слов, так как язык постоянно расширяется и изменяется.

Ключевые слова: веб-сервис, морфологический анализ, морфемный анализ, значения слов, русский язык, автоматическая обработка текста.

Matvey Mikhailovich Pravosudov, student, e-mail: matvey@pravosudov.com,
National Research Tomsk Polytechnic University,
Igor Petrovich Skirnevskiy, research engineer, e-mail: skirnevskiy@tpu.ru

INFORMATIONAL SYSTEM FOR ANALYSIS OF RUSSIAN LANGUAGE WORDS CALLED «SLOVOLIT»

Annotation. The present study aimed to show a service for automatic morphological, morphemic and meaning analysis. Today there is a limited set of systems provided the same functionality. Moreover, and most algorithms are not available for public access. The presented system could be used for school education or professional linguists conducting linguistic research. A new algorithm has been developed for morphemic analysis. The algorithm uses Tikhonov's and Efremova's dictionaries to detect morphemic structure of a word. The main idea of the algorithm is comparing morphs in a word with morphemes from a dictionary. The system was developed on a web-based platform and can be used on any device including a mobile phone. The main drawback of the software is limitation of language dictionaries.

Keywords: web-service, morphological analysis, morphemic analysis, word meaning, Russian language, automatic text processing.

ВВЕДЕНИЕ

Русский язык – один из самых сложных для изучения языков. В нем достаточно много правил, но еще больше исключений. Существует ряд инструментов, помогающих людям изучать язык: словари, справочники и учебники. В силу развития технологий появилась масса сервисов в интернете, позволяющих, например, находить значения слов или производить какой-либо анализ. Основных разборов (анализов) несколько:

– Морфологический – разбор слова по частям речи, определение морфологического состава, с последующей характеристикой отдельных слов, относящихся к той или иной части речи.

– Морфемный (структурный) – определение состава слова, например, приставки, корня, суффиксов и т. д.

– Фонетический – анализ слоговой структуры и звукового состава слова.

– Этимологический – определение происхождения слова.

Обычно сервисы по анализу слов работают не в одной «экосистеме», а раздробленно: каждый компонент ничего не знают о другом и работает независимо. Данное ограничение затрудняет комплексный анализ слов.

В данной статье описывается веб-сервис, позволяющий использовать унифицированный интерфейс для комплексного анализа нескольких слов. Главная часть интерфейса — командная строка запросов. Она поддерживает как ввод простого слова, так и расширение функционала при помощи вспомогательных синтаксических конструкций. Кроме того, интерфейс веб-сайта был разработан адаптивным, что позволяет легко использовать его на мобильных устройствах.

Такой сервис может быть использован как учащимися школ в качестве заменителя бумажных справочников, так и учителями русского языка и литературы для быстрого уточнения результатов разборов слов. Студенты и научные сотрдуники в сфере филологических наук смогут экономить время на обращении к бумажому словарю, благодаря быстрой работе сервиса. В силу того, что представленный веб-сервис адаптирован под мобильные устройства, его использование будет удобнее, нежели поиск разборов в нескольких книгах.

Веб-сервис является кросс-платформенным решением, позволяющим использовать его где угодно при наличии интернета. Благодаря адаптивности интерфейса, им пользоваться как на телефоне, так и на планшете или персональном компьютере.

Приложение отличается легкой возможностью масштабирования, а архитектура позволяет легко импортировать новые словари, которые могут подхватываться «на лету», благодаря модульному принципу компонентов. С помощью менеджера пакетов Composer на PHP процесс добавления новой библиотеки занимает секунды.

Кроме этого, функционал разрабатываемого приложения выходит далеко за рамки разбора в его классическом понимании и позволяет одновременно выполнять несколько разборов, или производить комплексный анализ слова с разных сторон в контексте русского языка.

Аналогичные сервисы, представленные на рынке, не предоставляют такого широкого набора возможностей; по больше части они не адаптированы под мобильные устройства и работают медленно. Поэтому, создание веб-сервиса «Словолит» обосновано и необходимо.

1. СМЕЖНЫЕ РАБОТЫ

В данной предметной области наиболее популярны темы синтаксического и морфологического анализа, так как первая используется в анализе и синтезе речи, а вторая являются ключевой частью в алгоритмах нечеткого поиска. Весьма подробно синтаксический анализ предложений русского языка рассматривает И. М. Ножов в диссертации «Морфологическая и синтаксическая обработка текста (модели и программы)» [1]. Автором были разработаны два метода синтаксического анализа предложения, построена прикладная модель синтаксического анализатора, а также разработаны программные компонента для автоматической обработки текста.

Одной из наиболее ярких работ по морфологическому анализу текста является программа «MyStem» за авторством И. Сегаловичем и В. Титовым [2]. Программное обеспечение позволяет строить гипотетические разборы для слов, не входящих в словарь. Документация библиотеки находится на официальном сайте [3].

В решении, описанном в данной статье, для морфологического анализа используется библиотека phpMorphy [4]. Она позволяет решить задачи лемматизации, получения всех форм слова, грамматической информации и изменения слова по образцу или форме. Также, библиотека, как и MyStem, предсказывает характеристики для слов, которых нет в словаре.

Так как «Словолит» является веб-приложением, разумно провести анализ текущего рынка. Веб-сервисы, предоставляющие инструменты для автоматической обработки текста, были проверены на доступность с мобильных устройств, а также определен набор доступных функций. Таблица сравнения приведена ниже.

Как видно из таблицы, анализы в сервисах не связаны между собой: невозможно производить несколько разборов слова на одном экране. Также, интерфейс многих сервисов не адаптирован для мобильных устройств, что является существенным ограничением. Словолит же анализирует несколько слов сразу и открывается на мобильных устройствах.

2. АРХИТЕКТУРА ПРИЛОЖЕНИЯ

Для взаимодействия программных компонентов используется архитектурный стиль REST. Он позволяет отделить интерфейс системы от бизнес-логики и данных, что упрощает расширение функционала системы, а также обновление отдельных компонентов.

Название	Морфол.	Морфем.	Фонет.	Этимол.	Компл. рез.	Адапт.
morphologyonline.ru	●●●	○○○	○○○	○○○	○○○	●●○
uchim.org	●●●	○○○	○○○	○○○	○○○	○○○
goldlit.ru	●●●	○○○	○○○	○○○	○○○	●○○
phoneticonline.ru	○○○	○○○	●●●	○○○	○○○	●●○
vnutrislova.net	●●●	●●●	●●●	○○○	○○○	●●○
slovonline.ru	●●●	●●●	●●●	●●●	○○○	●●○
yznaika.com	○○○	○○○	●○○	○○○	○○○	●○○
russkiy-na-5.ru	○○○	●○○	●○○	○○○	○○○	○○○
vasmer.lexicography.online	○○○	○○○	○○○	●●●	○○○	●●○
etymolog.ruslang.ru	○○○	○○○	○○○	●●●	●○○	○○○
gufo.me	○○○	○○○	○○○	●●●	●○○	●●●
aot.ru	●●●	○○○	○○○	○○○	○○○	○○○

Таблица 1. Сравнение онлайн-сервисов по автоматической обработке текста. ●●● — функция в сервисе реализована качественно; ●●○ — сервис не полностью покрывает набор слов или не очень удобен в использовании; ●○○ — функция сильно ограниченна или интерфейс сильно устаревший; ○○○ — функции нет.

В веб-сервисе реализованы методы API, которые по запросу отдают результат анализа слов. Для стандартизации этих данных используется формат JSON. Компонент, отвечающий за построение интерфейса, запрашивает JSON-данные о результате анализа слова у определенного API-метода, что представлено на Рис. 1.

Рис. 1. Пример ответа API-метода для морфемного анализа слова «предприниматель»

На рисунке опущена строчка в JSON, отвечающая за специфичный код для библиотеки Morfana, которая строит графические обозначения морфем в интерфейсе.

Веб-сервис использует фреймворк Laravel. Он написан по схеме проектирования приложений MVC — Model-View-Controller (Рис. 2). Схема обеспечивает разделение бизнес-логики, интерфейса и модели данных приложения.

Рис. 2. Серверная архитектура веб-приложения на фреймворке Laravel

3. РЕАЛИЗАЦИЯ ВЕБ-ПРИЛОЖЕНИЯ

Любой веб-сервис состоит из двух компонентов: интерфейс и серверная часть [5]. По запросу пользователя какой-либо страницы веб-сайта, сервер запускает необходимые процессы

бизнес-логики, получает данные и генерирует HTML-страницу, которая вместе с CSS-стилями отдается браузеру клиента. Браузер же отрисовывает эту страницу на экране.

Для упрощения верстки стилей CSS, которые и задают вид всех элементов сайта, созданы препроцессоры, которые добавляют в синтаксис CSS переменные, условные операторы, а также функции и многое другое. Подробнее о сравнении некоторых препроцессоров можно прочитать в [6], а посмотреть репозитории самых популярных на Github [7]. Для данного проекта был выбран препроцессор Sass [8], с синтаксисом SCSS, который компилируется с помощью библиотеки для языка программирования Ruby.

Серверных же языков программирования существует множество: Python, Ruby, PHP, NodeJS, Go и другие. Для решения поставленной задачи был выбран язык программирования PHP, как один из самых популярных, не смотря на споры [9] и [10] о перспективах PHP. Причиной выбора языка также послужил используемый в работе фреймворк Laravel [11], написанный на языке PHP.

Все разборы разделены по компонентам, для отделения алгоритмов анализа слов от остальной бизнес-логики, отвечающей, например, за выполнение составной команды поиска.

– **Морфемный анализ** `/async/morphemes/{word}`
  Так как алгоритм реализован с нуля, исходя из потребностей системы, ему посвящена следующая глава этой статьи.

– **Морфологический анализ** `/async/morph/{word}`
  Слово разбирается библиотекой [4], затем заменяются аббревиатуры граммем на их расшифровки. Например, КР_ПРИЛ — краткое прилагательное. К выдаче добавляется начальная форма слова.

– **Подбор синонимов** `/async/synonyms/{word}`
  Происходит поиск начальной форма слова из [4] в словаре Абрамова [12]. Формируется выдача.

– **Значение слова** `/async/meaning/{word}`
  Как и при подборе синонимов, происходит поиск начальной форма слова из [4] в словаре Ожегова [12]. Формируется выдача.

Использование веб-стека переносит нагрузку на сторону сервера, что уменьшает требовательность приложения к ресурсам аппаратной платформы. Также, исчезает проблема обновлений программы, так как используется всегда свежая версия.

4. АЛГОРИТМ МОРФЕМНОГО АНАЛИЗА

Морфемный анализ не так популярен среди исследователей, так как имеет небольшую практическую применимость, поэтому было принято решение разработать собственный алгоритм

получения структуры слова. Подробно алгоритм описан в материалах конференции «МСИТ 2016» [13].

Алгоритм построен на последовательном переборе уже известных морфем в данном слове. Для этого был использован оцифрованный Морфемно-орфографическии словарь русского языка А. Н. Тихонова [12], в котором размечен большой объем слов, но сами типы морфем не обозначены. Второй словарь — морфемы в формате JSON из словаря Т. Ф. Ефремовой [14], и данных сайта «Словород» [15]. Схема алгоритма представлена на Рис. 3.

Рис. 3. Упрощенная схема действия алгоритма морфемного анализа

Записи в словаре Тихонова выглядят как слово, разделенного слешами. Например, для слова частнопредпринимательский — **частн/о/предприним/а'/тель/ск/ий** система производит поиск по словарю и получает строчку такого вида. Далее алгоритм, используя преобразование строчки в массив с помощью функции explode получает массив морфем, но каких именно, неизвестно. С помощью данных морфемного словаря и последовательной проверки каждого элемента массива получается результирующий массив, который в дальнейшем используется API-методом для выдачи другим контроллерам системы.

Представленный алгоритм обладает ограничением: достоверно определить корень не всегда получается, так как полного словаря корней не существует. На данный момент существует и обновляется корнеслов Федора Шимкевича [15] на сайте «Словород», но он не является полным. В алгоритме же проблема обходится путем отметки начала и конца предполагаемого корневого массива (Рис. 4).

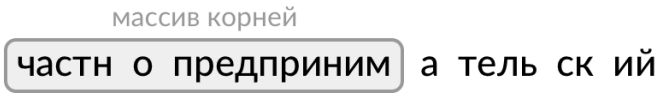

Рис. 4. Выделенный массив корней

Данный метод не работает в случаях, когда слово сложное (т. е. имеет несколько корней) и написано через дефис. В данном случае целесообразнее предварительно разбивать слово на несколько частей, а затем разбирать эти части.

5. ГРАФИЧЕСКИЙ ИНТЕРФЕЙС

Тренды в дизайне интерфейсов уже давно перешли от скевоморфизма к простым формам и так называемому «плоскому» дизайну. Приложения избавляются от лишних деталей, чтобы перенести информацию во главу угла, тем самым повышая получаемую пользу.

Во время проектирования интерфейса данного сервиса применен подход «mobile-first», что позволило улучшить пользовательский опыт на мобильных устройствах. Это подход, при котором проектирование сервиса начинается с мобильной версии, а не с версии для больших экранов, как это делается повсеместно. Подробнее об этом подходе написано в книге [16].

Дизайн выдачи результатов вдохновлен сервисом Wolfram Alpha, который акцентирует внимание на полезной информации. Интерфейс сервиса в двух разрешениях показан на Рис. 5.

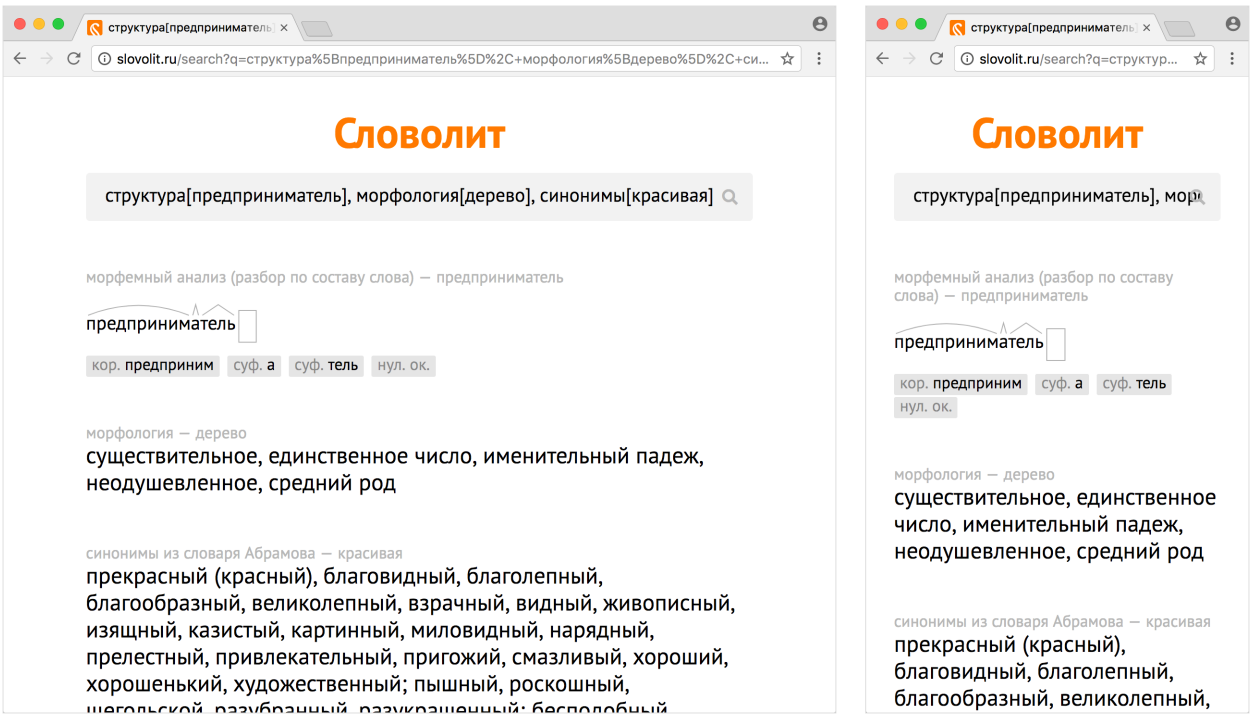

Рис. 5. Интерфейс веб-сервиса: в настольном и мобильном браузере

6. НЕРЕШЕННЫЕ ПРОБЛЕМЫ

Представленный в работе алгоритм морфемного разбора имеют существенное ограничение, связанное с отсутствием слова в словаре Тихонова. Так как язык постоянно расширяется, трудоемко иметь свежий словарь, покрытие разбираемых слов будет постепенно уменьшаться.

Другая проблема состоит в том, что алгоритм не может с максимальной точностью определить корень, потому что словаря корней также не существует, что следует из постоянного расширения языка.

Эти проблемы можно решить разработкой платформы, которая будет собирать данные о словах из различных источников, например, блогов, новостных сервисов или открытых корпусов. Путем частотного анализа и частичного разбора этих слов можно выявлять новые корни и пополнять словарь.

ЗАКЛЮЧЕНИЕ

Разработанный веб-сервис позволяет производить различные анализы слов, а также комбинировать результаты посредством использования различных команд в строке запросов. Сервис реализован в виде адаптивного веб-сайта, что позволяет использовать его как на настольных, так и на мобильных устройствах.

Разработанный алгоритм морфемного анализа может помочь другим специалистам при разработке текстовых анализаторов, например, для выявления суффиксов.

Представленный в данной статье веб-сервис, в отличие от других сервисов в сфере автоматической обработки слов, позволяет разбирать несколько слов одновременно, а также комбинировать результаты работы сервиса для разных слов на одной странице браузера.

В будущем планируется добавить алгоритм «предсказания» слов, которых нет в словаре, а также добавить в сервис фонетический и другие разборы. Кроме того, планируется разработка открытого АПИ для сторонних разработчиков для удобного пользования сервисом.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. *Ножов И. М.* Реализация автоматической синтаксической сегментации русского предложения [Текст] : дис. … канд. тех. наук : 05.25.05 / Ножов Игорь Михайлович. — М. : Российский государственный гуманитарный университет, 2003. — 147 с.
2. *Segalovich I.* A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine //. — 2003. — С. 273—280. — (MLMTA).
3. MyStem — Технологии Яндекса. — URL: `https://tech.yandex.ru/mystem/` (дата обр. 13.07.2017).
4. phpMorphy. — URL: `http://phpmorphy.sourceforge.net/dokuwiki/` (дата обр. 13.07.2017).
5. A Guide to Becoming a Full-Stack Developer in 2017. — URL: `https://medium.com/coderbyte/a-guide-to-becoming-a-full-stack-developer-in-2017-5c3c08a1600c` (дата обр. 13.07.2017).
6. Compare CSS PREprocessors. — URL: `http://csspre.com/compare/` (дата обр. 13.07.2017).
7. CSS preprocessors. — URL: `https://github.com/showcases/css-preprocessors` (дата обр. 13.07.2017).
8. Sass: Syntactically Awesome Style Sheets. — URL: `http://sass-lang.com/` (дата обр. 13.07.2017).

9. Why should I learn to program in PHP, why is it so important? — URL: https://www.quora.com/Why-should-I-learn-to-program-in-PHP-why-is-it-so-important (дата обр. 13.07.2017).

10. Is PHP dead? — URL: https://www.quora.com/Is-PHP-dead (дата обр. 13.07.2017).

11. *Otwell T.* Laravel - The PHP framework for web artisans. — URL: https://laravel.com/ (дата обр. 13.07.2017).

12. Словари русского языка для скачивания. — URL: http://www.speakrus.ru/dict/index.htm (дата обр. 13.07.2017).

13. *Правосудов М. М.* Программный алгоритм морфемного анализа слов русского языка : Сборник трудов XIV Международной научно-практической конференции студентов, аспирантов и молодых ученых //. Т. 2. — Изд-во ТПУ. 2016. — С. 276—277. — (Молодежь и современные информационные технологии).

14. Указатель морфов. — URL: http://rusgram.narod.ru/morf1t.html (дата обр. 13.07.2017).

15. *Гаршин И. К.* Словород: образование и история слов русского языка. Собирание и оживление славянских корней. — URL: http://www.slovorod.ru (дата обр. 13.07.2017).

16. *Wroblewski L.* Mobile First. — A book apart, 10.2011. — ISBN 978-1-937557-02-7.