ИНФОРМАТИКА

УДК 004.89:004.912 DOI: https://doi.org/10.54341/20778708_2022_4_53_84
EDN: XJBRGC

ВОЗМОЖНОСТИ И СРЕДСТВА БИБЛИОТЕКИ NEVOD ПРИ РЕШЕНИИ ЗАДАЧ ИЗВЛЕЧЕНИЯ ВРЕМЕННЫХ УКАЗАТЕЛЕЙ В ТЕКСТЕ

В.А. Савёнок[1], В.Б. Таранчук[2]

[1]*Белорусский государственный университет информатики и радиоэлектроники*, *Минск*
[2]*Белорусский государственный университет*, *Минск*

FEATURES AND TOOLS OF THE NEVOD LIBRARY IN SOLVING PROBLEMS OF EXTRACTING TEMPORAL MARKERS IN THE TEXT

V.A. Savionok[1], V.B. Taranchuk[2]

[1]*Belarusian State University of Informatics and Radioelectronics*, *Minsk*
[2]*Belarusian State University*, *Minsk*

Аннотация. Рассматриваются теоретические и методические вопросы семантического анализа текста в части извлечения фактов. На примере решения задачи извлечения временных указателей представлен способ поиска в тексте и его реализация в библиотеке Nevod. Анализируется функциональная полнота разработанной библиотеки путем сопоставления ее возможностей с инструментарием одного из лидеров в области распознавания сущностей – Microsoft.Recognizers.Text.

Ключевые слова: *семантический анализ текста*, *автоматическая обработка текста*, *тестовые наборы данных*, *поиск текста по шаблонам*, *пакет шаблонов*, *распознавание сущностей*, *временные указатели*, *библиотека Nevod*, *система компьютерной алгебры Mathematica.*

Для цитирования: *Савёнок*, *В.А*. Возможности и средства библиотеки Nevod при решении задач извлечения временных указателей в тексте / В.А. Савёнок, В.Б. Таранчук // Проблемы физики, математики и техники. – 2022. – № 4 (53). – С. 84–92. – DOI: https://doi.org/10.54341/20778708_2022_4_53_84. – EDN: XJBRGC

Abstract. Theoretical and methodological issues of semantic text analysis in terms of extracting facts are considered. Using the example of solving the problem of extracting temporal markers, the text search method, and its implementation in the Nevod library are presented. The functional completeness of the developed library is analyzed by comparing its capabilities with the tools of one of the leaders in the field of entity recognition – Microsoft.Recognizers.Text.

Keywords: *semantic text analysis*, *automatic text processing*, *test data sets*, *pattern-based text search*, *pattern package*, *entity recognition*, *temporal markers*, *Nevod library*, *Mathematica computer algebra system.*

For citation: *Savionok*, *V.A*. Features and tools of the Nevod library in solving problems of extracting temporal markers in the text / V.A. Savionok, V.B. Taranchuk // Problems of Physics, Mathematics and Technics. – 2022. – № 4 (53). – P. 84–92. – DOI: https://doi.org/10.54341/20778708_2022_4_53_84 (in Russian). – EDN: XJBRGC

Введение

С развитием компьютерных технологий и постоянным приростом объемов текстовой информации исследования в области автоматической обработки текстов сфокусировались на прикладном аспекте вопроса. Возможности большинства инструментов ориентированы на морфологический и синтаксический анализ с применением методов теории вероятностей и прикладной статистики.

Одним из главных направлений в области обработки текстов является выделение их смысловой составляющей – семантический анализ. В этом направлении решаются такие задачи, как поиск документов в локальных и глобальных сетях [1], автоматическое аннотирование и реферирование [2], классификация и кластеризация документов [3], синтез текстов и машинный перевод [4], [5], извлечение фактов [6], анализ тональности текста [7]. Результаты исследований применяются повсеместно, как в узкоспециализированных решениях и подсистемах, включаемых в экспертные системы [8], так и в масштабных системах поиска информации в глобальной сети. Можно констатировать, что, хотя научные и технические идеи в области обработки текстов развиваются интенсивно, некоторые задачи семантического анализа остаются нерешенными [9], [10].

Семантический анализ текста работает с представлением знаний в языковой форме. В основу компьютерной обработки ставится понимание, что слова в языке (тексте) отражают отдельные сущности, а грамматические отношения между словами передают их смысловую связь. Это обеспечивает преобразование текстового представления знаний в более формализованную форму – в виде специального словаря – онтологии. Словари такого типа должны оперировать

смыслами, не их частными обозначениями в виде слов, и, следовательно, описывать свойства и отношения понятий, а не слов [11], [12]. Процедуры семантического анализа опираются на функциональность используемого словаря. Эффективность словарной поддержки в системах семантического анализа является ключевым аспектом. При проектировании таких систем возникает вопрос, как правильно структурировать и представлять информацию в подобных словарях, чтобы поиск по ним был удобным и быстрым, к тому же, была возможность учитывать изменения в естественном языке – исчезновение старых и возникновение новых понятий, терминов, аббревиатур [13].

Онтологии применяются, в частности, на одном из ключевых этапов процесса семантического анализа – извлечении объектов и фактов. Данный этап позволяет установить взаимосвязи между участниками-фигурантами фактов для формирования целостной смысловой картины текста.

Наряду с онтологиями применяются и другие методы извлечения фактов: методы, основанные на машинном обучении, и методы, основанные на правилах [14]. Методы извлечения фактов на основе машинного обучения используют автоматическое извлечение признаков из текста и применяют такие алгоритмы машинного обучения, как, например, классический байесовский классификатор, дерево решений, или метод опорных векторов [15]. Методы, основанные на правилах, часто применяются для установления взаимосвязей между объектами – семантических отношений, которые носят различный характер: иерархия, агрегация, функциональные, семиотические отношения, отношения тождества, корреляции. Выделение отношений в рамках каждой из перечисленных групп требует составления соответствующих наборов предикатов различной формы [16].

Неотъемлемой составляющей задачи извлечения фактов и определения отношений между объектами является локализация во времени события, соответствующего факту. Информация, позволяющая локализовать событие на временной оси, передается посредством разнообразных по форме и содержанию текстовых выражений – временных указателей. Конечным результатом извлечения временных указателей из текста является их представление и интерпретация в рамках заданной в процессе семантического анализа формальной модели [17].

Одним из первых этапов извлечения фактов является выделение фрагментов текста, представляющих собой языковое отражение искомых фактов, что обуславливает необходимость решения задачи поиска в тексте конкретных задаваемых конструкций из слов и словосочетаний. Для этого применяется ряд методов и инструментов, в том числе средств лексического анализа, полнотекстового и морфологического поиска, которые оперируют на уровне символов и слов, а также содержат средства учета морфологии, словоформ [18]. В попытке воспользоваться некоторой формализованностью, присущей естественно-языковым конструкциям, указанные средства нередко основывают на грамматиках. Согласно иерархии Хомского, используют грамматики регулярного уровня и контекстно-свободные грамматики [19]. Примером средства извлечения сущностей, основанного на регулярных грамматиках, является [20]. Учет морфологических характеристик осуществляется в таких средствах, как [21] и [22]. Как показывают исследования, использование регулярных и контекстно-свободных грамматик при работе с естественными языками не позволяет охватить все многообразие языковых конструкций [23]. Поддержка морфологических аспектов для нескольких языков, наряду с работой на низком уровне символов, усложняет процесс описания языковых конструкций для поиска фрагментов текста, что в целом понижает производительность обработки текстов на естественных языках.

Резюмируя упомянутые аспекты текущего состояния в частях теории и методов, отметим, что создание новых методов семантического анализа текстов откроет новые возможности и позволит существенно продвинуться в решении многих задач компьютерной лингвистики, в частности, таких как машинный перевод, автореферирование, классификация текстов и других. Важной частью при этом является разработка новых инструментов, позволяющих автоматизировать семантический анализ. Представляется целесообразным разработать и реализовать средства текстового поиска, оперирующие на уровне слов и их сочетаний с минимальным учетом морфологических свойств. Ориентируясь на большие объемы данных, следует также оптимизировать способ работы с текстом, например, путем реализации потоковой обработки за один последовательный просмотр текста. Такой метод текстового поиска разработан в [24] и реализован в библиотеке Nevod [25]. Для оценки качества новых инструментов, предназначенных для извлечения фактов, и подтверждения их функциональной полноты обязательным является тестирование на эталонных наборах. Такие наборы следует готовить на базе типовых, но с дополнениями, отвечающими добавляемой функциональности. Первостепенной задачей при формировании эталонных наборов является определение их содержания, включения текстовых фрагментов, подтверждающих достоверность получаемых результатов извлечения фактов.

Анализ предложений и ИТ решений показывает наличие достаточно большого спектра систем, программных пакетов и приложений

семантического анализа текстов на английском, французском, немецком, испанском, итальянском, китайском и некоторых других языках. В то же время, можно констатировать, что у лидеров мирового уровня отсутствуют тестовые наборы для анализа текстов, в частности, на русском, белорусском языках [20], [26], [27]. Поэтому описываемые в данной работе методические рекомендации по подготовке наборов данных составлены без привязки к обрабатываемому языку и имеют общий характер.

Целью настоящей публикации является изложение:

– методических основ, особенностей использования разработанной библиотеки Nevod;

– вариантов настройки, возможного расширения библиотеки Nevod по результатам эксплуатации и сопоставления с аналогами, обоснование ее функциональной полноты с использованием специально сформированных наборов;

– подходов, методических и алгоритмических решений, средств подготовки представительных наборов данных для тестирования программных инструментов семантического анализа применительно к задачам извлечения фрагментов текста с временными указателями.

1 Общие сведения о библиотеке Nevod

Библиотека Nevod предназначена для поиска параметризированных шаблонов в тексте, которые представляют собой наборы правил с описаниями лексем, их сочетаний в виде последовательностей, альтернатив (вариаций), повторений и контекстных конструкций. Запатентованная технология позволяет проводить семантический анализ текста путем организации поиска сущностей и взаимосвязей между ними при использовании соответствующих наборов шаблонов [24].

Состав библиотеки Nevod. Библиотека включает компоненты поиска и редактор набора шаблонов, составные части показаны на рисунке 1.1.

Компоненты поиска обеспечивают ключевую функциональность библиотеки, к ним относятся:

– *Синтаксический анализатор языка описания шаблонов* – строит на основании текста описания шаблонов соответствующее дерево выражений и запрашивает загрузку требуемых наборов шаблонов.

– *Загрузчик набора шаблонов* – отвечает за поиск и загрузку текста описания наборов шаблонов, указанных в качестве зависимостей для пользовательского набора.

– *Компонент связывания наборов шаблонов* – устанавливает гиперсвязи между деревьями выражений, которые соответствуют наборам шаблонов, и проверяет доступность используемых ссылок.

– *Лексический анализатор текста для поиска* – осуществляет разбиение входного текста на последовательность лексем для анализа.

– *Движок поиска* – производит сопоставление последовательности лексем исходного текста для поиска со связанными деревьями выражений искомых шаблонов. Предварительно на основании деревьев выражений шаблонов составляется *поисковый индекс*, который используется при работе *автомата поиска*.

Редактор набора шаблонов реализован в виде универсального языкового сервера, поддерживающего разработанный Microsoft протокол LSP (Language Server Protocol) [28]. Использование данного протокола позволяет встраивать модуль в большинство современных сред разработки с минимальными затратами путем написания соответствующих расширений для взаимодействия с языковым сервером, например [29].

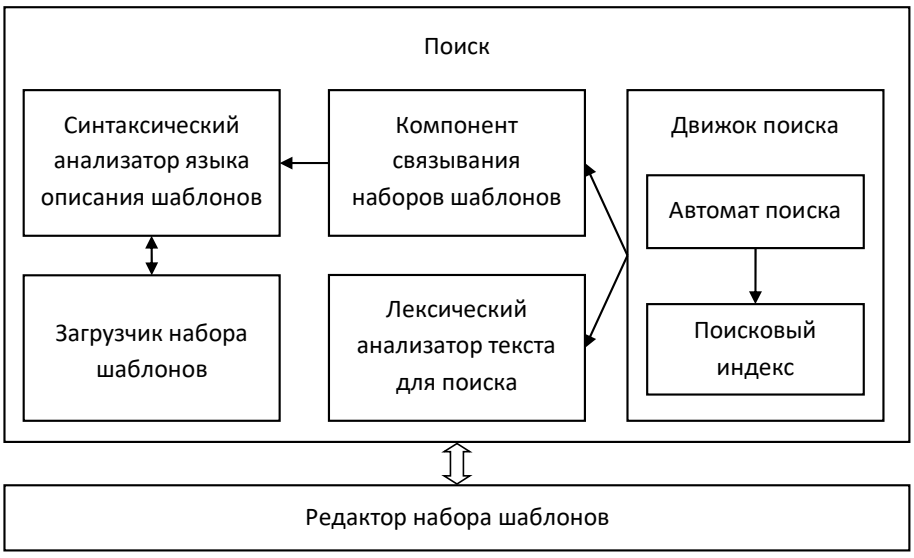

Рисунок 1.1 – Библиотека Nevod

При редактировании обеспечивается составление связанных наборов шаблонов с использованием базовых, контекстных и специальных операторов. К базовым операторам относятся следующие:

– последовательность – строгое следование подряд заданных выражений;

– вариация – возможность совпадения заданных выражений и невозможность совпадения других указанных выражений;

– повторение – последовательное совпадение выражения несколько раз подряд;

– опциональность – необязательное совпадение выражения;

– последовательность слов – совпадение двух выражений, разделенных лексемами «разрыв слова»;

– промежуток в словах – последовательное совпадение двух выражений, разделенных заданным количеством слов;

– сочетание – совпадение двух выражений в любом порядке, разделенных любым количеством лексем.

Контекстные операторы представлены следующими конструкциями:

– находится внутри – совпадение одного выражения в любой позиции по тексту совпадения другого выражения;

– не пересекает – совпадение выражения, которое не пересекает совпадения другого заданного выражения;

– содержит – совпадение выражения, текст которого содержит совпадения другого заданного выражения в любой позиции.

Оператор извлечения части совпадения является единственным представителем специальных операторов; он позволяет выделить в дополнительный атрибут результата часть текста совпадения, которая соответствует обозначенной части выражения.

Помимо операторов, в синтаксисе языка описания шаблонов определены специальные директивы, управляющие связыванием и компиляцией шаблонов (построением поискового индекса):

– @require – установление гиперсвязи с внешним набором шаблонов (связывание);

– @namespace – объявление пространства имён шаблонов (связывание);

– @pattern – обозначение определения шаблона, по умолчанию не является целью поиска (связывание и компиляция);

– @search – обозначение цели поиска в наборе шаблонов (компиляция);

– @where – определение секции вложенных шаблонов (связывание и компиляция).

В комплект библиотеки Nevod включены базовые (по умолчанию) наборы (комплекты) шаблонов, которые разработаны для поиска различных сущностей в тексте, таких как даты и время, временные интервалы, электронные адреса, телефонные номера, мировые денежные валюты и др. [30]. В данной работе сделан акцент на наборе шаблонов для поиска и распознавания в тексте абсолютных и относительных дат, которые могут выступать в качестве временных указателей.

2 О функциональной полноте библиотеки Nevod

Извлечение временных указателей средствами библиотеки Nevod предполагает проведение сопоставления возможностей с аналогичной библиотекой для распознавания сущностей в тексте Microsoft.Recognizers.Text (далее MRT) [20]. Библиотека MRT предоставляет возможность распознавания сущностей в текстах различных языков и широко используется в продуктах Microsoft, например, в предустановленных шаблонах для сервиса интеллектуального понимания языка LUIS (Language Understanding Intelligent Service), в платформе для создания диалоговых ботов Power Virtual Agents [31] и в когнитивных языковых сервисах в облачной инфраструктуре Azure – NER (Named Entity Recognition). Библиотека распространяется под лицензией открытого и свободного программного обеспечения MIT; наряду с исходным кодом в репозитории Microsoft на GitHub [20] публично доступны контрольные наборы данных для различных языков. Отметим, что в MRT отсутствует поддержка русского и белорусского языка.

Для поиска временных меток в тексте в MRT применяется модуль Microsoft.Recognizers.Text.DateTime, в частности, его компонент BaseDateExtractor. Данному компоненту соответствует контрольный набор, представленный в формате JSON, – файл DateExtractor.json [32]. Набор содержит 143 элемента, включающих абсолютные и относительные даты в различных вариантах записи, а также метаинформацию, которая используется для проверки корректности результатов извлечения. К элементу набора может быть прикреплен поисковый контекст – опорная дата, которая указывает точку во времени, используемую для перевода относительных временных меток в абсолютные. Пример элемента контрольного набора с комментариями по структуре изображен на рисунке 2.1.

Контрольный набор DateExtractor используется для подтверждения функциональной полноты библиотеки Nevod. Для сопоставления возможностей библиотек MRT и Nevod разработаны два программных модуля – mMRT и mNevod, обеспечивающие поиск и извлечение временных меток из текста. С использованием средств системы компьютерной алгебры Wolfram Mathematica проведено сравнительное тестирование программных модулей на упомянутом контрольном наборе. Получены следующие результаты: mMRT обработал корректно все элементы набора,

```
{
 "Input": "Cortana, please set up a Skype call sometime this friday
june twenty three with Jim", // входной текст для поиска
 // поисковый контекст:
 "Context": {"ReferenceDateTime":"2018-06-20T00:00:00"},
 "Results": [ // перечисление ожидаемых результатов
  // содержит текст совпадения, тип, позицию начала и длину совпадения
  {"Text":"this friday", "Type":"date", "Start":45, "Length":11},
  {"Text":"june twenty three", "Type":"date", "Start":57, "Length":17}
 ]
}
```

Рисунок 2.1 – Пример элемента контрольного набора

показав точность 100%; mNevod с базовым набором шаблонов [30] произвел корректное извлечение временных меток в 82,5% случаев.

Полученные результаты предполагают анализ содержимого контрольного набора. Для реализации такого анализа использовались средства системы Mathematica. Определены 17,5% нераспознанных mNevod ситуаций, они разделены на три категории:

1. Месяц и порядковый день месяца в словесной форме – «a baseball on may the eleventh», «i'll go back fourth of may» и т. п.

2. День недели в будущем относительно текущей даты – «Are you available two monday later?» и т. п.

3. Порядковый день недели в месяце – «i'll go back second sunday» и т. п.

Благодаря поддержке расширяемости набора правил в библиотеке Nevod, в базовый набор правил включены дополнительные шаблоны, которые позволили корректно обработать ранее не распознанные ситуации.

Для распознавания ситуаций категории 1 разработан дополнительный шаблон MonthWithOrdinalDay, который можно описать в виде нескольких альтернативных цепочек:

а) опциональный день недели, название месяца и номер дня (числовой или словесный), опциональный год;

б) опциональный день недели, номер дня (числовой или словесный) и название месяца, опциональный год;

в) номер дня (числовой или словесный) и относительный месяц («предыдущий», «следующий», «текущий» и т. п.).

С целью сокращения длины шаблона и снижения дублирования, из альтернатив а) и б) вынесен общий префикс (опциональный день недели) и суффикс (опциональный год), что позволило объединить обе ветки в цепочку с вариацией на второй позиции. Также при написании шаблона использовались уже содержащиеся в базовом наборе шаблонов элементы «день недели», «год», «относительный месяц» и др., что позволило сократить время составления нового правила и сохранить достаточный уровень наглядности полученного шаблона.

После добавления нового элемента в базовый набор шаблонов библиотеки Nevod, точность решения на его основе для контрольного набора MRT повысилась на 12,1% – до 94,4%.

Для обеспечения распознавания ситуаций категории 2 составлен шаблон OrdinalWeekDay, представляющий из себя простую цепочку из порядкового номера дня недели в месяце в словесной форме (от одного до пяти: «first», «second», … «fifth») и названия дня недели. Для учета различных вариантов сокращенного написания дня недели использован готовый шаблон из базового набора в библиотеке Nevod. Добавление нового шаблона повысило корректность работы модуля mNevod на 2,8% по сравнению с предыдущим результатом – до 97,2%.

Ситуации категории 3 устранены шаблоном RelativeWeekDay, который также представлен цепочкой из словесной записи числа, задающего количество пропускаемых недель, и названия дня недели, а также суффикса (вариации «from now» или «later»), указывающего на относительность временной метки. Данный шаблон обеспечивает распознание еще 1,4%, повысив общий результат до 98,6% правильно распознанных ситуаций из контрольного набора MRT.

При помощи средств Mathematica проведен анализ итоговых результатов. Обнаружено, что 1,4% ситуаций, которые не были распознаны после расширения базового набора правил библиотеки Nevod, представлен следующими элементами:

1) «Cortana, please set up a Skype call sometime this friday june twenty two with Jim»;

2) «6,107.31 August 2019 should not include the decimal».

В первом случае при использовании mNevod распознаны две временные метки: «this friday» и «june twenty two», в то время как в контрольном наборе MRT определено, что извлеченные выражения необходимо объединить. Библиотека MRT объединяет несколько идущих подряд текстовых дат в один результат извлечения на основании контекста – опорной даты для разрешения относительных временных меток. Программный модуль на основе библиотеки Nevod не производит объединения результатов извлечения, однако описанную ситуацию можно

предусмотреть путем добавления отдельного этапа условного слияния последовательно извлеченных временных меток. Следует отметить, что в некоторых случаях, например, при отсутствии какой-либо информации об опорной дате, объединенное выражение может представлять собой некорректную дату. Описанная ситуация учтена в контрольном наборе MRT и представлена другим элементом с такой же опорной датой: «Cortana, please set up a Skype call sometime this friday june twenty three with Jim». Здесь и mNevod, и mMRT извлекли две временные метки: «this friday» и «june twenty three».

Во втором случае при детальном рассмотрении тестового набора MRT выяснилось, что из фрагмента «6,107.31 August 2019» предполагается извлечение подстроки «August 2019», т. к. символ точка считается десятичным разделителем. Результирующая строка относится к *временным интервалам*, за обработку которых в библиотеке MRT отвечает отдельный модуль. Таким образом, при извлечении *временных меток* ожидается пустой результат. В базовом наборе правил библиотеки Nevod символ точка определен в качестве признака конца предложения, а десятичным разделителем выступает символ запятая. Данная формулировка привела к получению результата «31 August 2019». Путем точечного изменения существующих правил извлечения числовых дат достигнут ожидаемый результат – извлечен временной интервал «August 2019».

Заметим, что шаблоны, составленные для каждой категории ситуаций, первоначально не распознанных mNevod, являются независимыми: добавление или изъятие одного из шаблонов не влияет на работу других. Так, например, при изъятии из итогового набора правил шаблона, соответствующего ситуациям категории 1, и сохранении шаблонов для ситуаций категорий 2 и 3, точность работы модуля снизилась на 12,1% – на столько же процентов она повысилась при добавлении шаблона в исходный базовый набор правил. Написание независимых правил позволяет составлять наборы шаблонов, обладающие свойством аддитивности, что упрощает процесс их корректировки и совместного использования.

Таким образом, путем сравнительного анализа результатов обработки типового контрольного набора из библиотеки MRT подтверждена функциональная полнота библиотеки Nevod. В ходе проверки показаны дополнительные возможности библиотеки, в частности расширяемость набора правил.

Расширяемость набора шаблонов в библиотеке Nevod позволяет настраивать основанные на ней решения во время эксплуатации без необходимости взаимодействия с разработчиками библиотеки, что обеспечивает независимость решений и сокращает цикл обновления набора правил.

3 Методика формирования представительного тестового набора

При проверке и настройке средств извлечения фактов, в частности, временных указателей, одной из позиций для оценивания является ориентированность на распознавание, а не однозначную идентификацию сущностей в тексте. Исходный контрольный набор DateExtractor библиотеки MRT не позволяет в полной мере проанализировать функциональность соответствующих средств такого типа – он охватывает большинство вариантов написания дат в английском языке, включает аббревиатуры и общепринятые сокращения, но не учитывает возможность искажения входного текста. Представляется целесообразным составить новый контрольный набор данных, который бы учитывал данный аспект при оценке инструментов извлечения фактов.

Ориентируясь на инструменты для извлечения временных интервалов в тексте, используя фрагменты из DateExtractor, подготовлен новый тестовый набор. Для корректного сравнения (и с MRT в дальнейшем) из DateExtractor исключены элементы, составляющие упомянутые выше 1,4% нераспознанных библиотекой Nevod ситуаций. В полученный набор из 141 элемента внесены искажения (ошибки), наиболее типичные для ручного набора текста, таким образом, чтобы они затрагивали фрагменты текста, представляющие собой цель для извлечения.

В качестве типичных ошибок ручного ввода, которые не затрагивают итоговую длину слова, выбраны следующие виды ситуаций:

1) замена одиночной буквы,

2) перестановка пары соседних букв в слове.

Следует отметить, что при моделировании искажений вида 1) существует естественная эвристика, позволяющая ограничить множество букв, которые могут быть употреблены некорректно вместо заданной корректной буквы. В основе лежит предположение об использовании стандартного средства ввода текстовых данных для ЭВМ – клавиатуры. В таком случае наиболее часто встречающимися заменами будут являться соседние буквы по расположению клавиш на клавиатуре. Например, для слова «Monday» одним из частых вариантов такой ошибки для раскладки QWERTY является замена буквы «d» на букву «s» – «Monsay».

При моделировании ошибок вида 2) также возможно применение подобной естественной эвристики. Принимая во внимание слепой метод печати [33], логично предположить, что чаще всего ошибка с перестановкой будет возникать для символов, за набор которых отвечают пальцы разных рук. Схема зон ответственности разных пальцев при слепом методе печати изображена на рисунке 3.1. К примеру, для слова «Sunday» вариантом такого искажения в

раскладке QWERTY является перестановка букв «а» и «у» – «Sundya».

Рисунок 3.1 – Схема зон ответственности разных пальцев при слепом методе печати для американской раскладки QWERTY

Перечисленные виды искажений входного текста могут носить множественный характер: в одном слове может присутствовать как несколько ошибок одного вида, так и одновременно ошибки нескольких видов. При составлении контрольных наборов необходимо учитывать различные вариации ошибок. Далее показан процесс моделирования каждого вида искажений отдельно, без учета их комбинаций.

В соответствии с каждым видом ошибок, в полученный набор из 141 элемента внесены следующие искажения:

– произведена замена буквы «d» на букву «s» в слове «monday» (соответствует 2,8% набора, всего слово содержится в 3,5% набора);

– произведена перестановка букв «а» и «у» в слове «sunday» (соответствует 5,7% набора, всего слово содержится в 7% набора).

При оценке корректности обработки полученного набора программными модулями mNevod и mMRT, получены идентичные результаты: 91,4%. Благодаря расширяемости шаблонов в пакете Nevod добавлены правила для нивелирования соответствующих ошибочных ситуаций. Для распознавания замены буквы внесены различные варианты искажения слова «monday» на позиции, соответствующей букве «d». Для учета перестановки соседних букв в слове «sunday» добавлено правило, покрывающее возможные перестановки: с учетом описанной ранее эвристики, это варианты «sundya» и «usnday». Программный модуль, использующий библиотеку Nevod с обновленным набором правил, обработал корректно 100% представительного набора.

Таким образом, использование средств библиотеки Nevod позволяет адаптировать программный модуль под различные варианты искажений входных данных путем внесения точечных (локальных) изменений в существующие наборы шаблонов. Применение предложенных эвристик при составлении новых наборов правил позволит реализовать первоначальную обработку типовых ошибок ввода.

4 Используемые средства Wolfram Mathematica

Для сопоставления функциональных возможностей модулей mNevod и mMRT при решении задачи выделения временных указателей в тексте не только на основе контрольного набора DateExtractor, но и формирования других представительных наборов, в Wolfram Mathematica разработано сервисное приложение mDataWM. В нем реализованы программные инструменты, которые позволяют выделить подлежащий обработке набор данных от метаинформации, оценить и сопоставить качество результатов обработки модифицированного набора модулями mMRT и mNevod, а также исказить любой набор данных и проверить работоспособность библиотек. Приложение mDataWM обеспечивает создание тестовых наборов на любых языках и анализ результатов их обработки. Средства приложения mDataWM включают инструменты:

– искажения начального и формирование модифицированного набора данных;

– импорта / экспорта для обеспечения взаимодействия Mathematica с модулями mMRT и mNevod (работа с файлами и отделение данных от метаинформации);

– оценки качества результатов обработки тестового набора.

В приложении mDataWM использованы следующие функции ядра системы Mathematica [34]:

– Import[source],
– Export["dest.ext",expr,"format"],
– Map[f,expr],
– MapIndexed[f,expr],
– Association[key1→val1,key2→val2,...],
– AssociateTo[a, key→val],
– SortBy[list,f],
– KeyMemberQ[assoc,form],
– KeyDrop[assoc,{key1, key2,...}],
– KeyTake[assoc,{key1,key2,...}],
– RandomSample[{e_1,e_2,...},n],
– Select[list, crit],
– Delete[expr,n],
– StringReplace["string",s→sp],
– Count[list,pattern].

Заключение

Отмечены требования к новым средствам семантического анализа применительно к задаче извлечения временных указателей в тексте; представлены способы и инструменты поиска в тексте, их реализация в библиотеке Nevod. При подтверждении функциональной полноты библиотеки показано основное преимущество – расширяемость набора правил. Предложены несколько естественных эвристик, которые позволяют оптимизировать процесс внесения искажений в представительный набор для тестирования, анализа результатов извлечения фактов. Соответствующие инструменты подготовки специальных наборов данных для тестов разработаны с использованием средств системы компьютерной алгебры Wolfram Mathematica. Приведено описание методов и средств создания подобных

наборов, поясняется их содержание и особенности наполнения, когда наряду со значимыми и ключевыми словами, аббревиатурами, фразами включаются их искажения. Показано, что средствами библиотеки Nevod обеспечено распознавание, а не только однозначная идентификация сущностей в тексте.

ЛИТЕРАТУРА

1. *Половикова*, *О.Н.* Анализ способов формализаций документов для выполнения семантического поиска / О.Н. Половикова // Известия Алтайского государственного университета. – 2012. – № 1–2 (73). – С. 101–103.

2. *Батура*, *Т.В.* Методы и системы автоматического реферирования текстов: монография / Т.В. Батура, А.М. Бакиева; Ин-т систем информатики им. А.П. Ершова СО РАН. – Новосибирск: ИПЦ НГУ, 2019. – 110 с.

3. *Барахнин*, *В.Б.* Кластеризация текстовых документов на основе составных ключевых термов / В.Б. Барахнин, Д.А. Ткачев // Вестник Новосибирского государственного университета. Серия: Информационные технологии. – 2010. – Т. 8. – № 2. – С. 5–14.

4. *Липницкий*, *С.Ф.* Математическая модель синтеза текстов на основе слияния коммуникативных фрагментов / С.Ф. Липницкий // Проблемы физики, математики и техники. – 2018. – № 4 (37). – С. 106–110.

5. *Митренина*, *О.В.* Машинный перевод / О.В. Митренина // Прикладная и компьютерная лингвистика / И.С. Николаев, О.В. Митренина; под ред. Т.М. Ландо – М.: URSS, 2017. – Ч.2 – Гл. 1 – С. 156–189.

6. *Богатырев*, *М.Ю.* Извлечение фактов из текстов естественного языка с применением концептуальных графовых моделей / М.Ю. Богатырев // Известия Тульского государственного университета. Технические науки. – 2016. – № 7–1. – С. 198–208.

7. *Семина*, *Т.А.* Анализ тональности текста: современные подходы и существующие проблемы / Т.А. Семина // Социальные и гуманитарные науки. Отечественная и зарубежная литература. Серия 6: Языкознание. Реферативный журнал. – 2020. – № 4. – С. 47–63.

8. *Джарратано*, *Дж.* Экспертные системы. Принципы разработки и программирование / Дж. Джарратано, Г. Райли. – 4-е изд. – М.: Вильямс, 2007. – 1152 с.

9. *Сачков*, *В.Е.* Анализ проблемы частотного перекрытия слов при определении тематики текста в семантических вычислительных комплексах / В.Е. Сачков // Вестник современных исследований. – 2018. – № 6.1 (21). – С. 486–488.

10. *Листратова*, *О.К.* К проблеме анализа единиц решения при восприятии формы слова и его семантической интерпретации / О.К. Листратова // Вопросы современной филологии в контексте взаимодействия языков и культур: Материалы IV Международной научно-практической конференции, Оренбург, 26–27 мая 2021 года / Отв. за выпуск Е.А. Стуколова. – Оренбург: Оренбургский государственный педагогический университет, 2021. – С. 91–95.

11. *Система и способ создания и использования пользовательских семантических словарей для обработки пользовательского текста на естественном языке*: пат. 2 584 457 Российская Федерация, МПК G06F 17/28 / Е.Н. Яковлев, А.С. Старостин; заявитель общество с ограниченной ответственностью «Аби ИнфоПоиск» – № 2015103467/08; заявл. 03.02.15; опубл. 20.05.16 // Официальный бюл. / Федеральная служба по интеллектуальной собственности. – 2016. – № 14.

12. *Рубашкин*, *В.Ш.* Семантический (концептуальный) словарь для информационных технологий: методы формирования и ведения словаря / В.Ш. Рубашкин, Д.Г. Лахути // Научно-техническая информация. Серия 2: Информационные процессы и системы. – 2000. – № 7. – С. 1–9.

13. *Батура*, *Т.В.* Семантический анализ и способы представления смысла текста в компьютерной лингвистике / Т.В. Батура // Программные продукты и системы. – 2016. – № 4. – С. 45–57. – DOI: 10.15827/2311-6749.21.220.

14. *Горкун*, *О.П.* Подходы к извлечению объектов и фактов из неструктурированных текстов / О.П. Горкун // Advanced science: сборник статей VI Международной научно-практической конференции, Пенза, 12 января 2019 года. – Пенза: «Наука и Просвещение» (ИП Гуляев Г.Ю.), 2019. – С. 70–72.

15. *Семантический анализ для автоматической обработки естественного языка* [Электронный ресурс] / Научно-технический центр ФГУП «ГРЧЦ» (НТЦ), 2021. – Режим доступа: https://rdc.grfc.ru/2021/09/semantic_analysis/. – Дата доступа: 20.04.2022.

16. *Найханова*, *Л.В.* Основные типы семантических отношений между терминами предметной области / Л.В. Найханова // Известия высших учебных заведений. Поволжский регион. Технические науки. – 2008. – № 1 (5). – С. 62–71.

17. *Сулейманова*, *Е.А.* Семантический анализ контекстных дат / Е.А. Сулейманова // Программные системы: теория и приложения. – 2015. – Т. 6. – № 4 (27). – С. 367–399.

18. *Бутов*, *А.Л.* Метод и алгоритмы извлечения фактов в информационно-аналитических системах / А.Л. Бутов, А.Т. Миргалеев // Инновации в информационно-аналитических системах: сб. научн. трудов. – Курск: Науком 2013. – № 2. – С. 36–52.

19. *Гаршина*, *В.В.* Разработка контекстно свободных грамматик с использованием Томита-парсера для задач извлечение фактов из неструктурированных текстов / В.В. Гаршина, В.Е. Панин,

И.В. Коротких // Информатика: проблемы, методология, технологии: Сборник материалов XIX международной научно-методической конференции, Воронеж, 14–15 февраля 2019 года; под ред. Д.Н. Борисова. – Воронеж: Издательство «Научно-исследовательские публикации» (ООО «Вэлборн»), 2019. – С. 1447–1452.

20. *Microsoft.Recognizers.Text provides recognition and resolution of numbers*, *units*, *and date* / time expressed in multiple languages [Electronic resource]. – 2022. – Mode of access: https://git-hub.com/microsoft/Recognizers-Text. – Date of access: 15.04.2022.

21. *Томита-парсер – Технологии Яндекса* [Электронный ресурс]. – Режим доступа: https://yandex.ru/dev/tomita. – Дата доступа: 15.04.2022.

22. *RCO Fact Extractor SDK | RCO* [Электронный ресурс]. – 2022. – Режим доступа: http://www.rco.ru/?page_id=3554. – Дата доступа: 15.04.2022.

23. *Shieber*, *S.M*. Evidence against the context-freeness of natural language / S.M. Shieber // Studies in Linguistics and Philosophy. – 1985. – Vol. 8, № 3. – P. 333–343.

24. *Способ поиска в тексте совпадений с шаблонами*: пат. 037156 Респ. Беларусь, МПК G06F 17/27, G06F 17/24 / Д.А. Сурков, К.А. Сурков, Ю.М. Четырько, И.В. Шимко, В.А. Савёнок; заявитель общество с ограниченной ответственностью «Незабудка Софтвер» – № 201800581; заявл. 24.09.18; опубл. 31.03.20 // Официальный бюл. / Евразийская патентная организация. – 2021. – № 2.

25. *Nevod is a language and technology for pattern-based text search* [Electronic resource]. – Mode of access: https://github.com/nezaboodka/nevod. – Date of access: 15.04.2022.

26. *Solves basic Russian NLP tasks*, *API for lower level Natasha projects* [Electronic resource]. – Mode of access: https://github.com/natasha/natasha. – Date of access: 16.04.2022.

27. *Stanford CoreNLP*: A Java suite of core NLP tools [Electronic resource]. – 2022. – Mode of access: https://github.com/stanfordnlp/CoreNLP. – Date of access: 16.04.2022.

28. *Official page for Language Server Protocol* [Electronic resource]. – Mode of access: https://microsoft.github.io/language-server-protocol. – Date of access: 23.04.2022.

29. *Nevod language extension for VS Code* [Electronic resource]. – Mode of access: https://github.com/nezaboodka/nevod-vscode. – Date of access: 23.04.2022.

30. *Nevod Basic Patterns* [Electronic resource]. – Mode of access: https://github.com/nezaboodka/nevod-patterns. – Date of access: 23.04.2022.

31. *Intelligent Virtual Agents and Bots | Microsoft Power Virtual Agents* [Electronic resource]. – Mode of access: https://powervirtualagents.microsoft.com/en-us/. – Date of access: 15.04.2022.

32. *Recognizers Test Cases Specs for Date Extractor* [Electronic resource]. – Mode of access: https://github.com/microsoft/Recognizers-Text/blob/master/Specs/DateTime/English/DateExtractor.json. – Date of access: 18.04.2022.

33. *Селезнева*, *Ю.А*. Набор текста на ПК: Слепой десятипальцевый метод печати / Ю.А. Селезнева – СПб.: «Корона Принт», 2005. – 64 с.

34. *Wolfram Language & System Documentation Center* [Electronic resource]. – Mode of access: https://reference.wolfram.com/language/. – Date of access: 25.04.2022.

Поступила в редакцию 23.06.2022.

Информация об авторах

Таранчук Валерий Борисович – д.ф.-м.н., профессор

Савёнок Владислав Александрович – магистр технических наук