

На правах рукописи



Бермудес Сото Хосе Грегорио

**РАЗРАБОТКА И ИССЛЕДОВАНИЕ МЕТОДА И АЛГОРИТМОВ
СЕМАНТИЧЕСКОГО СРАВНЕНИЯ НАУЧНЫХ ТЕКСТОВ**

Специальность 05.25.05 – «Информационные системы и процессы»

АВТОРЕФЕРАТ

диссертации на соискание учёной степени
кандидата технических наук

Таганрог – 2018

Работа выполнена на кафедре системного анализа и телекоммуникаций Института компьютерных технологий и информационной безопасности Федерального государственного автономного образовательного учреждения высшего образования «Южный Федеральный Университет» в г. Таганроге.

Научный руководитель: **Рогозов Юрий Иванович**
доктор технических наук, профессор, ФГАОУ ВО «Южный Федеральный Университет», кафедра системного анализа и телекоммуникаций Института компьютерных технологий и информационной безопасности, заведующий

Официальные оппоненты: **Черников Борис Васильевич**
доктор технических наук, профессор, Российский экономический университет имени Г.В. Плеханова, кафедра информатики, доцент

Смирнов Иван Валентинович
кандидат физико-математических наук, Институт системного анализа ФИЦ ИУ РАН, отдел «Интеллектуальный анализ информации», заведующий

Ведущая организация: Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН)

Защита состоится 21 декабря 2018 г. в 14.00 на заседании диссертационного совета Д 212.208.25 Южного федерального университета по адресу: 347922, г. Таганрог, ул. Чехова, 2, ауд. И-409.

С диссертацией можно ознакомиться в Зональной научной библиотеке Южного федерального университета по адресу: 344090, г. Ростов-на-Дону, ул. Зорге 21-ж.

Диссертация в электронном виде доступна по адресу:

<http://hub.sfedu.ru/diss/announcement/4f7051e0-12b3-4bd0-8a0c-e63352cb3d23/>

Автореферат разослан 13 октября 2018 г.

Учёный секретарь
диссертационного совета



Ю.А. Брюхомицкий

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования.

В связи с возрастающей потребностью в обработке текстовой информации на естественном языке тема диссертационного исследования является актуальной. К задачам обработки текстовой информации на естественном языке относятся: информационный поиск, сравнение текстов, проверка на плагиат, вопросно-ответные системы, машинный перевод, извлечение информации, автоматизированное аннотирование и реферирование, диалоговые системы, анализ и синтез текста и другие.

Задачи обработки текстовой информации характеризуются разным набором входных данных и требуемой формой их представления, целевым результатом и предлагаемыми подходами. Полученные в данной области результаты и их практическая реализация показывают, что некоторые из этих задач требуют дальнейших исследований. Одной из таких задач является сравнение текстов, а именно задача выявления семантической близости двух на естественном языке.

Проблема автоматической обработки естественного языка при выявлении семантической близости заключается в том, что различные языки имеют различные семантические и грамматические особенности, а существующие алгоритмы успешно используются только для обработки одного отдельно взятого языка. Преодоление этой проблемы требует создания средств построения смысловых конструкций естественного языка, что на данный момент является не решенной задачей.

Для машинной обработки текстов на естественном языке необходимо, прежде всего, решить задачу создания средств преобразования языка (например русского, испанского, английского и т. д.) в формализованный, подобный языку программирования. Общие принципы создания средств для систем обработки текстов включают следующие компоненты: фрагментация или разделение, морфологический анализ, синтаксический и семантический анализ, из которых выход одного компонента является входом для следующего.

Существующие достижения в области обработки текстов на естественном языке включают: метод синтактико-семантических шаблонов (Чубинидзе К.А.), подход к автоматизации систем смысловой обработки текстов (Сбойчаков К.О.), метод концептуального анализа текстов в системах автоматической обработки научно-технической информации (Козачук М.В.) и др. Стоит отметить, что существующие подходы и методы не рассматривают в качестве критерия сравнения текстов их смысл, под которым в диссертационном исследовании в отличие от известных понимается текстовый пассаж, который не содержит анафорических связей, ассоциирующихся со словами другого текстового пассажа, содержащий по крайней мере один глагол, тип и категория которого выражает действие.

На основе предложенного выше определения смысла в диссертации сформулирован подход сравнения научных текстов, на базе которого разработаны метод и алгоритмы семантического сравнения научных текстов. Данные метод и алгоритмы могут быть использованы в приложениях автоматизированного обнаружения плагиата для повышения его эффективности.

Цель работы заключается в формулировке подхода сравнения научных текстов, а также разработке и исследовании метода и алгоритмов семантического сравнения на его основе, которые позволяют извлекать текстовые сегменты с полным смыслом и обнаруживать семантическое сходство.

Для достижения поставленной цели решаются следующие задачи:

1. Сформулировать определения значимого текстового пассажа и смысла научного текста;
2. На основе сформированных определений разработать подход сравнения научных текстов;
3. Реализация сформулированного подхода сравнения научных текстов в виде метода, состоящего из:
 - 3.1. Методики сегментации текстов на естественном языке, которая гарантирует извлечение значимых текстовых фрагментов, сохраняющих смысл текста;

3.2. Методики автоматизированного сравнения двух текстов на естественном языке, которая обнаруживает семантическое сходство, независимо от используемых слов;

4. Разработка алгоритмов сегментации и сравнения, позволяющих оценивать сходство научных текстов и реализующих предложенные методики.

5. Проведение экспериментальных исследований методик сегментации и сравнения текстов на естественном языке и оценка их по критерию выявления совпадений с человеческими мнениями.

Объектом исследования является информационная технология интерпретации текстов на естественном языке в части методов семантического, синтаксического и прагматического анализа текстовой информации.

Предметом исследования являются алгоритмы семантико-синтаксической обработки научных текстов на русском языке и методы автоматической обработки текста.

Научная новизна диссертационной работы:

В диссертации получены следующие новые научные и практические результаты:

1. Введено определение понятия значимого текстового пассажа как формального представления самостоятельной части текста посредством идентификации глаголов и анафорических связей.

2. Предложено формальное представление смысла научного текста в виде текстового пассажа, которое в отличие от известных подразумевает, что пассаж не содержит анафорических связей, ассоциирующихся со словами другого текстового пассажа, и содержащий по крайней мере один глагол, тип и категория которого выражает действие.

3. Предложен подход к формализации процедуры сравнения научных текстов, отличающийся от известных сравнением значимых текстовых пассажей, состоящих из элементов смысла.

4. Предложен метод формального сравнения научных текстов, отличающийся от известных введением сегментации по семантическим

критериям с учётом синонимов, что позволяет автоматизировано обнаружить семантическое сходство между двумя сравниваемыми текстами, и учитывать при этом как морфологическую структуру текста, так и его лексико-семантическое содержание.

5. Предложена методика формализованной сегментации текста, отличающаяся от известных использованием семантического критерия, что позволяет автоматизировано извлекать значимые текстовые фрагменты, сохраняющие смысл текста.

6. Предложена методика формализованного сравнения научных текстов, отличающаяся от известных комбинацией предложенного формального и существующего семантического представления, а также сопоставлением синонимов, что позволяет обнаруживать семантическое сходство, независимо от используемых слов.

Основные положения, выносимые на защиту:

1. Формализованное представление смысла, идентифицированного как глагол и анафорические связи, в виде значимого текстового пассажа;

2. Подход к формализации процесса семантического сравнения научных текстов на естественном языке как сравнение их формального представления в виде значимых текстовых пассажей;

3. Метод семантического сравнения научных текстов как их формальных представлений в виде значимых текстовых пассажей;

4. Методика формализованной сегментации текстов на значимые текстовые пассажи, сохраняющие смысл текста;

5. Методика семантического сравнения научных текстов на основе их формального представления в виде совокупности значимых текстовых пассажей;

6. Оценка семантического сравнения научных текстов, критерии правильности и глубины вычисления семантической близости значимых текстовых пассажей.

Соответствие специальности. Тематика работы соответствует паспорту специальности 05.25.05 – Информационные системы и процессы:

п.1. Методы и модели описания, оценки, оптимизации информационных процессов и информационных ресурсов, а также средства анализа и выявления закономерностей в информационных потоках. Когнитивные модели информационных систем, ориентированных на человеко-машинное взаимодействие.

п.4. Лингвистическое обеспечение информационных систем и процессов. Методы и средства проектирования словарей данных, словарей индексирования и поиска информации, тезаурусов и иных лексических комплексов. Методы семантического, синтаксического и прагматического анализа текстовой информации с целью ее формализации для представления в базах данных и организации интерфейсов информационных систем с пользователями. Формат внешнего и внутреннего представления данных, коммуникативные и иные форматы данных и документов.

Практическая значимость работы заключается в том, что подход, метод, методики и алгоритмы, разработанные автором для извлечения значительных пассажей и сравнения текстов, позволили повысить стабильность распознавания плагиата вне зависимости от процента замены слов и улучшить на 40% обнаружение подобия по сравнению с существующими системами при замене более 50% слов исходного текста.

Использование результатов. Основные теоретические и практические результаты диссертационной работы использованы в организациях: Национальный политехнический экспериментальный университет национальных вооружённых Боливарианских сил (УНЕФА – UNEFA); Национальный центр по совершенствованию преподавания науки (СЕНАМЕК – SENAMEC), и Кафедра системного анализа и телекоммуникаций Института компьютерных технологий и информационной безопасности Федерального государственного автономного образовательного учреждения высшего образования «Южный Федеральный Университет», что подтверждается актами о внедрении.

Обоснованность и достоверность полученных результатов подтверждается строгостью математических выкладок, использованием методов компьютерной лингвистики, теории вероятностей, теории графов, теории информационного поиска и современных технологий программирования, теории интеллектуальных систем, морфологического анализа, семантико-синтаксического и статистического анализа, разработкой действующей программы и результатами экспериментов.

Апробация результатов работы. Основные результаты, полученные в ходе работы, докладывались и обсуждались:

- 21-22 мая 2015 г. Международная конференция «Инновационные технологии и дидактика в обучении» (Innovative Technologies and Didactics in Teaching – ITDT 2015), Мадрид. Испания. Выступление с докладом.

- 16-18 декабря 2015 г. XIII Всероссийская Научная конференция молодых ученых аспирантов и студентов «Информационные технологии, системный анализ и управление» (ИТСАиУ – 2015), г. Таганрог. Россия. Выступление с докладом.

- 3-4 мая 2016 г. Международная конференция «Инновационные технологии и дидактика в обучении» (Innovative Technologies and Didactics in Teaching – ITDT 2016), Тенерифе. Испания. Выступление с докладом.

- 5-7 сентября 2016 г. VII Международная научно-техническая конференция «Технологии разработки информационных систем» (ТРИС – 2016), г. Геленджик. Россия. Выступление с докладом.

- 16-19 ноября 2016 г. XIV Всероссийская Научная конференция молодых ученых аспирантов и студентов «Информационные технологии, системный анализ и управление» (ИТСАиУ – 2016), г. Таганрог. Россия. Выступление с докладом.

- 2-3 мая 2017 г. Международная конференция «Инновационные технологии и дидактика в обучении» (Innovative Technologies and Didactics in Teaching – ITDT 2017), Берлин. Германия. Выступление с докладом.

- 4-5 сентября 2017 г. VIII Международная научно-техническая конференция «Технологии разработки информационных систем» (ТРИС – 2017), г. Геленджик. Россия. Выступление с докладом.

- 23-27 апреля 2018 г. Международная конференция «Инновационные технологии и дидактика в обучении» (Information innovative technologies – I2T 2018), Прага. Чешской Республики. Выступление с докладом.

Публикации. По материалам диссертации автором опубликовано 14 печатных работ, в том числе четыре статьи в изданиях из списка, рекомендованного ВАК, в которых отражены основные результаты диссертационного исследования, также получено свидетельство об официальной регистрации программы для ЭВМ № 2018614861 от 18.04.2018 г.

Структура и объем диссертационной работы. Диссертация состоит из введения, четыре главы и заключения. Основной текст изложен на 110 страниц, содержит 32 рисунка, 4 таблицы, список литературы включает 76 наименований. В приложениях содержатся: программный код, листы данных Excel с записями результатов, полученных в ходе экспериментов; свидетельство о государственной регистрации программы для ЭВМ № 2018614861 от 18.04.2018 г.; акты внедрения результатов работы.

СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы диссертации, обозначены цели и задачи работы, определён предмет исследования, представлены основные научные результаты, показаны теоретическая и практическая значимость работы.

В первой главе рассмотрены основные понятия и определения, представления и понимания естественного языка, дан краткий анализ используемой структуры русского языка. В разделе также рассматриваются характеристики и грамматические атрибуты, которые используются, в частности, в научных текстах, и лингвистический феномен анафоры. Проведён обзор существующих подходов и методов к обработке научных текстов на

естественном языке, выявлено, что все известные системы носят, как правило, узконаправленный характер и не рассматривают в качестве критерия сравнения текстов их смысл. Проанализированы современные методы представления и обработки текста на естественном языке. Выполняется обзор подходов к автоматизированной обработке текста, в частности методов сегментации и сравнения текста. Сформулированы постановка задачи и цели исследования.

Предлагаются основополагающие понятия «значимый текстовый пассаж» и «смысл», лежащие в основе разрабатываемого подхода. Если учесть, что под текстовым сегментом понимается определённый набор букв, слов или фраз, которые являются частью текста, то есть любой отрезок речи, характеризующийся относительной семантической независимостью и полученный в результате сегментации текста, то значимый текстовой пассаж можно определить как отдельную часть текста, обладающую какой-то целостностью. Значимый текстовой пассаж получают посредством идентификации семантических аспектов, так что значимый текстовой пассаж имеет законченный семантический смысл.

С другой стороны, смысл – внутреннее содержание, значение чего-то. Поэтому сформулируем определение значимого текстового пассажа следующим образом: это сегмент текста без анафорических связей, ассоциируемых со словами другого сегмента и в котором имеется по меньшей мере один глагол, тип и категория которого, выражает действие.

Отличия предлагаемых определений от существующих заключаются в следующем: Значимый текстовой пассаж имеет какой-то смысла текста, в то время как любой текстовой сегмент понимается определённый набор букв, слов или фраз. Слово “смысл” имеет традиционную концепцию, но для этого исследования относится к форме смысла, что пассаж обладает наличием глагола и всех анафорических отношений. Формальное определение понятия значимого текстового пассажа описано ниже.

На основе предлагаемого определения можно сформулировать подход сравнения научных текстов, позволяющий сравнивать смысловые значения текстовых пассажей и повысить процент обнаружения подобия.

В целях формулирования подхода был проведен анализ существующих подходов, моделей и методов обработки текстов на естественном языке с точки зрения их способности выделять значимые текстовые пассажи (рис. 1). Анализ подтвердил, что сами по себе эти методы не решают поставленную задачу.

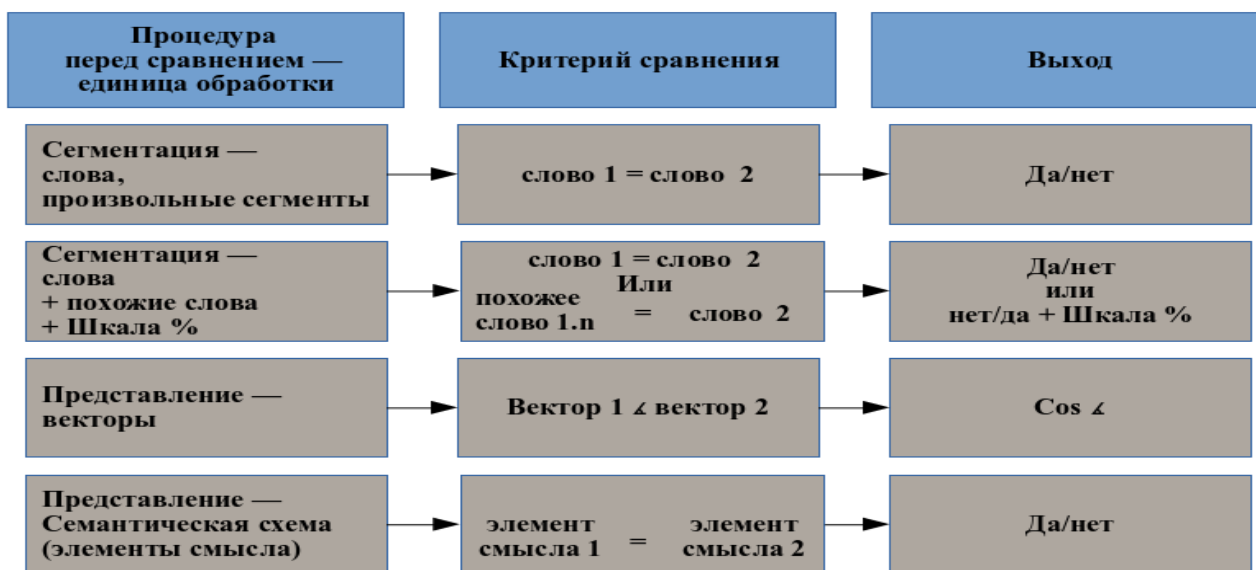


Рисунок 1 – Обобщение подходов к задаче текстового сравнения

Таким образом, чтобы система реализовывала семантическое сравнение текстов на основе выделенных значимых текстовых пассажей, необходимо разработать новый метод. В работе в качестве базовых будут использоваться методы синтаксического и семантического анализа текстов на естественном языке. Так же необходима модификация методов и совместное их использование для решения поставленной задачи. Необходимо разработать такой метод семантического сравнения текстов, который позволит оценивать тексты, написанные на естественном языке и определить степень их подобия к эталонным текстам, независимо от используемых слов и синтаксиса.

Во второй главе предлагается подход и метод семантического сравнения научных текстов. Основная идея подхода показана на рисунке 2.



Рисунок 2 – Предлагаемый подход для сравнения текстов

На основе подхода предлагается метод семантического сравнения научных текстов, который позволяет оценивать тексты, написанные на естественном языке и определить степень их подобия к эталонным текстам, независимо от используемых слов и синтаксиса. Это является основным отличием по отношению к существующим методам, основанным на точном выявлении слов и/или фраз. Предлагаемый метод показан на рисунке 3 и состоит из следующих методик, выстроенных в последовательность этапов: извлечение пассажей, разрешение анафор, семантическое представление, сравнение на близость и оценка. Автором предлагаются методики сегментации, сравнения на близость и оценки, остальные методики являются известными. Для реализации метода вводится формальное представление значимого текстового пассажа и описывается методика его сегментации.

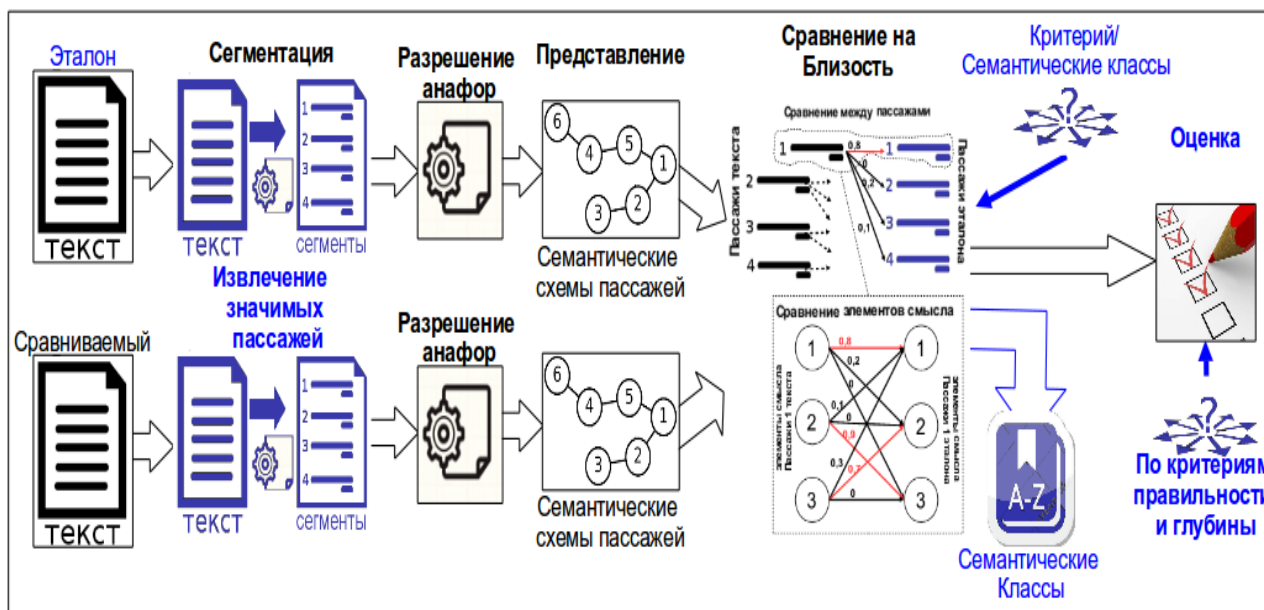


Рисунок 3 – Метод интеграции для сравнения текстов

Рассмотрим кратко этапы предлагаемого метода.

1. Методика извлечения пассажей (Этап 1). Из эталона, и сравниваемого текста, извлекаются текстовые пассажи, в соответствии с выявлением анафор.

Формально выделение значимых текстовых пассажей в предлагаемом методе следующее:

- Пусть D - документ, формируемый N фразы fj .

$$D = (fj, \dots fN)$$

- Определяются пассажи P_i , из документа D , следующим образом:

$$P_i = (fj, \dots fq); \dots P_n = (fk, \dots fN);$$

где: n – это количество пассажей, полученных в результате сегментации.

Но такие текстовые пассажи не случайные, а выделяются при помощи критерия остановки сегментации, указанным выше. Таким образом во всех пассажах P_{i+1} нет анафорических связей, связанных с предыдущим в P_i , в то время как в каждом пассаже P_i есть по меньшей мере один глагол типа A .

Таким образом, пусть A – это глагол в изъявительном наклонении (не инфинитив), или в условном наклонении, или в повелительном наклонении (который, выражает действие), h – любой анафорический элемент (личные местоимения, местоимения и наречия с грамматической функцией анафорической связи, относительные местоимения, указательные местоимения

и т.д.), c – предшествование предыдущей отсылки, и символ “:=” - однозначная зависимость, которая указывает, что h представляет c ($h:=c$), тогда определение пассажира P_i будет:

$$P_i = (f_j, \dots, f_q), / \forall_i ((\exists_A \in P_i) \wedge (\nexists_h \in P_{i+1} / (h:=c) \wedge c \in P_i))$$

2. Методика разрешения анафор (Этап 2). Для разрешения анафорических связей используется известный метод абдуктивной логики или объясняющего рассуждения.

3. Методика семантического представления (Этап 3). Схема строится из дерева зависимостей, создаваемых методом вопросов, в которых смысловая связь устанавливается от основного слова к зависимым. В методах, разработанных в предыдущих исследованиях, представлены семантические схемы предложений, используемые в информационном поиске для сравнения предложений или их фрагментов на близость. В этом смысле текстуальный пассаж может быть представлен с помощью функционала «смысловыразительности». Одна из схем семантического представления подтверждается на рисунке 4.

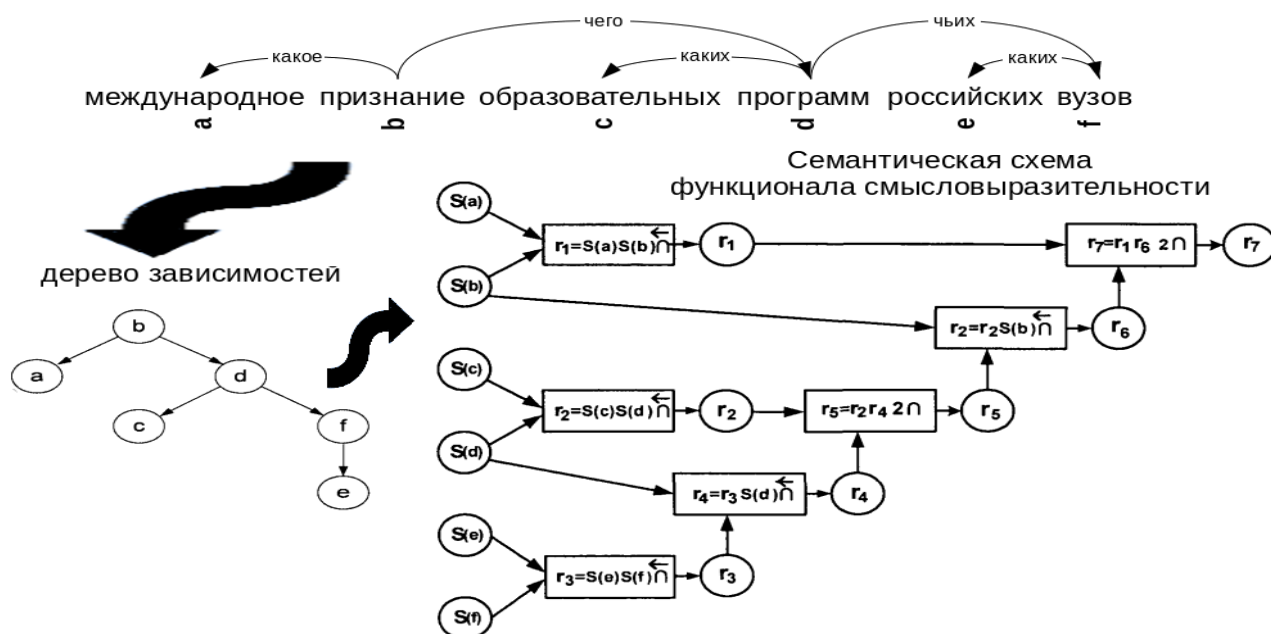


Рисунок 4 – Схема семантического представления

4. Методика сравнения на близость (Этап 4). Представлено определение степени подобия между текстовыми пассажами согласно критерию

семантической близости, основанном на семантических классах слов и predetermined эталоне. Критерий сравнения на близость, в соответствии с параметрами представлен в виде:

$$\Phi_{\text{семантик/класс}} = \frac{\sum_i^k \frac{p_i}{l}}{n}$$

где p – фактор совпадения между словами, участвующими в сравнении, для каждого элемента смысла, согласно семантическому классу в интервале $[0,1]$, $p = 1$, если слово идентично, $p = 0$ если слово вне семантического класса и $p = (0,1)$ в зависимости от степени синонимии; l – количество слов каждого элемента смысла; k – количество элементов смысла в текстовом пассаже сравниваемого текста, n – общее число элементов смысла в текстовом пассаже эталона.

Сравнение между пассажами, а затем сравнение между элементами смысла подтверждается на рисунке 5.

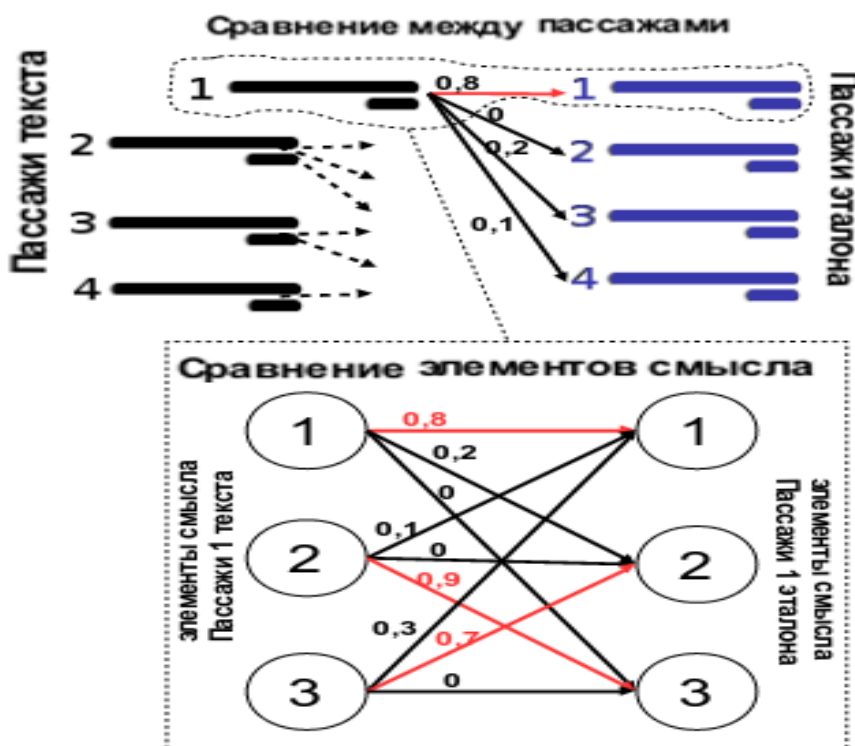


Рисунок 5 - Схема семантического сравнения

5. Методика оценка (Этап 5). Определение правильности и глубины, определяются по формулам:

$$C = \frac{\sum_1^q \Phi_i}{m}$$

где Φ – результат, полученный для каждого сравнения этапа 4; q – количество текстовых пассажей сравниваемого текста; и m – общее число текстовых пассажей эталона.

То есть коэффициент глубины S :

$$S = \frac{q}{m}$$

В то время как оценка может быть обозначена средним гармоническим двух ранее полученных коэффициентов:

$$F_R = \frac{2CS}{C + S}$$

Таким образом, в главе предложен подход и разработан метод сравнения текстов на уровне представления текстовых пассажей в семантических схемах, как наиболее эффективный по сравнению с имеющимися в настоящее время. Данный метод позволяет определить не только семантическую близость документов, представленных на естественном языке, но и дать количественную оценку сходства этих документов.

В третьей главе предложенные методики обработки текстов для текстового сравнения реализованы в виде алгоритмов. В первом пункте представлены общий алгоритм, основанный на сочетании существующих подходов, с особым включением сегментации текста в значительных текстовых пассажах, и использование семантических классов в самом текстовом сравнении.

Предлагаемые алгоритмы семантического сравнения текстов позволяют оценивать тексты, написанные на естественном языке и определять степень их подобия к эталонным текстам, независимо от используемых слов и синтаксиса.

Алгоритмы сегментации на значимых текстовых пассажах и текстового сравнения, интегрированы под общей методикой к обработке текстов и являются основами для проведения экспериментальных исследований методов сегментации и сравнения, которые позволяют проверить адекватность указанных алгоритмов. Схема сравнения представлена на рисунке 6.

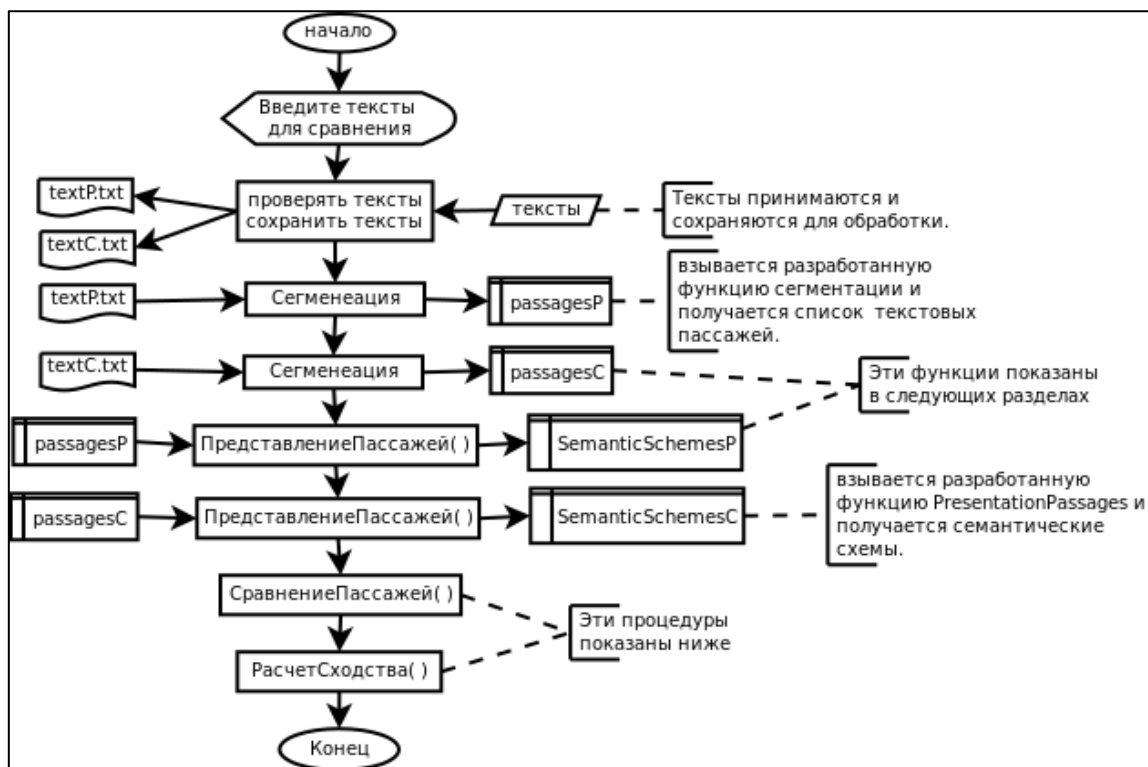


Рисунок 6 – Блок-схема текстового сравнения

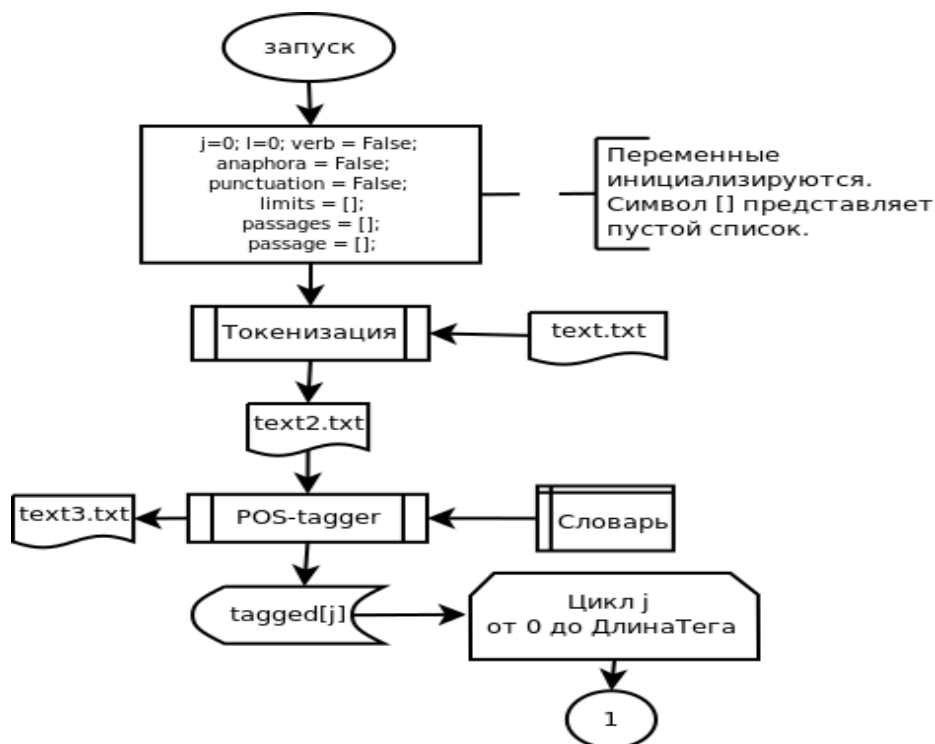


Рисунок 7 – Блок-схема сегментации часть 1

При сегментации требуется предшествующая сегментация по словам для того, чтобы идентифицировать грамматическую роль каждого слова и определить, какие слова выполняют функцию анафорической связи.

Также производится идентификация наклонений глаголов в сегменте, чтобы гарантировать, что эти сегменты содержат законченные значения. Эти предыдущие процессы известны как токенизация и POS-tagging, тогда алгоритм будет выглядеть, как показано в блок-схемах на рисунках 7 и 8:

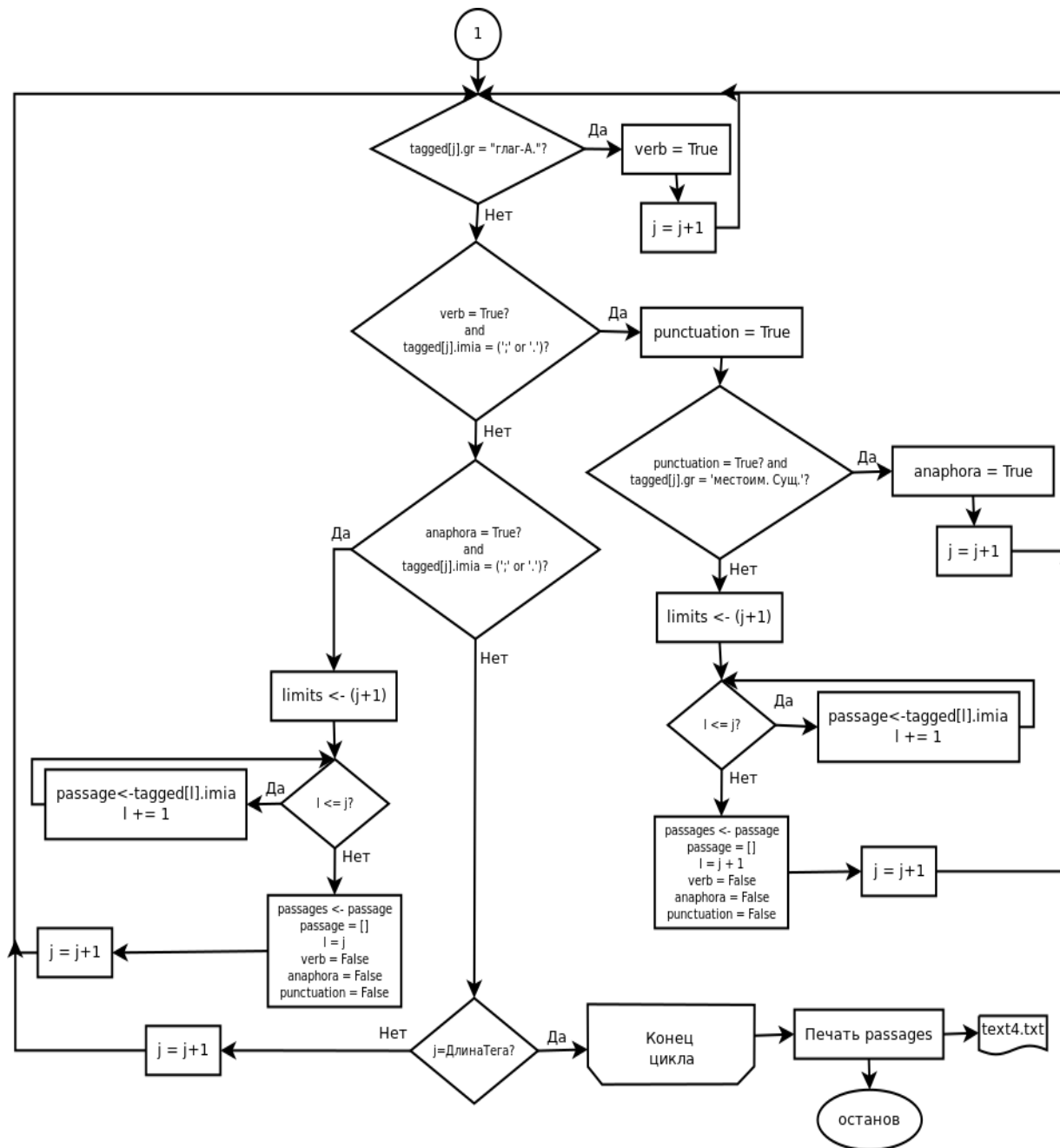


Рисунок 8 – Блок-схема сегментации часть 2

Определена степень подобия между текстовыми пассажами эталона и сравниваемого текста согласно критерию семантической близости. На входе алгоритма есть две семантические схемы текстового пассажа эталона и

сравниваемого текста соответственно. Алгоритмы совпадения и сравнения элементов представлены на блок-схемах на рисунках 9, 10, и 11.

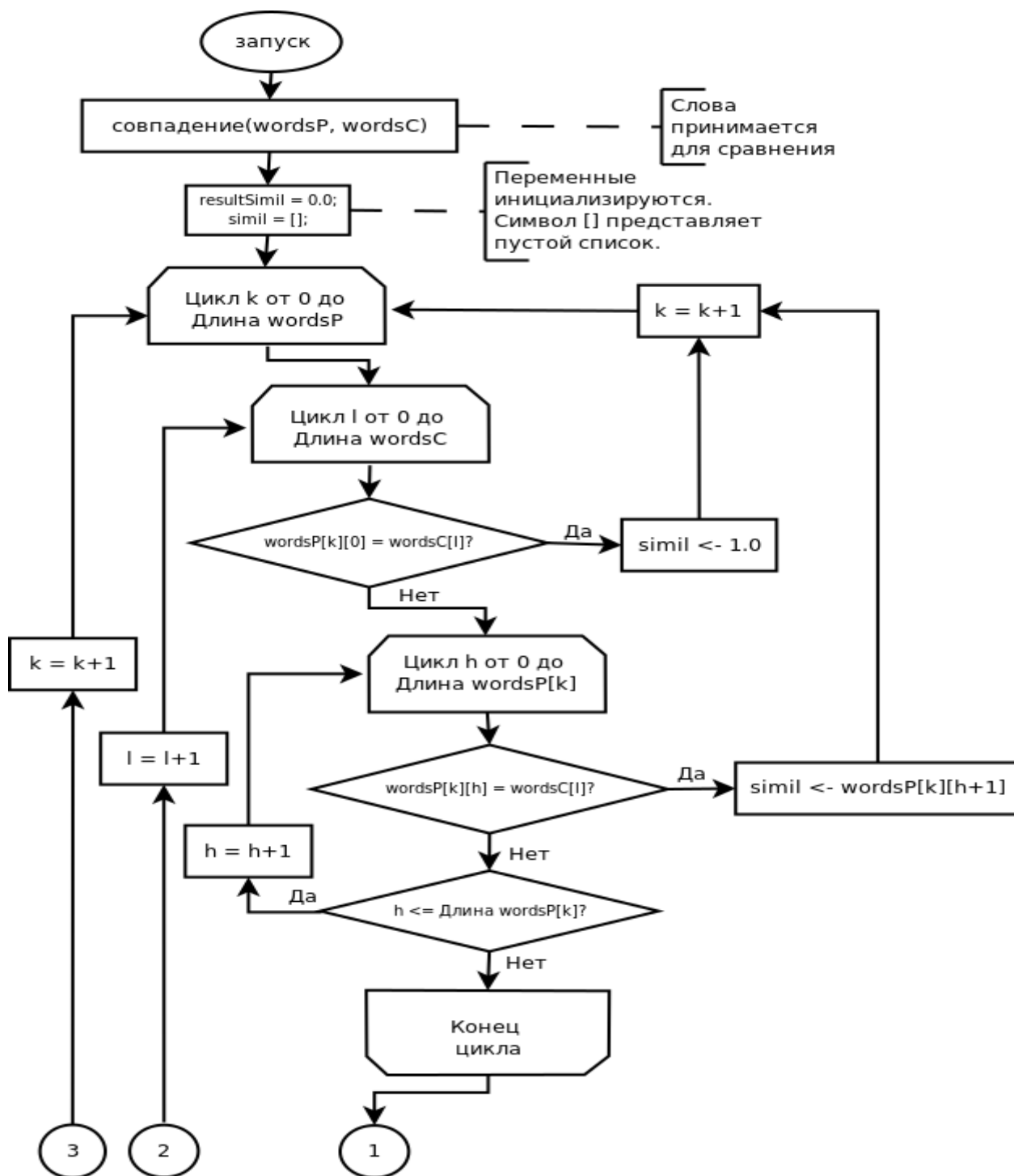


Рисунок 9 – Блок-схема совпадения часть 1

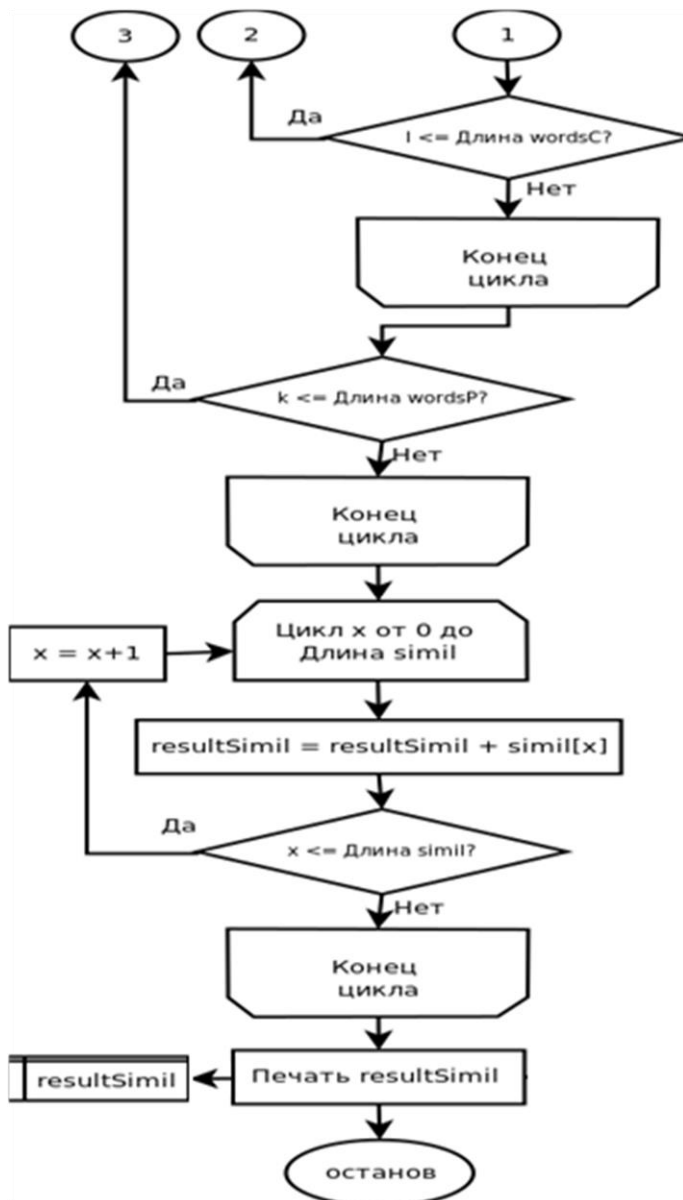


Рисунок 10 – Блок-схема совпадения часть 2

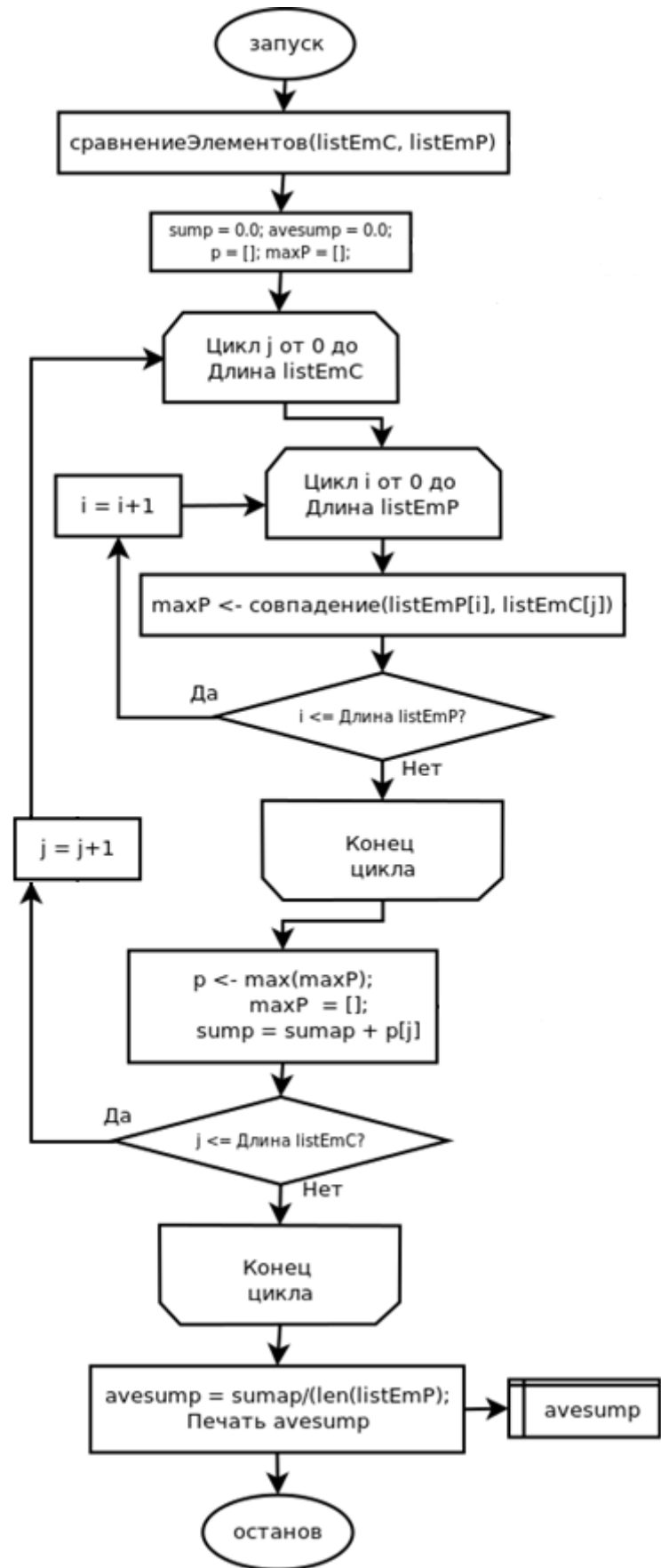


Рисунок 11 – Блок-схема сравнения элементов

Предложенные алгоритмы реализованы на языках Python и Ruby on Rails, получено свидетельство о Государственной регистрации программы для ЭВМ № 2018614861 от 18.04.2018 г. Эффективность разработанных алгоритмов оценивается в четвертой главе.

В четвёртой главе описываются эксперименты по сегментации и сравнению текста. Их результаты анализируются в сравнении с другими существующими методами, используя значимые текстовые пассажи и представление в семантических схемах, рассматривая семантические классы слов.

Для проведения экспериментов был разработан прототип автоматизированной системы сравнения семантического текста, разработанный на основе ранее описанных алгоритмов для лабораторной среды, в которой для выполнения текстового сравнения, нужно осуществить следующие шаги, а именно: токенизация, частеречная разметка, сегментация в значимые текстовые пассажи, построение деревьев зависимостей, построение семантических схем представления, сравнение сегментов и вычисление подобия. Все модули были реализованы за исключением построения деревьев зависимостей, которые были выполнены с помощью онлайн-приложения, имеющего API, который также выполняет разрешение анафорических связей. Эффективность рассмотренных алгоритмов сегментации и сравнения представлена на рисунках 12 и 13, и в таблицах 1 и 2.

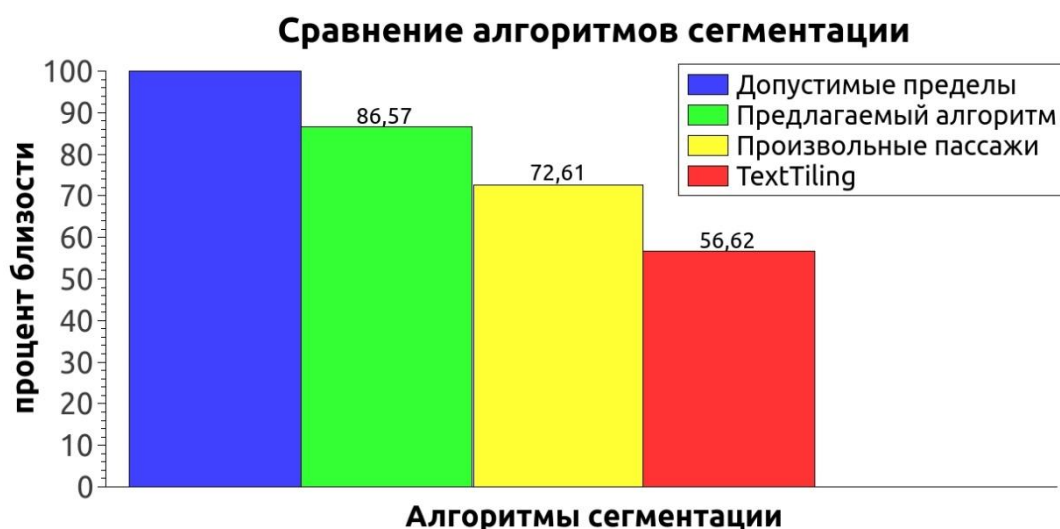


Рисунок 12 – Средние значения близости доступных пределов рассматриваемых алгоритмов

Из результатов, представленных на рисунке 12 следует что предложенный алгоритм превосходит алгоритм произвольных проходов на 14%, а алгоритм "TextTiling"- на 30% по отношению к проценту близости к пределам ручной сегментации.

Таблица 1 – Величины метрики "WindowDif" для 100 текстов

Метод/алгоритм	"WindowDif"
Предлагаемая методика значимых пассажей	0,1343
Произвольные пассажи	0,2739
"TextTiling"	0,4338

Из результатов, представленных в таблице 1, следует что предложенный алгоритм имеет в 2 раза меньшую погрешность, чем алгоритм произвольных пассажей для совпадений с допустимыми пределами, и в 3,2 раза меньшую погрешность, чем алгоритм "TextTiling".

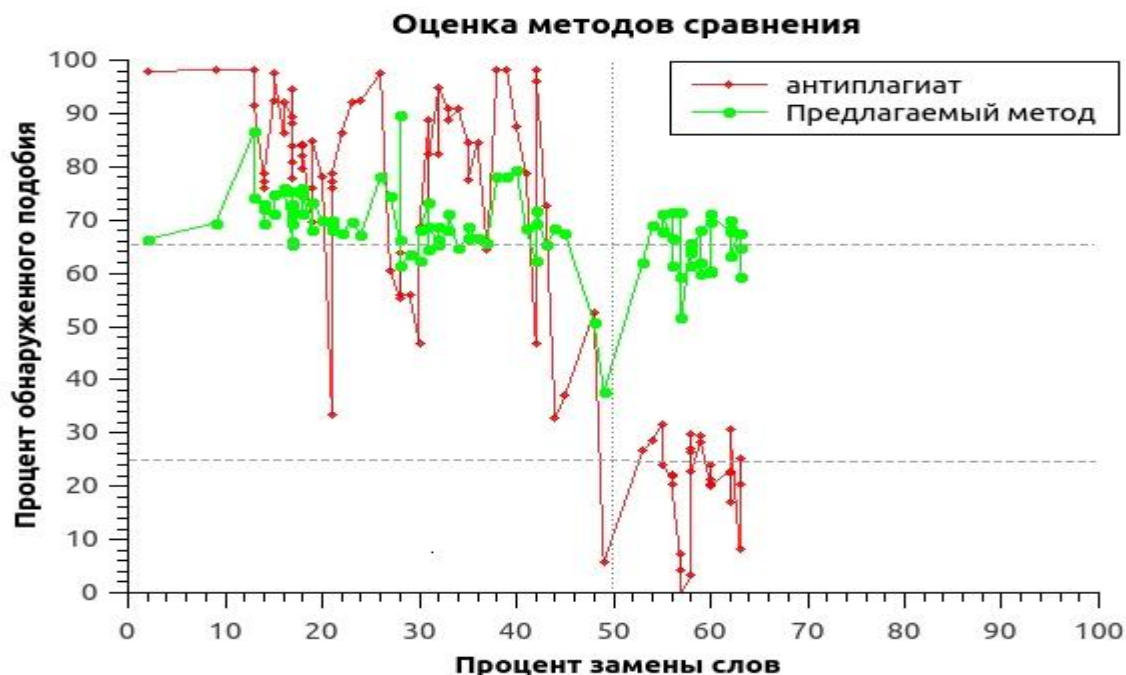


Рисунок 13 – Производительность методов сравнения для 100 текстов

Для сегментации предлагаемый метод для сот фрагментов текстов имеет самое низкое значение, что означает большую близость к допустимым пределам. Результаты текстового сравнения показали, что предложенный метод повышает стабильность распознавания плагиата вне зависимости от процента замены слов и улучшить на 40% обнаружение подобия по сравнению с существующими системами при замене более 50% слов исходного текста.

Таблица 2 Величины близости для 100 текстов

Метод/алгоритм	% близости
Предлагаемый Фсемантик/класс	67
Расстояние Левенштейна	61
Косинусное подобие	51
Коэффициент Джакарда	37

Предложенный метод в случае модифицированных текстов, приближается к эталону на 6% больше, чем расстояние Левенштейна, на 16% больше, чем косинусное подобие, и на 30% больше, чем остальные методы.

Особого упоминания заслуживают результаты, полученные и представленные для алгоритма расстояния Левенштейна, который имеет свою особенность. Если тексты вводятся с какими-то изменениями порядка абзацев по отношению к фрагментам, результаты уменьшаются от 20% до 40% в зависимости от сделанных изменений.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

1. Введены понятия «значимый текстовый пассаж» и «смысл», которые представляют собой базис для формулирования подхода сравнения научных текстов;

2. Введено определение понятия значимого текстового пассажа и представлена его формализация посредством идентификации глаголов и анафорических связей для обеспечения извлечения текстовых сегментов с законченным смыслом.

3. Предложено понятие смысла научного текста, которое в отличие от известных подразумевает, что текстовый пассаж не содержит анафорических связей, ассоциирующихся со словами другого текстового пассажа, и содержащий по крайней мере один глагол, тип и категория которого выражает действие.

4. Предложен подход сравнения научных текстов на естественном языке, отличающийся от известных сравнением значимых текстовых пассажей, состоящих из элементов смысла.

5. Предложен метод сравнения научных текстов, отличающийся от известных введением сегментации по семантическим критериям с учётом синонимов, что позволяет автоматизировано обнаружить семантическое сходство между двумя сравниваемыми текстами, и учитывать при этом как морфологическую структуру текста, так и его лексико-семантическое содержание.

6. Предложена методика сегментации текста, отличающаяся от известных использованием семантического критерия, что позволяет извлекать значимые текстовые фрагменты, сохраняющие смысл текста.

7. Проведена оценка алгоритмов, реализующих предложенный подход, которая показала, что повышается стабильность распознавания плагиата вне зависимости от процента замены слов и по сравнению с существующими системами на 40% увеличивается обнаружение подобия при замене более 50% слов исходного текста.

Существенной научной новизной в исследовании является введение определения значимого текстового пассажа, при котором сегменты, сравниваемые или обрабатываемые для любой последующей цели, будут содержать полный семантический смысл. Кроме того, после сопоставления сегментов текста применяются критерии глубины и полноты для вычисления подобия. Предлагаемый подход и метод семантического сравнения текстов позволяют оценивать тексты, написанные на естественном языке и определить степень их подобия к эталонным текстам, независимо от используемых слов и синтаксиса. Это является основным отличием по отношению к существующим методам, основанным на точном выявлении слов и/или фраз.

Результаты этого исследования и предлагаемые подход и метод имеют непосредственное применение в приложениях автоматизированного обнаружения плагиата в российской коммерческой организации «Акционерное общество Антиплагиат»; в институте системного анализа Федерального исследовательского центра "Информатика и управление" РАН, для повышения его эффективности, в дистанционном образовании в национальном политехническом экспериментальном университете национальных

вооружённых Боливарианских сил (УНЕФА–UNEFA); в Южном федеральном университете с целью улучшения методов оценки ответов.

Новые исследования, основанные на представленном подходе, могут способствовать разработке новых и инновационных методов и приложений для повышения эффективности автоматической обработки текста на естественном языке, в частности, при сравнении семантически подобных текстов, написанных с использованием другого словаря.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в изданиях ВАК

1. Бермудес С. Х. Г. Подход к созданию модели семантического сравнения текстов // Информатизация и связь. Москва, 2016, № 2 – с. 121-126.

2. Бермудес С. Х. Г. О методе извлечения значимых текстовых пассажей как базы для текстового сравнения // Информатизация и связь. Москва, 2016, № 3 – с. 213-219.

3. Бермудес С. Х. Г. О методе определения текстовой близости, основанном на семантических классах // Инженерный вестник Дона – 2016. – № 4. URL: <http://ivdon.ru/ru/magazine/archive/n4y2016/3832>

4. Бермудес С. Х. Г. Метод измерения семантического сходства текстовых документов // Известия ЮФУ. Технические науки. Тематический выпуск, «Информационно-измерительные системы». 2017, № 3 (188) – с. 17-29. URL: <http://izv-tn.tti.sfedu.ru/?p=23527>

Публикации в других изданиях

5. Фролов С. Н., Бермудес С. Х. Г. Обработка естественного языка в дистанционном образовании // Известия Юго-западного государственного университета. Серия Управление, вычислительная техника, информатика. 2015, № 2 (15) – с. 21-26. URL: https://www.swsu.ru/izvestiya/seriesivt/archiv/2_2015.pdf

6. Бермудес С. Х. Г. Natural language processing in distance education // Сборник трудов Международной конференции «Инновационные технологии и дидактика в обучении» (Innovative Technologies and Didactics in Teaching – ITDT 2015), Мадрид. 2015 – с. 179-185.

7. Бермудес С. Х. Г. Подход к созданию семантической системы резюмирования текстов // Сборник трудов XIII-ой Всероссийской Научной конференции молодых ученых аспирантов и студентов «Информационные технологии, системный анализ и управление» (ИТСАиУ – 2015), Таганрог. 2015 – с. 237-242.

8. Бермудес С. Х. Г. Approach of creation of semantic text comparison model // Сборник трудов Международной конференции «Инновационные технологии и дидактика в обучении» (Innovative Technologies and Didactics in Teaching – ITDT 2016), Тенерифе. 2016 – с. 157-161.

9. Бермудес С. Х. Г. Модель семантического сравнения текстов // Сборник трудов VII-ой Международной научно-технической конференции «Технологии разработки информационных систем» (ТРИС – 2016), Геленджик. 2016 – с. 160-165.

10. Бермудес С. Х. Г. Обзор методов и моделей разрешения анафор // Сборник трудов XIV-ой Всероссийской Научной конференции молодых ученых аспирантов и студентов «Информационные технологии, системный анализ и управление» (ИТСАиУ – 2016), Таганрог. 2016 – с. 226-233.

11. Бермудес С. Х. Г. Algorithm of segmentation of texts on meaningful passages // Сборник трудов Международной конференции «Инновационные технологии и дидактика в обучении» (Innovative Technologies and Didactics in Teaching – ITDT 2017), Берлин. 2017 – с. 81-84.

12. Бермудес С. Х. Г. Разработка метода и алгоритмов семантического сравнения текстовых документов // Сборник трудов VIII-ой Международной научно-технической конференции «Технологии разработки информационных систем» (ТРИС – 2017), Геленджик. 2017 – с. 154-174.

13. Bermúdez S. J. Sobre un modelo de comparación semántica de documentos textuales // “Ciencia e Ingeniería”. 2017. Vol. 38, Num. 3-2017. сс. 291-300. URL:[erevistas.saber.ula.ve/index.php/cienciaeingenieria/article/download/10037/9964](http://revistas.saber.ula.ve/index.php/cienciaeingenieria/article/download/10037/9964)

14. Bermúdez S. J. Approbation and results of experimental studies of the method of texts segmentation and semantic comparison // Сборник трудов Международной конференции «Информационные инновационные технологии» (Information innovative technologies – I2T 2018), – Прага, 2018. сс. 283-289.