УДК 004.9
DOI: 10.46548/21vek-2021-1056-0006

АВТОМАТИЗАЦИЯ СЕМАНТИЧЕСКОГО АНАЛИЗА ИНФОРМАЦИИ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ В РЕЗУЛЬТАТАХ ПОИСКОВОЙ ВЫБОРКИ

Шевнина Юлия Сергеевна, кандидат технических наук, доцент института
Системной и программной инженерии и информационных технологий
Томишинец Александр Михайлович, магистрант института
Системной и программной инженерии и информационных технологий
Национальный исследовательский университет МИЭТ
(124498, г. Москва, г. Зеленоград, площадь Шокина, дом 1, e-mails: yusm@rambler.ru, miracore.cr@gmail.com)

Аннотация. Работа посвящена разработке нейросетевого алгоритма семантического анализа информации на естественном языке, который позволяет формировать список терминов и отображать их на графе связности или на семантической карте. Показана реализация алгоритма в автоматизированной информационной системе семантического анализа текста на естественном языке. Для определения веса термина в документе и связанности терминов в алгоритме семантического анализа информации используется расчетная величина индекса обратной частоты документа, учитывающая, в отличие от существующих подобных алгоритмов, влияние повторений в тексте. Это позволило точнее определять семантические особенности входной текстовой информации, учитывать связанность терминов и определять многозначные слова. Разработанный алгоритм семантического анализа информации на естественном языке и его реализация в автоматизированной информационной системе позволили значительно повысить эффективность обработки поисковой текстовой выборки по ключевым словам и представить понятную визуализацию особенностей текста в виде графа связанности терминов или семантической карты. Разработанная автоматизированная информационная система семантического анализа текста на естественном языке имеет преимущество перед аналогами, поскольку позволяет одновременно формировать список терминов с учетом их повторяемости и многозначности и визуализировать текст в виде графа связанности и семантической карты.

Ключевые слова: естественный язык, семантический анализ текста, индекс обратной частоты, *UDPipe*, поиск текста по ключевым словам, анализ информации, многозначность терминов, визуализация текста, граф связанности, семантическая карта, учет редких терминов.

AUTOMATION OF SEMANTIC ANALYSIS OF INFORMATION IN NATURAL LANGUAGE AS RESULTS OF SEARCH SAMPLING

Shevnina Yulia Sergeevna, candidate of technical sciences, associate professor of the institute of
System and software engineering and information technologies
Tomishinets Alexander Mikhailovich, master's student of the institute of
System and software engineering and information technologies
National Research University MIET
(124498, Moscow, Zelenograd, Shokin square, building 1, e-mails: yusm@rambler.ru, miracore.cr@gmail.com)

Abstract. The work is devoted to the development of a neural network algorithm for semantic analysis of information in natural language, which allows you to form a list of terms and display them on a connectivity graph or on a semantic map. The implementation of the algorithm in an automated information system for semantic analysis of natural language text is shown. To determine the weight of a term in a document and the relatedness of terms in the algorithm for semantic information analysis, the calculated value of the index of the reciprocal frequency of the document is used, which, in contrast to existing similar algorithms, takes into account the effect of repetitions in the text. This made it possible to more accurately determine the semantic features of the input textual information, take into account the relatedness of terms and define polysemantic words. The developed algorithm for semantic analysis of information in natural language and its implementation in an automated information system made it possible to significantly increase the efficiency of processing a search text selection by keywords and to present a clear visualization of text features in the form of a graph of relatedness of terms or a semantic map. The developed automated information system for semantic analysis of text in natural language has an advantage over analogues, since it allows you to simultaneously form a list of terms, taking into account their repetition and polysemy, and to visualize the text in the form of a graph of connectivity and a semantic map.

Keywords: natural language, semantic text analysis, reverse frequency index, UDPipe, text search by keywords, information analysis, term ambiguity, text visualization, connectivity graph, semantic map, accounting for rare terms.

Введение. Сегодня в мире остаётся всё меньше регионов, в которых отсутствует доступ к всемирной сети Интернет. С каждым днём, благодаря огромному количеству пользователей, объём новой информации увеличивается экспоненциально [1-5]. Однако, среди всей этой информации лишь небольшая часть оказывается полезной, выделение полезной информации из общего потока представляет сложную задачу, для

решения которой было разработано множество поисковых систем. Предварительный семантический анализ информации представляет собой первую стадию отбора информации в поисковых системах, призванный сформировать у пользователя представление об информации до непосредственного ознакомления с ней. Одним из атрибутов, формирующих данное представление, являются ключевые слова. Они позволяют сравнивать между собой некоторое количество исходной информации, определяя наиболее релевантное пользовательскому поисковому запросу [6-9].

Исходя из этого, можно выделить три задачи, которые решает предварительный анализ:

– выбор наиболее подходящего источника информации;

– формирование первоначального представления об источнике;

– сравнения различных источников информации или версий одной и той же информации.

Предварительный анализ является трудозатратным для человека и требует дополнительных знаний, поскольку семантика исходной информации не полно отражается ключевыми словами. Для решения этой актуальной проблемы предлагается использовать автоматизированный семантический анализ информации, что позволит повысить эффективность анализа результатов поисковых запросов за счет нейросетевой предварительной семантической обработки большого количества текстовой информации [10-15].

Целью работы является создание алгоритма семантического анализа информации на естественном языке в поисковой выборке и его реализация в информационной системе.

Материалы и результаты исследования. Согласно существующим алгоритмам, анализ естественного языка требует предварительных синтаксических и морфологических манипуляций над исходным текстом и проводится в несколько этапов [16, 17]:

– разделение исходного текста на предложения;

– разбиение полученных предложений на слова (токенизация);

– определение начальной формы слов (лемматизация);

– определение части речи каждого из слов (частеречный анализ).

Данные манипуляции составляют базовую часть анализа текста, на её основе происходит выстраивание более сложных конструкций. На следующем этапе происходит поиск сложных и специфичных слов или терминов [1]. Далее происходит преобразование информации на естественном языке в данные с последующей визуализацией элементов текста либо структур. Текст или структура текста представляется графом, в котором вершины – ключевые слова, словосочетания или термины, выделенные из текстов, соединенные ребрами по определенным правилам.

На сегодняшний день существуют различные программные продукты для обработки текста большого объёма и формирования единого глоссария [18], такие программы применяются в частности для перевода текста на другой язык. Использование глоссария позволяет сохранить единство терминологии при переводе и уменьшить время выбора нужного варианта перевода [2]. Подобное программное обеспечение применяется для *SEO* анализа [3]. Наиболее известными являются программы: *Simple Concordance Program, MonoConc, MultiTerm Extract,* которые, поддерживая большое количество мировых языков, позволяют создавать списки терминов и определять для них синонимы. К популярным программам визуализации информации на основе естественного языка относятся: *Wordle, Taxedo, Many Eyes*. Эти программы дают возможность представлять исходный текст в виде облака тегов с различными настройками.

На осмновании проведенного аналитического обзора современных программных средств формирования списка терминов и его визуализации можно утверждать, что в настоящий момент не существует универсального программного продукта, способного осуществлять и поиск терминов и визуализацию текста. Кроме этого, облака тегов не предоставляют семантическую информацию о тексте [19]. Для решения этой проблемы требуется разработать алгоритм семантического анализа информации на естественном языке, который позволит формировать список терминов и отображать их на графе связности или на семантической карте (рис. 1).

Рисунок 1 – Алгоритм семантического анализа информации на естественном языке

Предварительная синтаксическая и морфологическая обработка текстовой информации осуществлялась с помощью библиотеки *UDPipe* [3] Для обучения нейронной сети получены и рассчитаны индексы

обратной частоты: 129 тысяч слов после просмотра 420 тысяч страниц Википедии. Для определения веса термина в документе и связанности терминов используется расчетное значение индекса обратной частоты документа [20]

$$idf(word) = log\ (\frac{DocCount}{DocCount_{word}}).$$

В процессе обучения нейронной сети была обоснована необходимость учета влияния повторений в тексте и коэффициента определения влияния повторений:

$$conc(word) = 1 + k \cdot log\ (\frac{Count_{word}}{DocCount_{word}})$$

Это позволило точнее определять семантические особенности входной текстовой информации, учитывать связанность терминов и определять многозначные слова.

Разработанный алгоритм предполагается использовать в автоматизированной информационной системе семантического анализа текста на естественном языке. К системе предъявляются следующие требования:

1) возможность анализа документов, содержащих текст на одном естественном языке, включающего:

– поиск ключевых слов в тексте;

– составление семантических карт из слов текста;

– создание графа связности для терминов;

– легенда встречаемости слов;

2) сравнение данных анализа нескольких текстов;

3) возможность загрузки результатов анализа текста;

4) ранение результатов анализа текста;

5) возможность экспортировать результаты анализа текста.

На основе сформулированных требований к системе построена функциональная модель информационной системы, представленная на рисунке 2. Основным актором является «*Пользователь ИС*», которому доступен весь функционал информационной системы.

Важной частью алгоритма семантического анализа информации на естественном языке является процедура поиска редких слов [21], который представлен в виде модели взаимодействия объектов (рис. 3). Пользователь ИС указывает текстовый файл, который обрабатывается с помощью словаря с индексами обратной частоты и данных, полученных из языковой модели настроенной на проведение токенизации, частеречного анализа. После обработки текста результирующие данные возвращаются на компоненты интерфейса и отображаются пользователю.

Для реализации требования хранения результатов анализа текста спроектирована структура базы данных. Модель базы данных представлена на рисунке 4, список таблиц базы данных представлены в таблице 1.

Для программной реализации автоматизированной информационной системы семантического анализа текста на естественном языке выбран язык программирования *C++*, интерфейсы реализованы с применением *Vue.js*.

Рисунок 2 – Функциональная модель автоматизированной информационной системы семантического анализа текста на естественном языке

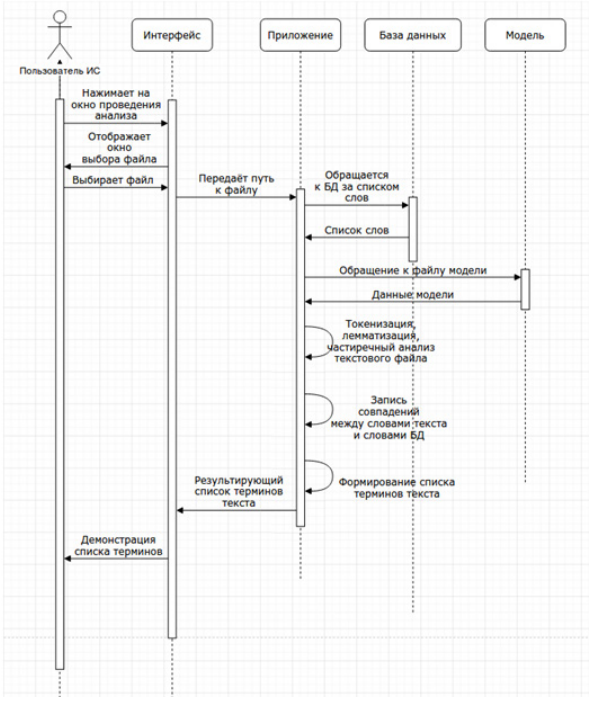

Рисунок 3 – Модель взаимодействия объектов автоматизированной информационной системы семантического анализа текста на естественном языке

Таблица 1 – Таблицы базы данных

Таблица	Описание
analyzes	Данные об анализе информации, необходимые для загрузки предыдущих анализов из архива. Данные анализа информации, получаются из обработки текста.
terms	Термины для анализа текста.
words	Слова, записанные в словаре базы данных, имеющие необходимые характеристики для воспроизведения результатов анализа текста.

Рисунок 4 – Модель базы данных

Заключение. Разработанный алгоритм семантического анализа информации на естественном языке и его реализация в автоматизированной информационной системе позволили значительно повысить эффективность обработки поисковой текстовой выборки по ключевым словам и представить понятную визуализацию особенностей текста в виде графа связанности терминов или семантической карты. Эмпирически полученные формулы по расчету обратной частоты документа позволили точнее определять семантические особенности входной текстовой информации, учитывать связанность терминов и определять многозначные слова, что является преимуществом по сравнению с аналогичными программными решениями [20].

СПИСОК ЛИТЕРАТУРЫ:

1. Сравнение программ по выделению терминов // englishelp.ru URL: https://www.englishelp.ru/translator/articles-for-translator/11-extract-terms-tools.html (дата обращения: 07.06.2020)
2. Визуализации, типы, сравнения // Научно-учебная группа «Методы анализа и визуализации веб-корпусов» URL: https://cs.hse.ru/vitext/visualize (дата обращения: 05.06.2020)
3. Анализатор для работы с текстом // UDPipe URL: http://ufal.mff.cuni.cz/udpipe (дата обращения: 03.06.2020)
4. Марков А.В. Проведение сравнительного анализа двух нейронных сетей EAST и PSENET в задаче детекции текста на изображениях прейскурантов // Евразийский союз ученых – М.: 2020. – С. 41-46.
5. Potaraev V. Analysis of relation types in semantic network used for text classification // Открытые семантические технологии проектирования интеллектуальных систем. – М.: 2020. – С. 305-308.
6. Злоказов К. Системно-функциональный семиотический подход к анализу поликодового текста: современное состояние и перспективы // Юрислингвистика. – М.: 2018. – С. 126-133.
7. Yuzhakova Yu.A., Ivanova D.S., Lavrentev V.A., Somova M.V. Addressing the issue of employing information technology in close reading // Russian linguistic Bulletin. – М.: 2020. – С. 67-70.
8. Воронин В.М., Курицин С.В., Наседкина З.А., Ицкович М.М. Использование латентного семантического анализа как альтернативы пропозиционального анализа в исследованиях понимания текста // Гуманизация образования. – М.: 2017. – С. 11-19.
9. Гусейнова К.Э. Критический дискурс-анализ как эффективный метод качественного анализа научных и образовательных текстов // Nauka.me. – М.: 2017. – С. 12.
10. Диковицкий В.В. Семантический анализ текста с применением нейросетевого анализа морфологии и синтаксиса // Труды Кольского научного центра РАН. – М.: 2017. – С. 109-115.
11. Рычагов С.А. Использование латентно-семантического анализа для автоматической классификации текстов // Международный журнал информационных технологий и энергоэффективности. – М.: 2017. – С. 28-33.
12. Tregubov A.S., Malyugina O.V. Neural network model for finding contradictions in natural language use using tripletloss function // Components of Scientific and Technological Progress. – М.: 2020. – С. 9-14.
13. Таныгина Е.А., Власова А.О., Миронюк Т.В. Экспериментальное исследование связи естественных языков и языков программирования высокого уровня // Известия Юго-Западного государственного университета. Серия: Лингвистика и педагогика. – М.: 2020. – С. 142-159.
14. Рожкин П.А. Разработка интеллектуальной системы глубинного машинного обучения для распознавания естественных языков // Инженерные кадры - будущее инновационной экономики России. – М.: 2017. – С. 118-121.
15. Aurora S. Natural language as a technological tool // Technology and Language. – М.: 2021. – С. 86-95.
16. Вострикова Е.В., Куслий П.С. Грамматикализация категориальной ошибки и естественный язык // Вопросы философии. – М.: 2020. – С. 116-126.
17. Уткин Л.В., Мелдо А.А., Ковалев М.С., Касимов Э.М. Простой общий алгоритм объяснения диагноза на выходе интеллектуальной системы диагностики в терминах примитивов естественного языка // Международная конференция по мягким вычислениям и измерениям. – М.: 2020. – С. 242-245.
18. Polyakov O.M. Linguistic data model for natural languages and artificial intelligence. Part 4. Language // Discourse. – М.: 2020. – С. 107-114.
19. Цитульский А.М., Иванников А.В., Рогов И.С. NLP – Обработка естественных языков // StudNet. – М.: 2020. – С. 467-475.
20. Субботин А.Н. Алгоритм классификации потоков текстовой информации на естественном языке // Научно-технический вестник Поволжья. – М.: 2020. – С. 18-20.
21. Трегубов А.С. Метод обучения нейронной сети с применением функции потерь tripletloss для задачи анализа текстов естественного языка // Евразийское Научное Объединение. – М.: 2020. – С. 138-139.

Статья поступила в редакцию 12.10.2021
Статья принята к публикации 07.12.2021