ИЗВЕСТИЯ РАН. ТЕОРИЯ И СИСТЕМЫ УПРАВЛЕНИЯ, 2024, № 4, с. 40–64

КОМПЬЮТЕРНЫЕ МЕТОДЫ

УДК 519.769

МЕТОДЫ РЕШЕНИЯ ЗАДАЧИ ТЕМАТИЧЕСКОЙ СЕГМЕНТАЦИИ ТЕКСТОВ НА ОСНОВЕ ГРАФОВ ЗНАНИЙ

© 2024 г. З. К. Авдеева[a, *], М. С. Гаврилов[a, b, **], Д. В. Лемтюжникова[a, *], А. Ф. Шарафиев[a, ****]**

[a]*Институт проблем управления им. В. А. Трапезникова РАН, Москва, Россия*

[b]*МАИ (национальный исследовательский ун-т), Москва, Россия*

**e-mail: avdeeva@ipu.ru*

***e-mail: cobraj@yandex.ru*

****e-mail: darabbt@gmail.com*

*****e-mail: whiskeydudev@gmail.com*

Поступила в редакцию 14.04.2024 г.

После доработки 12.05.2024 г.

Принята к публикации 15.07.2024 г.

Тематическая сегментация – это задача разделения неструктурированного текста на тематически связные сегменты (такие, в которых речь идет об одном и том же). Граф знаний – графовая структура, вершинами которой являются различные объекты, а ребрами – отношения между ними. Как задача тематической сегментации, так и задача автоматического построения графа знаний не будут новыми, поэтому существует множество алгоритмов для их решения. Однако методы решения задачи тематической сегментации с помощью графов знаний до сих пор исследованы мало. Более того, пока еще нельзя сказать, что задача тематической сегментации решена в общем виде, т.е.существуют алгоритмы, способные при должной настройке решить задачу с требуемым качеством на конкретном наборе данных. Предлагается новый метод решения задачи тематической сегментации на основе графов знаний. Применение графов знаний при сегментации позволяет использовать больше информации о словах в тексте: помимо того чтобы основываться на co-occurrance и семантических расстояниях (как классические алгоритмы), методы на базе графов знаний могут применять расстояние между словами на графе, инкорпорируя тем самым фактологическую информацию из графа знаний в процесс принятия решений о биении текста на сегменты.

Ключевые слова: тематическая сегментация, граф знаний, обработка естественного языка.

DOI: 10.31857/S0002338824040031 **EDN:** UENRQR

METHODS FOR SOLVING THE PROBLEM OF TOPIC SEGMENTATION OF TEXTS BASED ON KNOWLEDGE GRAPHS

Z. K. Avdeeva[a, *], M. S. Gavrilov[a, b, **], D. V. Lemtyuzhnikova[a, *], A. F. Sharafiev[a, ****]**

[a]*V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, Moscow, Russia*

[b]*Moscow Aviation Institute (National Research University), Moscow, Russia*

**e-mail: avdeeva@ipu.ru*

***e-mail: cobraj@yandex.ru*

****e-mail: darabbt@gmail.com*

*****e-mail: whiskeydudev@gmail.com*

Topic segmentation is the task of dividing unstructured text into thematically connected segments (such as those dealing with the same matter). The knowledge graph is a graph structure, the vertices of which are various objects, and the edges are the relationships between them. Both the task of topic segmentation and the task of automatically constructing a knowledge graph are not new, therefore there are many algorithms for solving them. However, methods for solving the problem of topic segmentation using knowledge graphs have so far been little studied. Moreover, yet it cannot be said that the problem of topic segmentation has been solved in a general way, that is, there are algorithms that, if properly configured, can solve the problem

with the required quality on a specific data set. In this paper, a new method for solving the problem of topic segmentation based on knowledge graphs is proposed. The use of knowledge graphs in segmentation allows us to use more information about words in the text: in addition to being based on co-occurrence and semantic distances (like classical algorithms), knowledge graph-based methods can apply the distance between words on the graph, thereby incorporating factual information from the knowledge graph into the decision-making process of partitioning the text into segments. In this paper, we propose a method for solving the problem of topic segmentation based on knowledge graphs.

Keywords: topic segmentation, knowledge graph, natural language processing.

Введение. В современном мире человеку приходится воспринимать и обрабатывать огромные потоки информации разной природы, основными из которых являются визуальная, слуховая и текстовая. Несмотря на то, что наиболее естественной является именно визуальная информация, людям нужно взаимодействовать с текстовыми данными: рабочими отчетами, литературными произведениями, постами в социальных сетях, субтитрами в фильмах и т.д.

В общем случае, поступающие человеку текстовые данные являются слабоструктурированными. В некоторых случаях, как, например, при чтении на электронном носителе скачанной из интернета книги, это не причиняет особого дискомфорта – достаточно того, что книга разделена на главы. Однако во многих ситуациях отсутствие видимой структуры в тексте затрудняет восприятие информации. Когда необходимо отыскать какую-то конкретную публикацию среди множества источников, поиск сильно усложняется, если это множество плохо структурировано.

Задача структурирования текстов возникает в большом количестве приложений, и для ее решения применяются различные подходы. Одной из ее устойчивых вариаций является задача тематической сегментации. Неформально она формулируется как отыскание разбиения документа на непересекающиеся последовательные подмножества таким образом, чтобы каждое подмножество характеризовалось высокой тематической связностью, другими словами, чтобы внутри каждого сегмента текст был об одном и том же.

Задача тематической сегментации имеет большое значение в области обработки естественного языка. Результаты решения данной задачи представляют интерес для использования в таких задачах, как суммаризация, моделирование диалога, в вопросно-ответных системах и др.

В статье предлагается формальная постановка задачи тематической сегментации в терминах графов, рассматриваются несколько алгоритмов для решения задачи тематической сегментации, исследуется качество их работы на научных статьях.

1. Задачи и алгоритмы тематической сегментации текстов и автоматического построения графов.

1.1. Обзор моделей, задач и алгоритмов. В общем виде предлагаемые методы решения задачи тематической сегментации текста представляют собой комбинацию из графа знаний, в которой отображаются предложения документа, и алгоритма поиска оптимальной группировки подграфов, соответствующих предложениям. Существует большой пласт методов и алгоритмов, которые решают задачу автоматического построения графа знаний. Приведем примеры некоторых работ, в которых они описываются.

В [1] граф знаний строится на основе синтаксического дерева. Из данного дерева извлекаются именные группы, соединенные заданными типами связи. Связи играют роль ребер. Помимо извлечения таких наборов, для каждого ребра рассчитывается вес на базе относительного расстояния между двумя вершинами в тексте, семантической схожести и энтропии. В [2] рассматривается метод построения онтологий на основе поиска семантических шаблонов. Семантический шаблон – это выражение на естественном языке, у которого есть смысл. Утверждается, что существующие методы построения онтологий ограничены поиском таксономических отношений, а нетаксономические отношения не рассматриваются. Предлагаемый в работе метод на базе семантических шаблонов призван помочь в решении этой проблемы. В [3] предлагаемый алгоритм осуществляет поиск и извлечение отношений с помощью структуры английских предложений. Для пар именованных фраз из текстов алгоритм извлекает фразы-отношения, удовлетворяющие определенным условиям: фраза не должна быть слишком длинной и должна удовлетворять заданным шаблонам (например, глагол-предлог). В [4] граф знаний строится следующим образом. Для заданного набора сущностей из сети Интернета набирается объем документов, которые содержат определенное число этих сущностей. Затем текстовые вхождения в документы, содержащие сущности из набора, подаются на вход

заранее обученной нейронной сети, которая, решая задачу классификации, предсказывает отношение между сущностями.

Существуют разные подходы сегментации, приведем четыре характерных на основе: лексического анализа, тематического моделирования, нейронных сетей и графа знаний. Сегментация на основе лексического состава заключается в сравнении употребляемого словаря в соседних блоках текста. К этому типу подходов относятся работы [5, 6]. В [5] документ токенизируется и разбивается на блоки с заданным числом токенов. Затем в соответствии с выбранной стратегией сравнения соседних блоков принимается решение о наличии или отсутствии тематической границы между ними. В статье рассматриваются три стратегии сравнения соседних блоков на основе: наличия одинаковых слов в обоих блоках, числа новых слов в последующем блоке относительно предыдущего и так называемых лексических цепочек, т.е. последовательностей предложений, в которых появляется то или иное слово. В [6] развивается подход публикации [5] с лексическими цепочками. В этой работе к ним добавляются определенным образом подсчитанные веса, в то время как в [5] просто рассматривалось, проходит или нет та или иная лексическая цепочка через блок. Суть второго направления заключается в интеграции задачи сегментации в задачу тематического моделирования текста. Здесь можно выделить два направления работ. В первом используется уже обученная тематическая модель, с помощью которой рассчитываются тематические векторные представления единиц документа. Во втором сегментирование документа встраивается в генеративный процесс документа. К этому типу подходов относятся работы [7, 8]. Так в [7] применяют обученную модель латентное размещение Дирихле (latent dirichlet allocation (LDA)), алгоритм для получения тематического вектора ранее неизвестного модели документа и динамическое программирование. Задача решается следующим образом. Принимается, что если сегмент является тематически более связным в сравнении с другим сегментом, то его функция правдоподобия должна быть больше. Таким образом, документ разделяется на всевозможные сегменты, для каждого из которых с помощью модели LDA выводится свой собственный тематический вектор. Затем с помощью динамического программирования находится та сегментация, которая максимизирует правдоподобие всего документа. В [8] рассматривается фиктивный генеративный процесс создания документов, как это делается, например, в модели LDA. Однако явным образом добавляется эвристика о том, что документ состоит из нескольких сегментов, каждый со своим тематическим распределением. Тематическое распределение для сегментов порождается с помощью процесса Дирихле-Пуассона и общим тематическим вектором для всего документа. Третий подход основывается на применении графа знаний. К этому типу подходов относится работа [9]. Представленный в [9] метод использует для решения задачи сегментации граф знаний. Непосредственно задача сегментации решается для длинных видеолекций, состоящих из слайдов, однако при этом применяется только текстовая информация, так что обобщить предложенный метод на только текстовые данные не составляет большого труда. Идея метода заключается в том, чтобы отобразить информацию, соответствующую каждому слайду, в подграф графа знаний. Данный граф знаний определенным образом строится на самом корпусе текстовой информации. Затем сравнивая, как подграфы соседних слайдов располагаются друг относительно друга, принимается решение о наличии или отсутствии тематической границы. Четвертый подход заключается в различных вариантах применения глубоких нейронных сетей. К этому типу подходов относятся работы [10–12]. Например, в [10] используется последовательно соединенные двунаправленные нейронные сети-кодировщики (bidirectional encoder representations from transformers (BERT)) и модель двунаправленной долгой краткосрочной памяти (bidirectional long short-term memory (Bi-LSTM)). Выходы последней пропускаются через блок пропускного подключения (skip-connection) и специальный блок многоцелевого внимания (multihop-attention), который рекуррентно реализует механизм внимания. Задача решается как классификация наличия в предложении тематического сдвига. В [11] применяются два последовательных трансформера. Первый трансформер используется для кодирования каждого предложения. Вектор предложения конкатенируется из двух частей – вектор прямого кодирования предложения и вектор, полученный кодированием конкатенации данного предложения с соседним. Совокупный вектор далее применяется в качестве входа следующего трансформера, чьи выходы уже используются для предсказания темы предложения и наличия в нем тематического перехода. В [12] для решения задачи сегментации был собран специальный датасет, состоящий из текстов Википедии, разбитых на секции. Каждая секция характеризовалась своим заголовком, а также меткой класса, проставленной в соответствии с заголовком. Архитектура модели представляла собой Bi-LSTM-кодировщик, который отображал каждое предложение текста в латентное представление. На полученных

векторах затем обучалось два классификатора: первый обучался предсказывать тему для предложения, второй – предсказывать слова, входящие в заголовок.

Приведенный обзор работ показывает, что направление решения задачи тематической сегментации с применением графов знаний исследовано на недостаточном уровне.

1.2. Постановка задачи. Неформально задача тематической сегментации текста звучит так: пусть есть некоторый документ, который необходимо разбить на части таким образом, чтобы каждая часть была тематически связной, т.е. в ней говорилось примерно об одном и том же. С формальной точки зрения на эту задачу можно смотреть с двух сторон. С одной стороны, искомой сегментацией можно считать разбиение документа, оптимальное по какому-либо критерию. В таком случае, "оптимальную" сегментацию необходимо искать среди множества всех возможных разбиений документа. Такую задачу сегментации можно назвать тематической сегментацией документа с глобальной оптимизацией. С другой стороны, практикуется иной подход, который заключается в сравнении примыкающих предложений (или блоков предложений), и на его основе принимается решение о наличии/отсутствии тематической границы между этими предложениями (блоками предложений). Другими словами, группировка предложений осуществляется с помощью локального сравнения. Такую задачу сегментации можно назвать тематической сегментацией документа с локальной оптимизацией. Мы предлагаем следующую формальную постановку для задачи глобальной тематической сегментации.

2. Типы задачи тематической сегментации.

2.1. Задача глобальной тематической сегментации. Формально эту задачу в терминах графов можно поставить следующим образом. Пусть есть некоторый документ D, состоящий из предложений d_i, длина документа в предложениях равна T. Считая, что существует заранее построенный граф знаний, поставим в соответствие каждому предложению подграф g_i из этого графа знаний следующим образом. Найдем в предложении термины, содержащиеся в графе знаний, и в качестве подграфа g_i возьмем подграф, порожденный множеством соответствующих этим терминам вершин.

Получим кортеж, состоящий из подграфов, обозначим его $S_1 = (g_1, g_2, ..., g_T)$. Применяя к этому кортежу операцию объединения подряд идущих подграфов в различных комбинациях, получим множество кортежей, элементами которых являются или сами подграфы – gi, или объединение подряд идущих подграфов, или и то и другое. Примеры таких кортежей – $S_2 = ((g_1, g_2), g_3, g_4, ..., g_T)$; $S_3 = (g_1, (g_2, g_3), g_4, ..., g_T)$; $S_i = ((g_1, ..., g_r), (g_{r+1}, ..., g_l), ..., (g_m, ..., g_T))$. Обозначим это множество $P = \{S_1, S_2, ..., S_n\}$. Полагая, что на множестве P задан некоторый функционал качества $L = L(S_i)$, будем называть оптимальной сегментацией кортеж S_i, на котором функция L принимает оптимальное значение (в частности, минимум).

Поскольку в данной постановке задачи рассматриваются всевозможные разбиения документа D, обозначим такую задачу сегментации текста как тематическая сегментация с глобальной оптимизацией.

2.2. Задача локальной тематической сегментации. Помимо этого, в литературе развивается другое направление методов, которые не подпадают под указанное определение. В этих методах решение о наличии или отсутствии тематической границы между предложениями принимается на основе локального сравнения двух соседних предложений (или в некоторых методах блоков предложений).

Формально такую задачу сегментации можно представить следующим образом. Как и в случае глобальной сегментации, документ, состоящий из предложений, отображается в кортеж подграфов общего графа знаний. Таким образом, пусть есть некоторый кортеж $S = (g_1, g_2, ..., g_T)$, соответствующий некоторому документу D. Считая, что существует некоторая функция $s(g_l, g_r)$, вычисляющая схожесть между подграфами g_l и g_r, будем относить к одному сегменту те подграфы, для которых $s(g_l, g_r) > TH$, где TH – некоторое задаваемое значение.

3. Графы знаний. Граф знаний (knowledge graph) – это графовая структура, в которой хранится информация о разных сущностях и взаимосвязях между ними, вершины в графе знаний суть некоторые концепции или понятия, а ребра – связи между концепциями. В работе используются три вида графов знаний: на основе тезаурусов по теории управления, классификатора и синтаксического графа. В контексте исследуемых методов графы знаний рассматриваются в качестве инструмента получения представлений для предложений. В табл. 1–3 представлен внешний вид тезаурусов и классификатора.

3.1. Описание используемых графов знаний. Граф знаний на основе тезаурусов по теории управления строится по табл. 1, 2 определений, которые представляют

Таблица 1. Представление тезауруса 1988 г.

Наименование термина/уровня	Английский термин	Определение	Примечание
Основные понятия			
Объект управления	Controlledobject; controllable object	Объект, для достижения желаемых результатов функционирования которого необходимы и допустимы специально организованные воздействия	1. Объект управления, подвергаемый управляющим воздействиям, можно называть управляемым объектом. 2. Объектами управления могут быть как отдельные объекты, выделенные по определенным признакам (например, конструктивным, функциональным), так и совокупности объектов-комплексов. 3. В зависимости от свойств или назначения объектов управления могут быть выделены технические, технологические, экономические, организационные, социальные и другие объекты управления и комплексы
Цель управления	Control aim	Значения, соотношения значений координат процессов в объекте управления или их изменения во времени, при которых обеспечивается достижение желаемых результатов функционирования объекта	

собой пары – термин и определение этого термина через другие термины. Вершинами в данном графе являются термины из табл. 1, 2. Вершины соединяются ребрами, если один термин определяется через другой. Отношение «определен через» в общем смысле не является симметричным (т.е. связь будет направленной), однако на первых этапах от направленности было решено отказаться.

Для построения таких графов использовалось два предметных тезауруса по теории управления, составленных в 1988 г. [13] и 2024 г. [14], которые отличаются набором и количеством вершин. Помимо этого, рассматривался совокупный граф, представляющий собой объединение графов, описанных выше. Графы объединялись таким образом, что непересекающиеся вершины, т.е. вершины, которые принадлежат только одному графу, добавлялись вместе со всеми связями между данными непересекающимися вершинами. После этого добавлялись пересекающиеся вершины и соответствующие связи с непересекающимися вершинами.

Граф классификатора строится на основе онтологии по теории управления, собранной экспертами. Классификатор представляет собой лес, содержащий три дерева. Каждое дерево имеет четыре уровня, отражающих общность находящихся на этом уровне терминов (т.е. чем ниже находится термин, тем он более конкретный). Пример последовательности терминов в порядке увеличения глубины: математический аппарат, алгебра и теория чисел, теория чисел, простое число. Граф знаний на базе классификатора строился следующим образом. В качестве вершин в графе брались вершины онтологии. Две вершины соединяются ребром, если они принадлежали одному дереву.

Синтаксический граф строится с помощью синтаксического дерева разбора предложений в обучающей выборке. Для каждого предложения из корпуса текстов выполняется синтаксический разбор. Из полученного синтаксического дерева извлекаются имена, глаголы и синтаксические связи между ними. Извлеченные связи и слова записываются в граф. Также

Таблица 2. Представление тезауруса 2024 г.

Термин	Английский термин	Определение
Абдукция	Abduction	Вид рассуждения, использующий абдуктивный вывод, т.е. вывод от следствия к причине. Правила абдуктивного вывода имеют следующий вид: из *A* следует *B*; *B* имеет место; следовательно, причиной *B* будет *A*. Поскольку причин явления *B* может быть много, заключение абдуктивного вывода представляется всего лишь гипотезой, а сам вывод – правдоподобным выводом. Поэтому абдуктивные выводы называют порождением гипотез
Абсолютная устойчивость	Absolute stability	Свойство нелинейного объекта сохранять асимптотическую устойчивость в целом для любых значений параметров нелинейной характеристики объекта из заданного класса нелинейных характеристик
Абстрагирование	Abstracting, abstraction	Процесс формирования образов реальности (представлений, понятий, суждений) посредством отвлечения и пополнения, т.е. путем использования (или усвоения),– лишь части из множества соответствующих данных и прибавления к этой части новой информации, не вытекающей из этих данных

в граф добавляется информация о местонахождении слов в корпусе (номер текста и номер слова) и число раз, которые эти слова и связи встретились.

Расстояние между двумя словами a и b на графе рассчитывается как сумма расстояния по синтаксическим ребрам и ссылочного расстояния. Для определения расстояний по ребрам за вес берется число, обратное количеству раз, которое ребро встретилось в корпусе, а ссылочное расстояние вычисляется по формуле; $1 / D(a,b) + S(a,b) + e(b)$, где $D(a,b)$ – кратчайшее расстояние между словами a и b в тексте, $S(a,b)$ – семантическая схожесть слов a и b, рассчитываемая как разность единицы и косинусного расстояния между векторными представлениями a и b, а $e(b)$ – энтропия Шеннона от слова b. Такой метод расчета расстояния по графу знаний учитывает частоты соупотребления слов, семантическую схожесть и количество информации, содержащейся в целевом слове. Графовое и ссылочное расстояния затем нормируются функцией сигмоиды, и их сумма используется как итоговое расстояние между словами. Итоговое расстояние нормировано в пределах от 0 до 1, и чем оно меньше, тем больше слова похожи друг на друга.

3.2. Характеристики используемых графов знаний. Для исследования взаимосвязей между характеристиками графов и качеством работы предлагаемых методов были подсчитаны следующие характеристики графов: средняя степень вершины; число связных компонент; количество изолированных вершин; коэффициент кластерности; кликовое, цикломатическое и хроматическое числа; степень ассортотивности; ранг матрицы смежности; остовное число. В табл. 4 приводится более подробное описание характеристик графов.

В теории графов коэффициент кластеризации – это мера степени, в которой узлы графа склонны объединяться в группы. Кликовое число графа – число вершин в наибольшей клике графа. Цикломатическое число графа указывает то наименьшее число ребер, которое нужно удалить из данного графа, чтобы получить дерево (для связного графа) или лес (для несвязного графа), т.е. добиться отсутствия у графа циклов. Хроматическое число графа – это минимальное число цветов, которыми можно правильно раскрасить вершины графа. Степень ассортативность – тенденция узлов графа соединяться с другими узлами того же типа.

Используемые в работе графы являются не эйлеровыми, не планарными, нерегулярными и не являются деревьями. Из всех графов только граф на основе классификатора будет хордальным, что следует из способа его построения.

Среди графов на базе тезауруса наибольшую среднюю степень вершины имеет граф на тезаурусе 2024 г., равную 8.852, наименьшую — граф на тезаурусе 1988 г., равную 7.007. Значение средней степени вершины для графа на классификаторе сильно отличается от графов на тезаурусе и равно 862.231. Граф на тезаурусе 1988 г. имеет одну связную компоненту, а граф на тезаурусе 2024 г. – четыре, однако этот граф также имеет три изолированные вершины, у совмещенного графа есть характеристики, как у графа на тезаурусе 2024 г. Исходя из построения графа на классификаторе, он имеет три компоненты связности. Количество вершин

Таблица 3. Представление классификатора

Факторы классификации	Уровень		Термины
	1	2	
Математический аппарат	Алгебра и теория чисел	Лычагин В.В.	–
		Теория чисел	Простое число, совершенное число, взаимно простые числа, алгебраические числа, p-адические числа, группа Галуа, дзета функция, конечные поля
		Линейная алгебра, теория матриц	Линейное пространство, линейная комбинация, векторное пространство, базис, линейный оператор, ранг матрицы, детерминант, обратная матрица
		Алгебраическая геометрия	Базис Гребнера, аффинные многообразия, проективные многообразия, алгебраические множества, бирациональная эквивалентность, спектр кольца, простой идеал
	Геометрия и топология	Лычагин В.В.	–
		Дифференциальная геометрия	Гладкое многообразие, диффеоморфизм, векторное поле, дифференциальная форма, внешний дифференциал, линейная связность, ковариантный дифференциал, кривизна связности, кручение связности
		Топология	Топологическое пространство, компакт, локально компактное пространство, фактор-пространство, нормальное пространство, хаусдорфово пространство, компактизация, связное пространство, непрерывное отображение, гомеоморфизм, гомотопия

в каждой компоненте равно 863, 1042, 417. Граф на классификаторе имеет самое большое значение коэффициента кластерности, равное один. Самое большое значение коэффициента среди графов на тезаурусах имеет граф на тезаурусе 1988 г., равное 0.495, самое маленькое значение у графа на тезаурусе 2024 г., равно 0.213.

Графы на тезаурусах имеют одинаковые значения для кликовых и хроматических чисел, равные пять и семь соответственно. Для графа на классификаторе обе метрики равны 1042. Среди графов на тезаурусах самое большое значение для цикломатического числа у совмещенного графа, будет 4233, а самое маленько – у графа на тезаурусе 1988 г., равное 687. Для графа на классификаторе эта метрика равна 998731. Среди всех графов самое большое значение для степени ассортотивности будет у графа на классификаторе, равное 1. Самое маленькое значение у графа на тезаурусе 1988 г. составляет –0.347. Матрица смежности графа на классификаторе имеет самый высокий ранг, равный 2322, в то время как самый низкий ранг, равный 173, имеет матрица смежности графа на тезаурусе 1988 г. Среди графов на тезаурусах самая большая длина пути (при расчете исключались изолированные вершины) у совмещенного графа (3.029), самая маленькая у графа на основе тезауруса 1988 г. (2.509). Исходя из построения, в графе на базе классификатора значение этой характеристики для каждой компоненты связности равно 1.

Средняя степень вершины в синтаксическом графе равняется 24.771, что больше, чем у всех графов на тезаурусах, но меньше, чем у графа классификатора. Синтаксический граф имеет шесть компонент связности и пять изолированных вершин. Коэффициент кластерности равен 0.209. Кликовое, хроматическое и цикломатическое числа равны 20, 30 и 88870 соответственно. Степень

Таблица 4. Характеристики графов

Характеристика	Год			Классификатор			Синтаксический граф
	2024	1988	1988+2024				
Средняя степень вершины	8.852	7.007	8.618	862.231			24.771
Число связных компонент	4	1	4	3			6
Количество изолированных вершин	3	0	3	0			5
Кластерность	0.213	0.495	0.262	1.0			0.209
Кликовое число	5	5	5	1042			20
Хроматическое число	7	7	7	1042			30
Цикломатическое число	3567	687	4233	998731			88870
Степень ассортотивности	–0.216	–0.347	–0.214	1.0			–0.127
Ранг матрицы смежности	742	173	878	2322			7004
Компонента	1	1	1	1	2	3	1
Размер компоненты	1037	274	1275	863	1042	417	7800
Средняя длина пути	2.991	2.509	3.029	1	1	1	2.864

ассортотивности равна –0.127 и является самой близкой к нулю. Ранг матрицы смежности синтаксического графа равен 7004. По своим характеристикам, если не учитывать степень ассортотивности и ранг матрицы смежности, данный граф находится где-то между графом классификатора и графами на тезаурусах. Средняя длина пути в синтаксическом графе равна 2.864.

4. Методы тематической сегментации текстов. Как уже было сказано, задача тематической сегментации заключается в разбиении документа на тематически связные сегменты. Для решения задачи тематической сегментации текста предлагаются следующие методы: на основе оценки всевозможных сегментаций, на базе функции тематической связности, на основе синтаксического графа и на базе накопительного графа. Рассмотрим каждый из них.

4.1. Методы на основе оценки всевозможных сегментаций. Задача глобальной тематической сегментации решается следующим образом.

Пусть $S_i = ((g_1,...,g_r),(g_{r+1},...,g_l),...,(g_m,...,g_T)) = (G_1,G_2,...,G_h)$ – некоторый кортеж из множества P. Приведем общий вид функции:

$$L(S_i) = \sum_{i=1}^{h-1} L'(G_i,G_{i+1}),$$

где $L'(G_i,G_{i+1})$ – некоторая функция, определенная на парах подряд идущих элементов кортежа S_i. Для функции L' был выбран следующий функциональный вид:

$$L'(G_i,G_{i+1}) = \sum_{n_p \in G_i} \sum_{n_q \in G_{i+1}} F(n_p,n_q),$$

где n_p,n_q – вершины, принадлежащие $G_i,G_{(i+1)}$ соответственно, F – функция взаимодействия между вершинами n_p, n_q. Другими словами, значение функционала L на S_i равно сумме значений некоторой функции L' на парах подряд идущих элементов кортежа. В свою очередь L' равна сумме попарных взаимодействий между вершинами, принадлежащими разным элементам пары. Функция F задает взаимодействие между двумя вершинами и является объектом исследования данной статьи.

Для графов на тезаурусах используется следующая функция взаимодействия:

$$F(n_p,n_q) = \begin{cases} G\dfrac{w_{n_p}w_{n_q}}{r^2}, r < TH, \\ -G\dfrac{w_{n_p}w_{n_q}}{r^\gamma}, r \geq TH, \end{cases}$$

где w_{n_p}, w_{n_q} – частоты терминов, соответствующих вершинам n_p, n_q; r – расстояние между вершинами n_p, n_q; G, γ, TH – гиперпараметры.

Для графов на онтологии применяется следующая функция взаимодействия:

$$F(n_p, n_q) = \begin{cases} G\ w_{n_p} w_{n_q}, Comp(n_p) = Comp(n_q), \\ -G\ w_{n_p} w_{n_q}, Comp(n_p) \neq Comp(n_q), \end{cases}$$

где $Comp(n)$ – функция, возвращающая индекс компоненты связности, в которой лежит вершина n.

Эвристика здесь следующая. Если вершины находятся достаточно близко друг к другу, то они должны стараться объединить два подграфа в один, в противном случае они должны сопротивляться этому объединению. Поскольку для графа на онтологии понятие расстояния теряет какой-либо смысл, в соответствующих формулах этот множитель просто убирается.

4.2. Методы на основе расчета весов промежутков. Рассматриваются несколько типов алгоритмов решения задачи локальной тематической сегментации: на базе функции тематической связности, на основе синтаксического графа и накопительного графа. За основу всех предлагаемых методов берется классический алгоритм TextTiling [5]. Напомним, что идея этого алгоритма заключается в том, что текст разбивается на блоки фиксированной длины, после чего с помощью сравнения тем или иным образом соседних блоков (или групп блоков) принимается решение о наличии/отсутствии тематической границы между ними. При этом в статье предлагаются различные способы того, как можно сравнивать соседние блоки. Представленные алгоритмы берут за основу идею сравнения соседних блоков. Непосредственно в публикации [5] блоки состоят из фиксированного количества слов, что может идти вразрез с синтаксической структурой текста (границы блоков проходят внутри предложений). Данные методы опираются на сравнение предложений как минимальных единиц документа. Для локальной оптимизации справедлива та же эвристика, что и для глобальной оптимизации с той лишь разницей, что в этом случае рассматривается только кортеж, получаемый непосредственно после отображения предложений в подграфы.

4.2.1. Метод на основе функции тематической связности. В качестве метрики схожести соседних графов была выбрана следующая функция, которую будем называть функцией тематической связности (topic coherence):

$$\Delta L^i = \frac{1}{2}(\Delta L(g_i, g_{i+1}) + \Delta L(g_{i+1}, g_i)),$$

$$\Delta L(g_m, g_c) = \sum_{v \in g_c} \Delta L(g_m, g_c, v),$$

$$\Delta L(g_m, g_c, v) = \begin{cases} w_v, v \in N(g_m), \\ -w_v, v \notin N(g_m), \end{cases}$$

где ΔL^i – оценка прироста тематической связности при объединении подграфов g_i, g_{i+1}; g_i, g_{i+1} – подграфы предложений d_i, d_{i+1} соответственно; g_m, g_c – основной и дополняющий подграфы; v – вершина, принадлежащая g_c; w_v – частота термина, соответствующего вершине v в соответствующем предложении.

Рассматривая два соседних графа, сначала один из них выбирается в качестве основного, а второй – в качестве дополняющего. Если очередная вершина лежит в единичной окрестности основного графа, то к метрике схожести добавляется частота этой вершины, так как считается, что в дополняющем графе говорится про то же, что и в основном. Если очередная вершина не лежит в единичной окрестности основного графа, то ее частота вычитается из итогового значения. После прохода по всем вершинам дополняющего графа, графы меняются ролями и аналогичный подсчет повторяется для другого графа. В качестве итогового значения метрики используется среднее значений для двух случаев.

4.2.2. Метод на основе синтаксического графа. Для расчета графовых представлений из предложения извлекаются слова, соответствующие вершинам заранее построенного синтаксического графа, затем по графу находятся расстояния между словами, входящими в графы соседних предложений. Расстояние между словом и подграфом A рассчитывается как минимальное из всех расстояний между этим словом и вершинами из подграфа A. Расстояние от подграфа A до подграфа B вычисляется как среднее всех расстояний между вершинами подграфа A и графом B. Вес промежутка между двумя предложениями рассчитывается как среднее значение расстояний от A до B и от B до A.

Когда расчет промежутков завершен, выполняется сегментация, пороговое значение вычисляется как сумма среднего значения всех промежутков и их среднеквадратичного отклонения. Все промежутки, значение которых больше, чем пороговое значение, считаются разрывами сегментов.

4.2.3. Метод на основе накопительного графа. В данном методе вместо расчета схожести между соседними предложениями, вычисляется схожесть предложения-кандидата с накопленным сегментом. Это позволяет лучше учитывать контекст и аккумулировать информацию о рассматриваемом сегменте.

В этом методе графовое представление предложения считается как структура связей между именованными группами в предложении. Графовые представления предложений рассчитываются с помощью алгоритма 1.

Алгоритм 1. Расчет графовых представлений предложения.

Вход: предложение.

Выход: группа графовое представление предложения.

Шаг 1. Выполним синтаксический разбор предложения с помощью синтаксического парсера.

Шаг 2. Извлечем из синтаксического дерева имена существительные.

Шаг 3. Между соответствующими вершинами, имеющими прямую связь в дереве разбора, добавим ребра.

Шаг 4. Если в предложении два имени связаны через глагол, также запишем эту связь в итоговый граф.

Шаг 5. Дополним граф словами, имеющими в данном предложении наибольший вес, согласно мере tf-idf.

Для решения задачи сегментации используется алгоритм 2.

Алгоритм 2. Алгоритм решения задачи тематической сегментации текста с помощью накопительного графа.

Вход: текст, разбитый на предложения.

Выход: группы из предложений.

Шаг 1. Рассчитаем для каждого предложения графовое представление с помощью алгоритма 1.

Шаг 2. Создадим граф сегмента, представляющий собой графовое представление первого предложения. Установим пороговое значение, равным 0.

Шаг 3. Берем первое доступное графовое представление предложения из списка предложений, пока они есть.

Шаг 3.1. При рассмотрении данного предложения находим схожесть между сегментом и предложением по формуле N_i/N_a, где N_i – количество вершин графового представления приведенного предложения, входящих в граф сегмента, а N_a – это количество всех вершин графового представления этого предложения.

Шаг 3.2. Если величина схожести больше порогового значения.

Шаг 3.2.1. Производим слияние сегмента и предложения.

Шаг 3.2.2. Записываем величину схожести в список истории.

Шаг 3.2.3. Записываем индекс предложения в список индексов предложений сегмента.

Шаг 3.2.4. Помечаем это предложение как недоступное.

Шаг 3.2.5. Рассчитываем новое пороговое значение как сумму среднего и среднеквадратичного отклонения истории.

Шаг 3.3. Если величина схожести меньше либо равна пороговому значению.

Шаг 3.3.1. Считаем величину схожести сегмента с последующим предложением.

Шаг 3.3.2. Если величина схожести больше порогового значения.

Шаг 3.3.2.1. Добавляем оба предложения в сегмент.

Шаг 3.3.2.2. Добавляем оба значения схожести в историю.

Шаг 3.3.2.3. Записываем индексы предложений в список индексов предложений сегмента.

Ш а г 3.3.2.4. Помечаем эти предложения как недоступные.

Ш а г 3.3.3. Если величина схожести меньше либо равна пороговому значению.

Ш а г 3.3.3.1. Добавляем накопленные индексы предложений сегмента в список сегментов.

Ш а г 3.3.3.2. Очищаем историю величин схожести и очищаем сегмент.

Ш а г 4. Переходим к шагу 3.

Ш а г 5. Возвращаем группы предложений в соответствии со списком сегментов.

Результатом работы этого алгоритма является сегментация s, представляющая собой последовательность 0 и 1 длиной $n - 1$, где n – это количество предложений в тексте. Единице соответствует разрыв сегмента, нулю — его отсутствие.

Также рассматривалось предположение, что учет длины приведенного сегмента может положительно сказаться на качестве сегментации. Для его проверки была проведена следующая модификация данного метода.

Вместо статистической меры по истории промежутков пороговое значение вычисляется по формуле $2 / (1 + e^{-l/c}) - 1$, где l – длина сегмента, а c – гиперпараметр, определяющий скорость роста порогового значения с длиной. Эвристика данной формулы заключается в том, что в естественном тексте тематические сегменты не могут совпадать со всем текстом (если он имеет большой объем предложений), а значит, существует некоторое граничное (неизвестное нам) значение для распределения истинных размеров сегментов. Таким образом, чем больше сегмент имеет длину, тем более вероятно, что он будет иметь длину выше пороговой, и тем менее вероятно, что к данному сегменту можно добавить еще предложения.

5. Эксперименты.

5.1. О п и с а н и е д а н н ы х. Данные представляют собой статьи сотрудников научной организации в формате pdf, из которых выбрали 6000 русскоязычных статей. Из них было отобрано 65 статей на 27 тем. Каждую статью вручную разделили на предложения. Затем из этих статей были собраны новые статьи следующим образом. Сначала все предложения, относящиеся к одной теме, объединили в группу. При генерации нового документа сначала случайным образом выбиралась тема, в которой есть свободные предложения, затем из этой группы извлекалось случайное количество предложений, которые добавлялись в генерируемый документ. Этот процесс продолжался до тех пор, пока генерируемый документ не стал содержать заданного объема предложений. Документы генерировались, пока были не пустые темы. Таким образом создали 63 статьи, состоящие минимум из 70 предложений. В табл. 5, 6 представлена статистика синтезированного датасета по темам и описание тем и их кодировок.

Следовательно, для создания тестового датасета было отобрано 65 статей. Оставшиеся 5935 статей из начального набора использовали для построения синтаксического графа.

В качестве базовых методов, с которыми проводилось сравнение предлагаемых алгоритмов, рассматривались алгоритмы локальной оптимизации, применяющие модели векторного представления текста. В качестве таких алгоритмов выбрали широко известный алгоритм частота терминов по обратной частоте документов (term frequency inverse document frequency (TF-IDF)) [15] и языковую модель BERTSciRus (science russian). С их помощью для каждого предложения рассчитывались векторные представления, которые затем использовались для вычисления косинусной близости соседних предложений.

5.2. П о к а з а т е л и к а ч е с т в а. Для оценки качества работы алгоритмов сегментации применяются два основных показателя – вероятностная ошибка (Pk) [16] и разность сегментов (window difference (WD)) [17].

Pk вычисляется с использованием скользящего окна. При движении окна по документу алгоритм определяет, принадлежат ли предложения на концах этого окна одному или разным сегментам в эталонной сегментации. Если эта принадлежность не соответствует предсказанной сегментации, тогда счетчик неправильных окон увеличивается на один. Итоговое количество неправильных окон нормируется на число всевозможных окон в документе. Размер скользящего окна обычно устанавливается равным половине средней длины эталонного сегмента.

Показатель WD работает схожим образом, однако он оценивает несоответствие в эталонной и предсказанной сегментациях в количестве границ в пределах окна. Другими словами, если в эталонной сегментации некоторое окно содержит N границ, а предсказанная сегментация – M и $M \neq N$, то счетчик числа неправильных окон увеличивается на один.

Хронологически показатель WD появился позже показателя Pk и был задуман как более лучшая характеристика, исправляющая ряд проблем второй. По этой причине в качестве ос-

Таблица 5. Статистика датасета по темам

Обозначение темы	**Общее число предложений**	**Общее число блоков**
упр_груп_бпла	375	56
управл_соц_обр	30	4
углеводор	71	13
нейросемант_ис	501	71
упр_соцсеть	338	44
учеба_vr	22	5
множдос_нелин	266	38
имит_мод	129	19
инфобез	57	8
интел_анализ	442	62
онколог_вакцин	490	71
адапт_резв_комп	98	16
времряд_активиы	155	21
эконом_росс	368	51
гис_соц_обр	394	55
упр_косм_роб	335	49
графика_01_т	48	7
интер_эргон	130	20
CRM	91	16
устойч_стабил_колеб	114	17
тепл_ламп	111	16
упр_спин_глобул	82	12
модел_газ	88	14
быстрореаг_произв	196	29
решен	24	4
геосинх_косм_апп	59	9
сист_полет	257	38

новного показателя используется WD, в то время как Pk – как вспомогательный инструмент, например для отладки алгоритмов.

5.3. А н а л и з р е з у л ь т а т о в. Полная таблица с результатами работы алгоритмов представлена в табл. 7. Исходя из анализа этой таблицы можно заключить, что в среднем методы для решения задачи глобальной тематической сегментации показывают одни из самых плохих результатов и в целом работают хуже, чем алгоритмы для решения задачи локальной сегментации. Лучше работает алгоритм, использующий функцию тематической связности. При этом такие алгоритмы работают лучше базовых моделей, если применяются графы знаний на тезаурусах, в противном случае алгоритм проигрывает базовой модели BERT. Из “глобальных” алгоритмов сильно выбивается алгоритм на основе классификатора, который превосходит аналогичные алгоритмы на тезаурусах, а также базовые модели.

Алгоритм, основанный на синтаксическом графе, показывает результаты, сравнимые с результатами работы алгоритмов на функции тематической связности и лучше, чем у базовой модели. И наконец, алгоритм, имеющий самые хорошие метрики качества, – алгоритм на основе накопительного графа.

Таблица 6. Расшифровка тем

Наименование темы	Тема
упр_груп_бпла	Управление группой БПЛА
управл_соц_обр	Управление в сфере социального образования
углеводор	Разведка и добыча углеводородов
нейросемант_ис	Нейросемантические системы
упр_соцсеть	Информационное управление в социальных сетях
учеба_vr	Учебные пособия на основе виртуальной реальности
множдос_нелин	Множества достижимости нелинейных систем
имит_мод	Имитационноемоделирование
инфобез	Информационнаябезопасность
интел_анализ	Интеллектуальныйанализ
онколог_вакцин	Вакцинотерапияонкологии
адапт_резв_комп	Адаптивное резервирование комплексов взаимосвязанных программных модулей
времряд_активы	Прогнозирование временных рядов (активов)
эконом_росс	Экономика России
гис_соц_обр	Геоинформационные системы для социально-образовательной сферы
упр_косм_роб	Управление космическим роботом
графика_01_т	Программный комплекс «графика-01-т»
интер_эргон	Эргономика интерфейсов
CRM	CRM-системы
устойч_стабил_колеб	Устойчивость и стабилизация колебаний
тепл_ламп	Тепловые лампы
упр_спин_глобул	Управление системой спинов
модел_газ	Моделирование сильно нестационарных потоков газа
быстрореаг_произв	Концепция «быстро реагирующего производства»
решен	Системы для оптимизации процесса принятия решений
геосинх_косм_апп	Система геосинхронных низкоорбитальных спутников
сист_полет	Системы управления полетом

Такую ситуацию можно объяснить следующими соображениями. Во-первых, все методы для решения задачи глобальной сегментации опираются только на тезаурусы и классификатор. То же можно сказать и про “локальные” алгоритмы, которые используют функцию тематической связности. Поскольку графы на тезаурусах и классификаторе были собраны вручную, они не отражают всевозможного лексического содержания тем. Другими словами, некоторые предложения отображаются либо в небольшие подграфы (малое число вершин), либо совсем не отображаются в подграфы (число вершин равно нулю). Более того, случается, что предложения, принадлежащие разным темам, могут отображаться или в близкие (минимум попарного расстояния между вершинами подграфов равен единице), или в пересекающиеся подграфы. Видно, что при увеличении числа вершин качество работы алгоритмов, использующих функцию тематической связности, улучшается.

Аномально хорошие результаты “глобального” алгоритма на классификаторе в сравнении с аналогичными алгоритмами на тезаурусах могут свидетельствовать о том, что механизм решения глобальной задачи сегментации хорошо сочленяется с графом классификатора.

Таблица 7. Результаты тестирования алгоритмов

Наименование метода	WD	Pk
Методы на основе перебора всевозможных разбиений		
Тезаурус-1988	0.790	0.432
Тезаурус-2024	0.738	0.612
Тезаурус-1988+2024	0.744	0.617
Классификатор	0.629	0.606
Методы на основе оценки весов промежутков		
Синтаксический граф	0.644	0.383
Базовая модель (BERT)	0.664	0.343
Базовая модель (tfidf)	0.998	0.318
Topic Coherence (Тезаурус-1988)	0.652	0.510
Topic Coherence (Тезаурус-2024)	0.626	0.486
Topic Coherence (Тезаурус-1988+2024)	0.623	0.476
Topic Coherence (классификатор)	0.660	0.528
Topic Coherence (синтаксический граф)	0.668	0.529
Методы на основе накопительного графа		
Без штрафа за длину	0.485	0.404
Со штрафом за длину	0.457	0.376

Методы, применяющие накопительный граф по своей задумке, не используют граф, построенный заранее, вместо этого собирая новые графы во время своего выполнения. Таким образом, концепции (соответствующие вершинам), которые алгоритм извлекает из текста, являются более релевантными самому тексту в сравнении с графами на тезаурусах и классификаторе. Если сравнивать результаты этого подхода с базовыми моделями, можно заметить, что подход значительно выигрывает в качестве. При этом рассматривалось два вида подобного подхода: без штрафования за длину сегмента и с штрафами. Алгоритм, использующий штрафы за длину сегмента, показывает результаты лучше, чем без штрафов, однако для его настройки применяется ряд гиперпараметров, что требует разделять выборку на несколько частей – валидационную и тестовую.

Наконец, алгоритм на основе синтаксического графа, использует заранее построенный в автоматическом режиме граф на базе большого числа релевантных тестовому датасету документов, закладывая через различные расстояния в этот граф информацию о синтаксических и семантических связях. Согласно результатам, такой подход работает чуть лучше по сравнению с моделью BERT. Алгоритм на основе функции тематической связности также показывает не самые хорошие результаты при использовании синтаксического графа. Это может свидетельствовать о том, что алгоритм построения синтаксического графа захватывает слишком много шумовой информации.

Анализируя табл. 5 с метриками графов и результатами, можно сделать следующие выводы. Сравнивая алгоритм локальной сегментации с функцией оценки Topic Coherence для графов на тезаурусах, видно, что увеличение вершин оказывает положительную динамику на результаты сегментации. Однако это справедливо только для графов на тезаурусах, поскольку граф классификатора и синтаксический граф имеют гораздо большее совокупное число вершин и при этом показывают худшие результаты. Помимо этого можно заметить, что данный метод демонстрирует самые лучшие результаты на совмещенном графе тезаурусов 1988 и 2024 гг. Видно при этом, что средняя степень вершины и кластерность не имеют явного влияния на качество сегментации, так как совокупный граф находится по этим метрикам между графами тезаурусов 1988 и 2024 гг. Алгоритм работает тем лучше, чем больше средняя длина пути. Помимо этого видно, что лучшая сегментация имеет самую близкую к нулю степень ассортотивности (среди графов на тезаурусах), что может говорить о том, что для лучшей сегментации не

должно быть тенденции смежности вершин с одинаковыми степенями. Что касается графа на классификаторе, то можно сделать вывод, что его структура в целом не подходит для алгоритма локальной сегментации с функцией тематической сязности. По построению синтаксический граф ближе по структуре к графам на тезаурусах, но в сравнении с ними алгоритм на основе синтаксического графа дает самые плохие результаты. Однако из сравнения характеристик этих графов трудно сделать какой-то вывод. Кластерность синтаксического графа меньше, чем у графов на тезаурусе, но совмещенный граф, дающий лучший результат, имеет промежуточное значение кластерности. У синтаксического графа существует самая большая степень ассортотивности в сравнении с графами на тезаурусах, но степень ассортотивности совмещенного графа ниже, чем у синтаксического графа. На данном этапе подобное поведение списывается на сильную зашумленность синтаксического графа (большое число неинформативных вершин и ребер).

Анализируя связь характеристик с результатами алгоритмов глобальной сегментации, можно увидеть различия в поведении алгоритмов на тезаурусах. Если в случае локальной сегментации лучше всего себя показал совмещенный граф, то в данном случае самые высокие метрики принадлежат алгоритму на основе тезауруса 2024 г. Можно заметить некоторую корреляцию между кластерностью графа и результатами работы, а именно чем ниже кластерность графа такой структуры, тем лучше работает алгоритм. Похожее замечание можно сделать относительно средней степени вершины,т.е. чем она больше, тем лучше результат.

Таблица 8. Статистика по темам метода локальной сегментации на основе функции тематической связности с графом на тезаурусе 1988 г.

topic	inner_mistakes_rate	outer_mistakes_rate
im_top_best		
учеба_vr	0	0
тепл_ламп	0	13.33
быстрореаг_произв	1.23	13.21
углеводор	1.59	12
времряд_активиы	3.1	17.07
im_top_worst		
имит_мод	21.1	13.89
управл_соц_обр	19.23	12.5
нейросемант_ис	18.84	12.98
адапт_резв_комп	18.29	7.14
графика_01_т	14.63	33.33
pr_top_best		
учеба_vr	0	0
геосинх_косм_апп	7.41	0
адапт_резв_комп	18.29	7.14
упр_соцсеть	3.74	7.32
сист_полет	11.42	10.29
pr_top_worst		
решен	10	42.86
графика_01_т	14.63	33.33
упр_косм_роб	13.43	27.47
интел_анализ	9.26	20.69
упр_спин_глобул	5.71	20

Таблица 9. Статистика по темам метода локальной сегментации на основе функции тематической связности с графом на тезаурусе 2024 г.

topic	inner_mistakes_rate	outer_mistakes_rate
im_top_best		
решен	0	57.14
углеводор	0	28
тепл_ламп	0	20
устойч_стабил_колеб	2.06	15.15
онколог_вакцин	2.63	20.63
im_top_worst		
адапт_резв_комп	23.17	28.57
интер_эргон	21.05	18.92
гис_соц_обр	17.4	28.16
имит_мод	16.51	36.11
времряд_активиы	13.95	26.83
pr_top_best		
геосинх_косм_апп	9.26	0
учеба_vr	5.88	0
сист_полет	4.57	10.29
модел_газ	6.76	11.54
управл_соц_обр	3.85	12.5
pr_top_worst		
решен	0	57.14
имит_мод	16.51	36.11
быстрореаг_произв	7.41	32.08
адапт_резв_комп	23.17	28.57
гис_соц_обр	17.4	28.16

В табл. 8–17 приводятся статистики работы алгоритмов по темам. Алгоритмы сегментации могут совершать два вида ошибок. Первый – когда алгоритм ставит границу между предложениями, относящимися к одной теме. Второй – когда алгоритм не разбивает предложения, относящиеся к разным темам. В табл. 8–17 предлагаемые алгоритмы ранжируются по этим ошибкам. В таблицах столбец inner mistakes rate означает количество ошибок первого типа, которые алгоритм допустил для той или иной темы, деленное на общее количество предложений данной темы; столбец outer mistakes rate означает долю тех случаев, когда алгоритм не отделил данную тему от какой-либо другой. Строка im_top_best/im_top_worst показывает начало блоков, содержащих топ пяти тем, на которых алгоритм отработал лучше всего/хуже всего в терминах ошибок первого вида. Строки pr_top_best/pr_top_worst имеют аналогичное значение, но для ошибок второго типа.

Из табл. 8–17 можно увидеть, что алгоритмы, использующие графы тезаурусов 2024 г. и совмещенный, имеют схожий профиль тем по качеству работы. В целом, такое поведение ожидаемо, поскольку граф тезауруса 2024 г. содержит в 10 раз больше вершин, чем граф 1988 г., т.е. при их совмещении основная информация берется из графа тезауруса 2024 г. При этом профили лучших тем по внутренним ошибкам немного отличаются для алгоритма решения задачи глобальной сегментации.

Можно заметить, что для “локального” и “глобального” алгоритмов на классификаторе схожи профили лучших тем по внутренним ошибкам, но отличаются профили лучших тем по внешним ошибкам. Часто можно заметить, что темы “решен” и “учеба_vr” появляются в профилях лучших

Таблица 10. Статистика по темам метода локальной сегментации на основе функции тематической связности с композитным графом на тезаурусах 1988 и 2024 гг.

topic	inner_mistakes_rate	outer_mistakes_rate
im_top_best		
решен	0	57.14
углеводор	0	28
тепл_ламп	0	20
устойч_стабил_колеб	2.06	15.15
управл_соц_обр	3.85	12.5
im_top_worst		
адапт_резв_комп	25.61	25
интер_эргон	21.05	21.62
имит_мод	16.51	36.11
гис_соц_обр	15.93	28.16
времряд_активиы	14.73	31.71
pr_top_best		
геосинх_косм_апп	9.26	0
учеба_vr	5.88	0
модел_газ	6.76	11.54
управл_соц_обр	3.85	12.5
упр_груп_бпла	7.84	12.5
pr_top_worst		
решен	0	57.14
имит_мод	16.51	36.11
быстрореаг_произв	7.41	33.96
времряд_активиы	14.73	31.71
упр_соцсеть	11.22	29.27

тем алгоритмов для обоих типов ошибок. Скорее всего, такое очень частое появление связано с тем, что популяции этих тем (количество предложений) в датасете самые малочисленные, а значит, на этих темах как минимум не получится в принципе сделать больше ошибок, чем для более многочисленных тем.

Из этих же таблиц видно, что “глобальные” алгоритмы хорошо работают на теме «графика_01_т». “Локальные” алгоритмы сегментации на основе тезаурусов в совокупности (объединение по множеству «хороших» темминус множество «плохих» тем обоих типов) хорошо работают на темах 'онколог_вакцин', 'тепл_ламп', 'углеводор', 'устойч_стабил_колеб', 'учеба_vr'по ошибкам первого типа и на темах 'геосинх_косм_апп', 'модел_газ', 'сист_полет', 'упр_груп_бпла', 'учеба_vr' по ошибкам второго типа. «Глобальные» алгоритмы сегментации на тезаурусах хорошо работают с темами 'быстрореаг_произв', 'управл_соц_обр', 'учеба_vr'по ошибкам первого типа и с темами 'быстрореаг_произв', 'интел_анализ', 'управл_соц_обр', 'эконом_росс' по ошибкам второго типа. Методы на базе классификатора хорошо работают на темах 'CRM', 'инфобез', 'устойч_стабил_колеб', 'учеба_vr', 'эконом_росс' по ошибкам первого типа и 'CRM', 'времряд_активиы', 'интел_анализ', 'инфобез', 'множдос_нелин', 'учеба_vr' по ошибкам второго типа.

Хуже всего алгоритмы на тезаурусах работают с темами 'адапт_резв_комп', 'времряд_активиы', 'геосинх_косм_апп', 'гис_соц_обр', 'графика_01_т', 'имит_мод', 'интер_эргон', 'множдос_нелин', 'нейросемант_ис', 'решен', 'тепл_ламп', 'углеводор', 'управл_соц_обр', 'устойч_стабил_колеб' по ошибкам первого типа и 'CRM', 'адапт_резв_комп', 'быстрореаг_произв',

Таблица 11. Статистика по темам метода локальной сегментации на основе функции тематической связности с графом на классификаторе

topic	inner_mistakes_rate	outer_mistakes_rate
im_top_best		
инфобез	0	20
учеба_vr	0	11.11
CRM	1.33	10
упр_спин_глобул	1.43	5
эконом_росс	3.44	10.64
im_top_worst		
упр_косм_роб	19.43	20.88
управл_соц_обр	19.23	12.5
геосинх_косм_апп	16.67	17.65
имит_мод	16.51	16.67
решен	15	14.29
pr_top_best		
времряд_активиы	6.2	4.88
упр_спин_глобул	1.43	5
интел_анализ	13.76	9.48
CRM	1.33	10
множдос_нелин	8.33	10
pr_top_worst		
тепл_ламп	7.37	30
модел_газ	4.05	26.92
быстрореаг_произв	8.02	26.42
графика_01_т	12.2	25
упр_соцсеть	11.22	21.95

'времряд_активиы', 'гис_соц_обр', 'графика_01_т', 'имит_мод', 'интел_анализ', 'интер_эргон', 'инфобез', 'множдос_нелин', 'модел_газ', 'нейросемант_ис', 'решен', 'упр_косм_роб', 'упр_соцсеть', 'упр_спин_глобул', 'устойч_стабил_колеб' по ошибкам второго типа. Алгоритмы на основе графа классификатора хуже всего работают на темах 'геосинх_косм_апп', 'имит_мод', 'решен', 'упр_косм_роб', 'упр_соцсеть', 'управл_соц_обр' по ошибкам первого типа и 'адапт_резв_комп', 'быстрореаг_произв', 'гис_соц_обр', 'графика_01_т', 'модел_газ', 'тепл_ламп', 'упр_соцсеть', 'упр_спин_глобул' по ошибкам второго типа. Эти результаты напрямую свидетельствуют о соответствии используемых графов темам из собранного датасета и будут применены в дальнейшем при рассмотрении/построении новых графов знаний.

Заключение. Исследована модель задачи тематической сегментации текстов в терминах графов знаний. Предложена формализация этой задачи и методы ее решения. Изучена связь характеристик графов знаний с результатами работы алгоритмов. Проведен анализ тем, на которых хорошо работают предлагаемые методы тематической сегментации. Методы представляют собой связку из графа знаний и алгоритма поиска оптимальной сегментации. Для сравнения методов использовались две базовые модели – на основе алгоритма tf-idf и BERT. Алгоритмы на основе накопительных графов, решающие задачу сегментации с локальной оптимизацией, показывают наилучшие результаты, как в сравнении с базовыми моделями, так и с остальными предлагаемыми алгоритмами.

Таблица 12. Статистика по темам метода глобальной сегментации на основе тезауруса 1988 г.

topic	inner_mistakes_rate	outer_mistakes_rate
im_top_best		
быстрореаг_произв	11.73	20.75
времряд_активиы	13.18	31.71
инфобез	14.29	20
тепл_ламп	16.84	26.67
учеба_vr	17.65	22.22
im_top_worst		
углеводор	38.1	32
графика_01_т	31.71	16.67
решен	30	14.29
нейросемант_ис	29.07	34.35
имит_мод	27.52	30.56
pr_top_best		
управл_соц_обр	23.08	12.5
решен	30	14.29
графика_01_т	31.71	16.67
эконом_росс	17.81	19.15
инфобез	14.29	20
pr_top_worst		
устойч_стабил_колеб	24.74	42.42
CRM	18.67	36.67
нейросемант_ис	29.07	34.35
упр_соцсеть	22.11	32.93
интер_эргон	20.18	32.43

Алгоритмы для решения “глобальной” сегментации на базе тезаурусов показали наоборот самые плохие результаты, за исключением алгоритма на основе классификатора. Такое поведение представляет определенный интерес для дальнейшего исследования, поскольку может быть ключом к отысканию как подходящего вида критерия оптимизации, так и структуры графа знаний, для решения “глобальной” задачи сегментации.

Несмотря на то, что алгоритмы, использующие синтаксический граф, показали не самые хорошие результаты, это направление также представляет огромный интерес. Если решить задачу с очисткой графа от шумовой информации, потенциально можно получить хорошие результаты, так как данный граф может содержать в себе гораздо больше информации, чем во всех остальных графах вместе взятых.

Алгоритмы на основе функции тематической связности показали средние результаты. Однако лучше всего они работают с графами на тезаурусах. При этом можно видеть, что при увеличении числа вершин в этих графах и средней длины пути, а также при приближении к нулю степени ассортотивности качество работы алгоритмов растет. Граф на основе классификатора лучше всего работает при решении задачи глобальной сегментации. Графы на тезаурусах лучше работают с функцией тематической связности для решения задачи локальной сегментации. Метод, строящий граф “на лету” работает лучше, чем алгоритмы, использующие вручную построенные графы. Главным итогом данной статьи является вывод о том, что возможно применять графы знаний для решения задачи тематической сегментации.

Таблица 13. Статистика по темам метода глобальной сегментации на основе тезауруса 2024 г.

topic	inner_mistakes_rate	outer_mistakes_rate
im_top_best		
решен	0	28.57
интер_эргон	1.75	5.41
имит_мод	1.83	8.33
быстрореаг_произв	1.85	1.89
упр_соцсеть	2.72	7.32
im_top_worst		
тепл_ламп	12.63	0
времряд_активиы	12.4	14.63
устойч_стабил_колеб	9.28	3.03
геосинх_косм_апп	9.26	5.88
множдос_нелин	9.21	14.29
pr_top_best		
графика_01_т	4.88	0
тепл_ламп	12.63	0
быстрореаг_произв	1.85	1.89
устойч_стабил_колеб	9.28	3.03
интел_анализ	4.23	4.31
pr_top_worst		
решен	0	28.57
модел_газ	6.76	15.38
времряд_активиы	12.4	14.63
множдос_нелин	9.21	14.29
инфобез	4.08	13.33

Дальнейшее направление исследований видится в следующем. Во-первых, необходимо понять, почему «глобальный» метод сегментации с помощью графа классификатора показывает настолько хорошие результаты в сравнении с аналогичными методами на других графах. Во-вторых, как уже было сказано, большой интерес представляет синтаксический граф. Планируется доработать алгоритм его генерации. Например, можно, выделив заранее с помощью другого алгоритма термины из данных, ограничить множество вершин множеством выделенных терминов. В-третьих, необходимо исследовать другие возможные функции взаимодействия между вершинами.

Таблица 14. Статистика по темам метода глобальной сегментации на основе тезаурусов 1988 и 2024 гг.

topic	inner_mistakes_rate	outer_mistakes_rate
im_top_best		
решен	0	28.57
управл_соц_обр	0	12.5
углеводор	1.59	8
имит_мод	1.83	8.33
инфобез	2.04	6.67
im_top_worst		
тепл_ламп	12.63	0
адапт_резв_комп	10.98	17.86
времряд_активиы	10.08	14.63
графика_01_т	9.76	0
геосинх_косм_апп	9.26	5.88
pr_top_best		
графика_01_т	9.76	0
тепл_ламп	12.63	0
быстрореаг_произв	6.79	1.89
интел_анализ	3.44	4.31
геосинх_косм_апп	9.26	5.88
pr_top_worst		
решен	0	28.57
адапт_резв_комп	10.98	17.86
модел_газ	4.05	15.38
упр_спин_глобул	4.29	15
времряд_активиы	10.08	14.63

Таблица 15. Статистика по темам метода глобальной сегментации на основе классификатора

topic	inner_mistakes_rate	outer_mistakes_rate
im_top_best		
инфобез	0	0
учеба_vr	0	0
упр_спин_глобул	0	5
устойч_стабил_колеб	1.03	0
тепл_ламп	1.05	0
im_top_worst		
геосинх_косм_апп	10	0
имит_мод	9.52	0
упр_косм_роб	8.04	0
управл_соц_обр	7.69	0
упр_соцсеть	6.8	0
pr_top_best		
модел_газ	5.41	0
инфобез	0	0
управл_соц_обр	7.69	0
учеба_vr	0	0
геосинх_косм_апп	10	0
pr_top_worst		
упр_спин_глобул	0	5
адапт_резв_комп	1.22	3.57
быстрореаг_произв	6.59	1.89
гис_соц_обр	6.19	0.97
модел_газ	5.41	0

Таблица 16. Статистика по темам метода локальной сегментации на основе синтаксического графа

topic	inner_mistakes_rate	outer_mistakes_rate
im_top_best		
решен	0	0
модел_газ	2.7	15.38
упр_спин_глобул	2.86	20
адапт_резв_комп	4.88	10.71
учеба_vr	5.88	22.22
im_top_worst		
тепл_ламп	29.47	33.33
углеводор	29.31	24
инфобез	24.49	0
геосинх_косм_апп	20	17.65
эконом_росс	18.61	20.21
pr_top_best		
решен	0	0
управл_соц_обр	15.38	0
графика_01_т	12.2	0
инфобез	24.49	0
имит_мод	6.36	8.33
pr_top_worst		
CRM	12	36.67
тепл_ламп	29.47	33.33
углеводор	29.31	24
учеба_vr	5.88	22.22
онколог_вакцин	17.18	20.63

Таблица 17. Статистика по темам метода локальной сегментации на основе накопительного графа

topic	inner_mistakes_rate	outer_mistakes_rate
im_top_best		
учеба_vr	5.88	55.56
решен	10	42.86
имит_мод	11.82	52.78
упр_косм_роб	14.69	46.15
управл_соц_обр	15.38	62.5
im_top_worst		
углеводор	31.03	48
эконом_росс	30.28	48.94
времряд_активиы	27.61	43.9
нейросемант_ис	27.44	48.85
упр_спин_глобул	27.14	45
pr_top_best		
устойч_стабил_колеб	22.68	24.24
инфобез	22.45	40
интер_эргон	20.91	40.54
графика_01_т	17.07	41.67
решен	10	42.86
pr_top_worst		
геосинх_косм_апп	24	64.71
управл_соц_обр	15.38	62.5
быстрореаг_произв	19.76	58.49
учеба_vr	5.88	55.56
онколог_вакцин	16.23	55.56

СПИСОК ЛИТЕРАТУРЫ

1. *Chen H., Luo X.* An Automatic Literature Knowledge Graph and Reasoning Network Modeling Framework Based on Ontology and Natural Language Processing // Advanced Engineering Informatics. 2019. V. 42. https://doi.org/: 10.1016/j.aei.2019.100959
2. *Dahab M., Hassan H.* TextOntoEx: Automatic Ontology Construction from Natural English Text // Expert Systems with Applications. 2008. V. 34(2). P. 1474–1480. https://doi.org/10.1016/j.eswa.2007.01.043
3. *Oren E., Anthony F., Christensen J., Soderland S.* Mausam. Open Information Extraction: The Second Generation // Intern. Joint Conf. on Artificial Intelligence.Barcelona, 2011. https://doi.org/:10.5591/978-1-57735-516-8/IJCAI11-012
4. *Ristoski P., Gentile A.L., Alba A., Gruhl D., Welch S.* Large-scale Relation Extraction from Web Documents and Knowledge Graphs with Human-in-the-loop // J. Web Semantics. 2019. V. 60. https://doi.org/: 100546. doi:10.1016/j.websem.2019.100546
5. *Hearst A.M.* TextTiling: Segmenting Text IntoMulti-paragraph Subtopic Passages // Computational Linguistics. 1997. V. 23(1). P. 33–64.
6. *Galley M., McKeown K., Fosler-Lussier E.* Discourse Segmentation of Multi-Party Conversation // Proc. 41st Annual Meeting on Association for Computational Linguistics (ACL '03). 2003. V. 3. P. 562–569. https://doi.org/:10.3115/1075096.1075167
7. *Misra H., Yvon F., Jose J.M.* Text Segmentation via Topic Modeling: An Analytical Study //Proc. 18th ACM Conf. on Information and Knowledge Management (CIKM '09). Hong Kong, 2009. V. 1. P. 1553–1556. https://doi.org/:10.1145/1645953.1646170
8. *Du L., Buntine W., Jin H.* A Segmented Topic Model Based on the Two-parameter Poisson-Dirichlet Process // Machine Language. 2010. V. 81(2). P. 5–19. https://doi.org/:10.1007/s10994-010-5197-4
9. *Das A., Das P.* Incorporating Domain Knowledge To Improve Topic Segmentation Of Long MOOC Lecture Videos // arXiv:2012.07589 [cs.CL]. https://doi.org/10.48550/arXiv.2012.07589
10. *Nouar F., Belhadef H.* A Deep Neural Network Model with Multihop Self-attention Mechanism for Topic Segmentation of Texts // Innovative Systems for Intelligent Health Informatics. 2021. V. 72. P. 407–417. https://doi.org/:10.1007/978-3-030-70713-2_38
11. *Lo K., Jin Y., Tan W., Liu M., Du L., Buntine W.L.* Transformer over Pre-trained Transformer for Neural Text Segmentation with Enhanced Topic Coherence // Findings of the Association for Computational Linguistics: EMNLP 2021. 2021. V. 1. P. 3334–3340. https://doi.org/:10.18653/v1/2021.findings-emnlp.283
12. *Arnold S., Schneider R., Cudr'e-Mauroux P., Gers F.A.* SECTOR: A Neural Model for Coherent Topic Segmentation and Classification // Transactions of the Association for Computational Linguistics. 2019. V. 7. P. 169–184. https://doi.org/:10.1162/tacl_a_00261
13. Теория управления. Терминология / Под ред. М. М. Гальперина. М.: Наука, 1988. 56 с.
14. Теория управления: словарь системы основных понятий / Под общ. ред. *Д. А. Новикова.* М.: ЛЕНАНД, 2024. 128 с.
15. *Jones K.S.* A Statistical Interpretation of Term Specificity and Its Application in Retrieval // Journal of Documentation. 2004. V. 60(5). P. 493—502. https://doi.org/:10.1108/EB026526
16. *Beeferman D., Berger A. L., Lafferty J.D.* Statistical Models for Text Segmentation // Machine Learning. 1998. V. 34. P. 177–210. https://doi.org/:10.1108/EB026526
17. *Pevzner L., Hearst M.A.* A Critique and Improvement of an Evaluation Metric for Text Segmentation // Computational Linguistics. 2002. V. 28. P. 19–36. https://doi.org/:10.1162/089120102317341756