УДК 004.89 **DOI** 10.24412/2413-7383-2025-3-38-113-123

А. В. Ниценко, В. Ю. Шелепов, С. А. Большакова
Федеральное государственное бюджетное научное учреждение
«Институт проблем искусственного интеллекта», г. Донецк
283048, г. Донецк, ул. Артема, 118 б

АВТОМАТИЧЕСКОЕ РАЗБИЕНИЕ ТЕКСТА НА СЕМАНТИЧЕСКИ ОДНОРОДНЫЕ ФРАГМЕНТЫ (АБЗАЦЫ)*

A. V. Nicenko, V. Ju. Shelepov, S. A. Bolshakova
Federal Budgetary State Scientific Institution «Institute of Artificial Intelligence Problems»,
Donetsk, Artema str., 118-b.

AUTOMATIC TEXT SEGMENTATION INTO SEMANTICALLY HOMOGENEOUS FRAGMENTS (PARAGRAPHS)

Работа посвящена актуальной теме сегментации русскоязычного текста на семантически однородные фрагменты. В статье выполнен обзор ряда существующих подходов к задаче сегментации русского текста. Предложен алгоритм автоматического разбиения текста на абзацы, как тематически однородные фрагменты за счет использования отношения, учитывающего частоту встречаемости слова и длину отрезка текста, где оно встречается. С помощью разработанного программного обеспечения проведены эксперименты и сравнение предложенного алгоритма с другими методами сегментации. В результате установлено, что предложенный алгоритм демонстрирует лучшие показатели, нежели другие сравниваемые подходы.

Ключевые слова: семантическая сегментация текста, абзац, частота слова, отрезок встречаемости.

The work is devoted to the topical topic of segmentation of the Russian-language text into semantically homogeneous fragments. The paper provides an overview of a number of existing approaches to this task. An algorithm is proposed for automatically dividing text into paragraphs as thematically homogeneous fragments by using a ratio that takes into account the frequency of occurrence of a word and the length of the text segment where it occurs. Using this software, experiments were conducted and the proposed algorithm was compared with other text segmentation methods. As a result, it was found that the proposed algorithm demonstrates better performance than other compared approaches.

Keywords: semantic text segmentation**,** paragraph, word frequency, segment of occurrence.

* **Поддержка исследований.** Работа выполнена при финансовой поддержке Министерства науки и высшего образования РФ в рамках НИР №Г/Р 123092600030-4.

Введение

В настоящее время проблема семантической сегментации текста становится все более актуальной благодаря экспоненциальному росту объемов данных, зачастую представленных в виде текстовых документов. В том случае, когда в тексте присутствует определенная семантическая разметка (заголовки, абзацы и т. д.), сегментация не представляет проблем. Сложнее, когда такой информации нет или сегментацию нужно выполнить более детально. В таком случае появляется необходимость в алгоритмах, которые позволяют осуществить это автоматически.

В данной статье речь идет об автоматической сегментации русскоязычных текстов. Под сегментацией мы будем подразумевать разделение текста на абзацы так, чтобы предложения внутри каждого абзаца были объединены некоторой общей темой. Эта задача является необходимым этапом для последующих задач, таких как тематическая классификация документов, реферирование, извлечение, индексирование и поиск информации, построение связанной с текстом онтологии. В научной литературе существует большое количество работ, описывающих различные подходы к задаче автоматического разбиения сплошного текста на тематически однородные фрагменты, оформляемые в виде абзацев.

Обзор литературы

Существует несколько подходов к решению данной задачи. Часть из них основаны на лексической когезии – связи между частями текста через схожую лексику [1]. Согласно этому принципу, границы тем определяются точками лексических изменений [2], [3].

TextTiling [4] – один из наиболее ранних методов сегментации текстов по темам сравнивает блоки предложений, используя меры подобия (BOW, TF-IDF). Лексические изменения затем определяются по минимальным значениям коэффициента подобия. Данный метод определяет границы тем путем нахождения точек, в которых мера лексической когезии значительно изменяется.

Minimum Cut (MinCut) [5] моделирует документ в виде графа, где узлы – это предложения в документе, а ребра имеют значение веса, соответствующее значению подобия между двумя связанными узлами. Для определения сегментов находится минимальный разрез графа.

В методе [6] предполагается, что изменение темы в тексте происходит в местах, где частые повторения слов начинаются и заканчиваются, образуя цепочку. С помощью косинусного коэффициента подобия определяется уровень лексического сцепления между двумя лексическими цепями.

Методы на основе лексической когезии предназначены для работы с очень большими документами, содержащими достаточный объем статистически значимой лексической информации. Они хорошо справляются с задачей тематической сегментации научных текстов, но мало подходят, например, для сегментации текстов стенограмм заседаний или телефонных разговоров [1].

Другие подходы основываются на байесовских структурах. В работе [7] используется Байесовская сеть для сегментации, где каждая тема рассматривается, как отдельная языковая модель. Используется также латентное распределение Дирихле и двоичные переменные для указания сдвига темы между предложениями [8]. В [9] для тематической сегментации используется иерархическая байесовская модель.

Методы, использующие байесовские структуры, обладают способностью учитывать как лексическую связность, так и определять ключевые фразы и тем самым превосходят упомянутые ранее методы. Они хорошо справляются с сегментацией речи и научных текстов – лучше методов, основанных на лексической когезии [1].

В работе [10] предлагается новый метод решения задачи тематической сегментации для русскоязычного текста на основе графов знаний. Применение графов знаний при сегментации позволяет использовать больше информации о словах в тексте. Методы, основанные на базе графов знаний, могут применять расстояние между словами на графе, интегрируя тем самым фактологическую информацию из графа знаний в процесс принятия решений о разбиении текста на сегменты.

Применение к задаче сегментации объединенного подхода DNN-HMM (*Deep Neural Network-Hidden Markov Model*) – «глубокие нейронные сети-скрытые марковские модели» оказалось весьма успешным и позволило существенно улучшить точность сегментации [11-13]. Глубокая нейронная сеть (DNN) вычисляет апостериорную вероятность темы для слова, а HMM моделирует переходы между темами. Алгоритм Витерби отображает последовательность слов в последовательность тем. Изменение темы определяет границу сегмента текста. В работе [14] для сегментации по разделам текстов научных статей на русском языке используется многослойный персептрон.

Современные методы используют также иерархические рекуррентные нейронные сети (RNN) и сети трансформерной архитектуры, где сначала предложения представляются в виде векторов, а затем двунаправленная LSTM-сеть на уровне предложений моделирует переход между темами, основанный на этой последовательности векторов [15-22]. Методы данной группы могут применяться в любых типах задач и для текстов любого типа, при условии наличия достаточного объема обучающих данных и способны эффективнее других справляться с задачей тематической сегментации.

Ощутимым недостатком данных методов является необходимость предварительного обучения нейронных сетей, входящих в архитектуру моделей, на больших объемах данных. По этой причине сегментация русскоязычного текста на данный момент является сложной задачей ввиду отсутствия в открытом доступе достаточного набора данных для обучения и тестирования.

Описание предлагаемого алгоритма разбиения текста на тематически однородные фрагменты

Ниже предлагается простой и, как представляется, достаточно эффективный способ решения обсуждаемой проблемы, основанный на использовании часто встречающихся в рассматриваемом тексте существительных (назовем их ключевыми словами) и местах их концентрации. Мы используем словарь [23], содержащий более четырех миллионов русских словоформ с полной грамматической разметкой, добавив к этой разметке лемму слова. Представление этого словаря в виде дерева, обеспечивает почти мгновенный поиск в словаре словоформы текста и ее лемматизацию.

Рассмотрим некоторый фрагмент сплошного текста, подлежащий разбиению на абзацы. Формируется список лемм всех его слов, которые являются существительными и встречаются в тексте не менее двух раз (условие L). Количество вхождений словоформ, входящих в парадигму определенной леммы, будем называть частотой слова, которую обозначим через a. Помимо этого фиксируются номера первого и последнего предложения, где встречается слово: b и c соответственно. После чего вычисляется отношение

$$\frac{a}{c-b+1} \tag{1}$$

Отношение (1) тем больше, чем больше частота слова и чем меньше отрезок текста, на котором это слово сосредоточено.

Алгоритм находит (первый) максимум отношения (1), определяет соответствующие номера *b* и *c* и выделяет в качестве абзаца отрезок текста от предложения с номером *b* до предложения с номером *c* включительно. Затем проводится подсчет количества предложений в образовавшихся предыдущих и последующих фрагментах текста (не входящих в уже выделенные абзацы) и к наибольшему из них при выполнении условия L применяется вышеописанная процедура. Если для него условие L не выполняется, то же делается для второго по размеру из образовавшихся фрагментов. Процесс продолжается до тех пор, пока после выделения очередного абзаца для всех оставшихся фрагментов перестает выполняться условие L.

Следующий этап – анализ полученного разбиения на предмет анафор и абзацев, состоящих их одного предложения. Алгоритм анализирует те абзацы, которые в первом своем предложении содержат личные местоимения *он, она, оно*, а также указательные слова *этот, эта, это, там, туда, оттуда, столько*, поскольку именно они выполняют анафорическую функцию. Если такое предложение найдено и в нем нет предшествующего существительного в том же роде и числе, то абзац присоединяется к предыдущему. Объединяются также соседние абзацы, на стыке которых оказался общий фрагмент прямой речи. При этом, если начало прямой речи выделено с помощью тире, идущего вслед за двоеточием, то она, естественно, оформляется отдельным абзацем. К сожалению, конец этого абзаца пока приходится отмечать вручную. Наконец, абзац, состоящий из одного предложения, присоединяется к меньшему из соседних.

Описанный алгоритм реализован на языке программирования C++ в экспериментальном программном обеспечении для сегментации текста. С помощью данного программного обеспечения было проведено тестирование и оценка результатов работы алгоритма.

Экспериментальные результаты

Проиллюстрируем этапы работы предложенного алгоритма на примере 15-ой главы романа Даниеля Дефо «Робинзон Крузо» в пересказе К.И. Чуковского. Вот фрагмент этой главы, представленный в виде сплошного текста:

«Конечно, было бы хорошо иметь лодку на этой стороне острова, поближе к моему дому, но как привести ее оттуда, где я оставил ее? Обогнуть мой остров с востока – от одной мысли об этом у меня сжималось сердце и холодела кровь. Как обстоит дело на другой стороне острова, я не имел никакого понятия. Что, если течение по ту сторону такое же быстрое, как и по эту? Разве не может оно швырнуть меня на прибрежные скалы с той же силой, с какой другое течение уносило меня в открытое море. Словом, хотя постройка этой лодки и спуск ее на воду стоили мне большого труда, я решил, что все же лучше остаться без лодки, чем рисковать из-за нее головой. Нужно сказать, что теперь я стал гораздо искуснее во всех ручных работах, каких требовали условия моей жизни. Когда я очутился на острове, я совершенно не умел обращаться с топором, а теперь я мог бы при случае сойти за хорошего плотника, особенно если принять в расчет, как мало было у меня инструментов. Я и в гончарном деле (совсем неожиданно!) сделал большой шаг вперед: устроил станок с вертящимся кругом, отчего моя работа стала и быстрее и лучше; теперь вместо корявых изделий, на которые было противно смотреть, у меня выходила очень неплохая посуда довольно правильной формы. Но никогда я, кажется, так не радовался и не гордился своей изобретательностью, как в тот день, когда мне удалось сделать трубку. Конечно, моя трубка была первобытного вида –

из простой обожженной глины, как и все мои гончарные изделия, и вышла она не очень красивой. Но она была достаточно крепка и хорошо пропускала дым, а главное – это была все-таки трубка, о которой я столько мечтал, так как привык курить с очень давнего времени. На нашем корабле были трубки, но, когда я перевозил оттуда вещи, я не знал, что на острове растет табак, и решил, что не стоит их брать. К этому времени я обнаружил, что мои запасы пороха начинают заметно убывать. Это чрезвычайно встревожило и огорчило меня, так как нового пороха достать было неоткуда. Что же я буду делать, когда у меня выйдет весь порох? Как я буду тогда охотиться на коз и птиц? Неужели я до конца моих дней останусь без мясной пищи?»

В этом фрагменте текста содержится 18 предложений. Ниже приведена таблица 1, в которой приведены результаты вычислений. Во втором столбце таблицы после леммы указана частота (параметр *a*), затем в скобках приведены номера первого и последнего предложения, где встречается слово (параметры *b* и *c*). В предпоследнем столбце выведены значения отношения (1), в последнем – первый максимум этой величины отмечен тремя звездочками.

Таблица 1 – Результаты вычислений

Лемма	***a***	***(b, c)***	***a/(c-b+1)***	***Max a/(c-b+1)***
1 шаг – фрагмент [1;18]				
остров	5	(1, 13)	0.38	
трубка	4	(10, 13)	1.00	***
лодка	3	(1, 6)	0.50	
порох	3	(14, 16)	1.00	
время	2	(12, 14)	0.67	
дело	2	(3, 9)	0.29	
день	2	(10, 18)	0.22	
изделие	2	(9, 11)	0.67	
работа	2	(7, 9)	0.67	
сторона	2	(1, 3)	0.67	
течение	2	(4, 5)	1.00	
2 шаг – фрагмент [1;9]				
остров	4	(1, 8)	0.50	
лодка	3	(1, 6)	0.50	
дело	2	(3, 9)	0.29	
работа	2	(7, 9)	0.67	
сторона	2	(1, 3)	0.67	
течение	2	(4, 5)	1.00	***
3 шаг – фрагмент [6;9]				
лодка	2	(6, 6)	2	***
работа	2	(7, 9)	0.67	
4 шаг – фрагмент [7;9]				
работа	2	(7, 9)	0.67	***
5 шаг – фрагмент [14;18]				
порох	3	(14, 16)	1.00	***
6 шаг – фрагмент [1;3]				
остров	3	(1, 3)	1,5	***
сторона	2	(1, 3)	0.67	
7шаг – фрагмент [17;18]				
-	-	-	-	-

Для приведенного фрагмента текста алгоритм выдает следующий результат:

«Конечно, было бы хорошо иметь лодку на этой стороне острова, поближе к моему дому, но как привести ее оттуда, где я оставил ее? Обогнуть мой остров с востока – от одной мысли об этом у меня сжималось сердце и холодела кровь. Как обстоит дело на другой стороне острова, я не имел никакого понятия.

Что, если ТЕЧЕНИЕ по ту сторону такое же быстрое, как и по эту? Разве не может оно швырнуть меня на прибрежные скалы с той же силой, с какой другое ТЕЧЕНИЕ уносило меня в открытое море.

Словом, хотя постройка этой ЛОДКИ и спуск ее на воду стоили мне большого труда, я решил, что все же лучше остаться без ЛОДКИ, чем рисковать из-за нее головой.

Нужно сказать, что теперь я стал гораздо искуснее во всех ручных работах, каких требовали условия моей жизни. Когда я очутился на острове, я совершенно не умел обращаться с топором, а теперь я мог бы при случае сойти за хорошего плотника, особенно если принять в расчет, как мало было у меня инструментов. Я и в гончарном деле (совсем неожиданно!) сделал большой шаг вперед: устроил станок с вертящимся кругом, отчего моя работа стала и быстрее и лучше; теперь вместо корявых изделий, на которые было противно смотреть, у меня выходила очень неплохая посуда довольно правильной формы.

Но никогда я, кажется, так не радовался и не гордился своей изобретательностью, как в тот день, когда мне удалось сделать ТРУБКУ. Конечно, моя ТРУБКА была первобытного вида – из простой обожженной глины, как и все мои гончарные изделия, и вышла она не очень красивой. Но она была достаточно крепка и хорошо пропускала дым, а главное – это была все-таки ТРУБКА, о которой я столько мечтал, так как привык курить с очень давнего времени. На нашем корабле были ТРУБКИ, но, когда я перевозил оттуда вещи, я не знал, что на острове растет табак, и решил, что не стоит их брать.

К этому времени я обнаружил, что мои запасы ПОРОХА начинают заметно убывать. Это чрезвычайно встревожило и огорчило меня, так как нового ПОРОХА достать было неоткуда. Что же я буду делать, когда у меня выйдет весь ПОРОХ?

Как я буду тогда охотиться на коз и птиц? Неужели я до конца моих дней останусь без мясной пищи?»

Окно программы показано на рисунке 1.

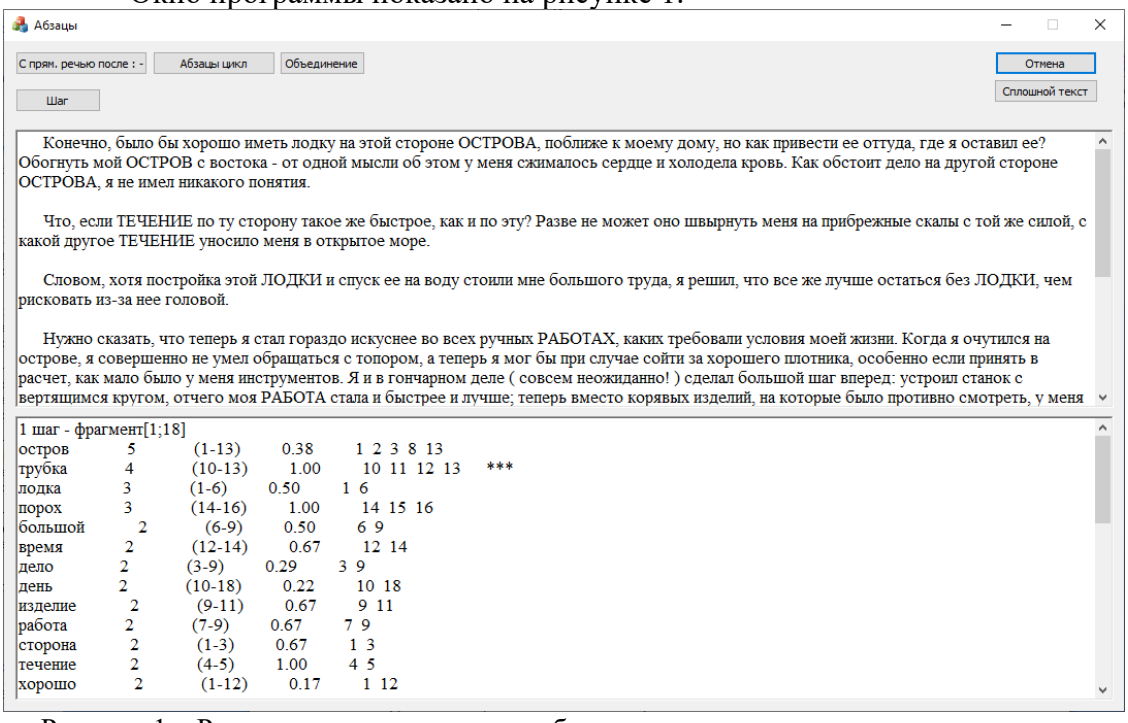

Рисунок 1 – Результат программного разбиения текста на тематически однородные фрагменты

Оценка результатов

Для оценки качества предлагаемого алгоритма было проведено сравнение результатов его работы с эталонной ручной разметкой с помощью двух популярных оценочных характкристик – Pk [24] и WindowDiff [25]. Для сравнения использовалось двоичное представление сегментации в виде последовательности символов «0» и «1», где «1» соответствует предложениям текста, за которыми следует граница сегмента (абзац), «0» – всем остальным предложениям. Обе характеристики используют фиксированное скользящее окно и оценивают, как расположены границы в пределах окна относительно друг друга. Характеристика Pk – это вероятность того, что при прохождении скользящего окна по предложениям текста, предложения на границах окна будут ошибочно классифицированы как принадлежащие к одному сегменту (или наоборот). Характеристика WindowDiff заключается в подсчете количества границ сегментов между началом и концом заданного окна и назначении штрафа, если это количество разное для экспериментального и эталонного разбиения. Оба показателя оценивают разницу между экспериментальной и эталонной сегментацией. Чем меньше значение, тем лучше сегментация. Для вычисления характеристик использовалась реализация из библиотеки Natural Language Toolkit Library [26].

В качестве набора данных для тестирования мы использовали текст художественного произведения (Д. Дефо «Робинзон Крузо» в пересказе К.И. Чуковского). С помощью разработанного программного обеспечения мы произвели разбиение на абзацы трех глав из данного произведения. В качестве эталона мы использовали текст данных глав, вручную размеченных человеком на абзацы, объединенные общей тематикой. В таблице 2 приведены результаты сравнения усредненных показателей разработанного алгоритма с другими известными методами тематической сегментации текста (на основе данных из работ [7],[10],[14],[20]).

Таблица 2 – Результаты сравнения с другими методами

Метод сегментации	**Язык документов**	**Тип документов**	**WinDiff**	**Pk**
TextTiling[4]	Англ.	Мед. литература	0.4	0.38
MinimumCut[5]	Англ.	Мед. литература	0.382	0.368
LCSEG[6]	Англ.	Мед. литература	0.385	0.37
BAYESSEG[7]	Англ.	Мед. литература	0.353	0.339
Граф знаний[10]	Рус.	Научные статьи	0.524	0.476
DNN[14]	Рус.	Научные статьи	0.37	0.27
BiLSTM[20]	Англ.	Стенограммы	0.43	0.41
BERT[20]	Англ.	Стенограммы	0.35	0.34
S-BERT[20]	Англ.	Стенограммы	0.349	0.336
Предлагаемый алгоритм	Рус.	Худ. литература	**0.347**	**0.292**

Заключение

В работе предложен метод автоматического разбиения текста на абзацы как тематически однородные фрагменты за счет использования отношения (1), учитывающего частоту встречаемости слова и длину отрезка текста, где оно встречается. Такой подход характеризуется малой вычислительной сложностью и не требует специальных лингвистических знаний, кроме грамматического словаря и простых правил, учитывающих анафорические ссылки. Опираясь на значения вычисленных характеристик качества, можно утверждать, что предложенный алгоритм сегментации текста демонстрирует показатели, улучшенные в сравнении с другими известными алгоритмами.

Список литературы

1. Чернобаев И.Д., Суркова А.С. Обзор методов тематической сегментации текстовых данных. *Информационные системы и технологии ИСТ-2018: Материалы докладов XXIV Международной научно-технической конференции, посвященной 100-летию Нижегородской радиолаборатории* (г. Нижний Новгород, 20 апреля 2018 года). Нижний Новгород: НГТУ им. Р.Е. Алексеева, 2018. С. 1079-1083.
2. Kirana R.P., Mukhrizal M., Jayanti F.G. Types of Lexical Cohesion and Grammatical Cohesion in Thesis Abstracts. *Jadila: Journal of Development and Innovation in Language and Literature Education*. 2020. vol. 1. no. 1. pp. 57-68. DOI: 10.52690/jadila.v1i1.14.
3. Мурай О.В. Когезия и когерентность в английской речи в публицистических текстах. *Современная наука: актуальные проблемы теории и практики. Серия: Гуманитарные науки*. 2021. №12-2. С. 171-174. DOI: 10.37882/2223-2982.2021.12-2.28.
4. Hearst M. A. TextTiling: Segmenting text into multiparagraph subtopic passages. *Computational* linguistics. 1997. vol. 23. no. 1. pp. 33-64.
5. Malioutov I., Barzilay R. Minimum cut model for spoken lecture segmentation. *Proceedings of the ACL*. 2006. pp. 25-32.
6. Galley M., McKeown K., Fosler-Lussier E., Jing H. Discourse segmentation of multi-party conversation. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. 2003. pp. 562-569.*
7. Eisenstein J., Barzilay R. Bayesian unsupervised topic segmentation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics*, 2008. pp. 334-343.
8. Kharisudin I., Masri'an H. Topic Modeling on WhatsApp User Reviews Using Latent Dirichlet Allocation // Scandinavian Journal of Immunology. 2022. vol. 9. no. 1. pp. 51-62. DOI: 10.15294/sji.v9i1.34941.
9. Du L., Buntine W., Johnson M. Topic segmentation with a structured topic model. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2013. pp. 190-200.
10. Авдеева З.К., Гаврилов М.С., Лемтюжникова Д.В., Шарафиев А.Ф. Методы решения задачи тематической сегментации текстов на основе графов знаний. *Известия Российской академии наук*. Теория и системы управления. 2024. № 4. С. 40-64. DOI: 10.31857/S0002338824040031.
11. Yu J., Shao H. Broadcast news story segmentation using sticky hierarchical Dirichlet process. *Applied Intelligence*. 2022. vol. 52. no. 11. pp. 12788-12800. DOI: 10.1007/s10489-021-03098-4.
12. Yu J. A DNN-HMM Approach to Story Segmentation //INTERSPEECH. 2016. pp. 1527-1531.
13. Vetráb M. Using Hybrid HMM/DNN Embedding Extractor Models in Computational Paralinguistic Tasks // Sensors. 2023. vol. 23, no. 11. pp. 5208. DOI: 10.3390/s23115208.
14. Баранов А. М., Юдина Т. А. Алгоритм сегментации научных статей, сочетающий принципы обучения с учителем и без учителя. *Новые информационные технологии в автоматизированных системах*. 2019. №22. URL: https://cyberleninka.ru/article/n/algoritm-segmentatsii-nauchnyh-statey-sochetayuschiy-printsipy-obucheniya-s-uchitelem-i-bez-uchitelya (дата обращения: 18.09.2025).
15. Javeed A. An LSTM model for extracting hierarchical relations between words for better topic modeling // *Journal of Physics: Conference Series*. 2021. vol. 1780, no. 1. pp. 12-19. DOI: 10.1088/1742-6596/1780/1/012019.
16. Inzer S., Cheng K., Leung A., Shen X. Multi-scale Hybridized Topic Modeling: A Pipeline for Analyzing Unstructured Text Datasets via Topic Modeling. *SIAM Undergraduate Research Online.* 2023. vol. 16. DOI: 10.1137/22s1536832.
17. Lomakina L. S., Surkova A. S., Zhevnerchuk D. V., Chernobaev I. D. Text structures synthesis on the basis of their system-forming characteristics. *IV International Research Conference "Information Technologies in Science, Management, Social Sphere and Medicine*" (ITSMSSM 2017). Advances in Computer Science Research (ACSR). 2017. vol. 72. pp. 108-113.
18. Memon M. Q., Lu Yu., Chen P. An ensemble clustering approach for topic discovery using implicit text segmentation. *Journal of Information Science*. 2021. vol. 47, no. 4. pp. 431-457. DOI: 10.1177/0165551520911590.
19. Sonata I., Heryadi Ya., Tho C. Topic Segmentation using Transformer Model for Indonesian Text. *Procedia Computer Science*. 2023. vol. 227. pp. 159-167. DOI: 10.1016/j.procs.2023.10.513.
20. Iikura R. Okada M., Mori N. Improving BERT with Focal Loss for Paragraph Segmentation of Novels. *Distributed Computing and Artificial Intelligence, 17th International Conference. DCAI 2020. Advances in Intelligent Systems and Computing*. 2021. Vol. 1237. DOI: 10.1007/978-3-030-53036-5_3.

21. Solbiati A., Heffernan K., Damaskinos G., Poddar S., Modi S., Cali J. Unsupervised Topic Segmentation of Meetings with BERT Embeddings. *arXiv preprint arXiv:2106.12978*. 2021. DOI: 10.48550/arXiv.2106.12978.
22. Sokol V., Krykun V., Bilova M. Topic segmentation methods comparison on computer science texts // *Вестник Национального технического университета "ХПИ". Серия Системный анализ, управление и информационные технологии*. 2021. no. 2 (6). pp. 59-66. DOI: 10.20998/2079-0023.2021.02.10.
23. Хаген М. *Полная парадигма. Морфология.* URL: https://ru.z-lib.fm/book/3305205/3a85b7/ (дата обращения: 15.09.2025).
24. Choi F. Y. Y. *Advances in domain independent linear text segmentation*. 2000. arxiv preprint arXiv: cs/ 0003083.
25. Pevzner L., Hearst M. A. A critique and improvement of an evaluation metric for text segmentation // *Computational Linguistics*. 2002. vol. 28. no.1. pp. 19-36.
26. Веб-сайт библиотеки NLTK. URL: https://www.nltk.org (дата обращения 18.09.2025).

References

1. Chernobaev I.D., Surkova A.S. Overview of methods of thematic segmentation of text data // Information Systems and Technologies IST-2018: Proceedings of the XV International Scientific and Technical Conference dedicated to the 100th anniversary of the Nizhny Novgorod Radio Laboratory. Nizhnij Novgorod: NGTU im. R.E. Alekseeva, 2018. pp. 1079-1083.
2. Kirana R.P., Mukhrizal M., Jayanti F.G. Types of Lexical Cohesion and Grammatical Cohesion in Thesis Abstracts // Jadila: Journal of Development and Innovation in Language and Literature Education. 2020. vol. 1. no. 1. pp. 57-68. DOI: 10.52690/jadila.v1i1.14.
3. Muraj O.V. Cohesion and coherence in English speech in journalistic texts // Modern science: actual problems of theory and practice. Series: Humanities. 2021. no. 12-2. pp. 171-174. DOI: 10.37882/2223-2982.2021.12-2.28.
4. Hearst M. A. TextTiling: Segmenting text into multiparagraph subtopic passages // Computational linguistics. 1997. vol. 23. no. 1. pp. 33-64.
5. Malioutov I., Barzilay R. Minimum cut model for spoken lecture segmentation // Proceedings of the ACL. 2006. pp. 25-32.
6. Galley M., McKeown K., Fosler-Lussier E., Jing H. Discourse segmentation of multi-party conversation // Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. 2003. pp. 562-569.
7. Eisenstein J., Barzilay R. Bayesian unsupervised topic segmentation // Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2008. pp. 334-343.
8. Kharisudin I., Masri'an H. Topic Modeling on WhatsApp User Reviews Using Latent Dirichlet Allocation // Scandinavian Journal of Immunology. 2022. vol. 9. no. 1. pp. 51-62. DOI: 10.15294/sji.v9i1.34941.
9. Du L., Buntine W., Johnson M. Topic segmentation with a structured topic model // Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2013. pp. 190-200.
10. Avdeeva Z.K., Gavrilov M.S., Lemtjuzhnikova D.V., Sharafiev A.F. Methods for solving the problem of thematic segmentation of texts based on knowledge graphs // Proceedings of the Russian Academy of Sciences. Theory and management systems. 2024. № 4. С. 40-64. DOI: 10.31857/S0002338824040031.
11. Yu J., Shao H. Broadcast news story segmentation using sticky hierarchical Dirichlet process // Applied Intelligence. 2022. vol. 52. no. 11. pp. 12788-12800. DOI: 10.1007/s10489-021-03098-4.
12. Yu J. A DNN-HMM Approach to Story Segmentation //INTERSPEECH. 2016. pp. 1527-1531.
13. Vetráb M. Using Hybrid HMM/DNN Embedding Extractor Models in Computational Paralinguistic Tasks // Sensors. 2023. vol. 23, no. 11. pp. 5208. DOI: 10.3390/s23115208.
14. Baranov A. M., Judina T. A. An algorithm for segmenting scientific articles that combines the principles of teaching with and without a teacher // New information technologies in automated systems. 2019. No. 22. Available at: https://cyberleninka.ru/article/n/algoritm-segmentatsii-nauchnyh-statey-sochetayuschiy-printsipy-obucheniya-s-uchitelem-i-bez-uchitelya (accessed: 18.09.2025).
15. Javeed A. An LSTM model for extracting hierarchical relations between words for better topic modeling // Journal of Physics: Conference Series. 2021. vol. 1780, no. 1. pp. 12-19. DOI: 10.1088/1742-6596/1780/1/012019.

16. Inzer S., Cheng K., Leung A., Shen X. Multi-scale Hybridized Topic Modeling: A Pipeline for Analyzing Unstructured Text Datasets via Topic Modeling // SIAM Undergraduate Research Online. 2023. vol. 16. DOI: 10.1137/22s1536832.
17. Lomakina L. S., Surkova A. S., Zhevnerchuk D. V., Chernobaev I. D. Text structures synthesis on the basis of their system-forming characteristics // IV International Research Conference "Information Technologies in Science, Management, Social Sphere and Medicine" (ITSMSSM 2017). Advances in Computer Science Research (ACSR). 2017. vol. 72. pp. 108-113.
18. Memon M. Q., Lu Yu., Chen P. An ensemble clustering approach for topic discovery using implicit text segmentation // Journal of Information Science. 2021. vol. 47, no. 4. pp. 431-457. DOI: 10.1177/0165551520911590.
19. Sonata I., Heryadi Ya., Tho C. Topic Segmentation using Transformer Model for Indonesian Text // Procedia Computer Science. 2023. vol. 227. pp. 159-167. DOI: 10.1016/j.procs.2023.10.513.
20. Iikura R. Okada M., Mori N. Improving BERT with Focal Loss for Paragraph Segmentation of Novels // Distributed Computing and Artificial Intelligence, 17th International Conference. DCAI 2020. Advances in Intelligent Systems and Computing. 2021. Vol. 1237. DOI: 10.1007/978-3-030-53036-5_3.
21. Solbiati A., Heffernan K., Damaskinos G., Poddar S., Modi S., Cali J. Unsupervised Topic Segmentation of Meetings with BERT Embeddings // arXiv preprint arXiv:2106.12978. 2021. DOI: 10.48550/arXiv.2106.12978.
22. Sokol V., Krykun V., Bilova M. Topic segmentation methods comparison on computer science texts // Bulletin of the National Technical University "KHPI". A series of System analysis, management and information technology. 2021. no. 2 (6). pp. 59-66. DOI: 10.20998/2079-0023.2021.02.10.
23. Hagen M. The complete paradigm. Morphology. Available at: https://ru.z-lib.fm/book/3305205/3a85b7/ (acessed: 15.09.2025).
24. Choi F. Y. Y. Advances in domain independent linear text segmentation. 2000. arxiv preprint arXiv: cs/ 0003083.
25. Pevzner L., Hearst M. A. A critique and improvement of an evaluation metric for text segmentation // Computational Linguistics. 2002. vol. 28. no.1. pp. 19-36.
26. NLTK. Available at: https://www.nltk.org (accessed: 18.09.2025).

RESUME

A.V. Nicenko, V. Ju. Shelepov, S.A. Bolshakova

Automatic text segmentation into semantically homogeneous fragments (paragraphs)

Background: Currently, the problem of semantic segmentation of text is becoming increasingly relevant due to the exponential growth of big data, often presented in the form of text documents. When there is a certain semantic markup in the text segmentation is not a problem. It is more difficult when there is no such information or segmentation needs to be performed in more detail. In this case, there is a need for algorithms that allow this to be done automatically.

Materials and methods: The paper provides an overview of a number of existing approaches to this task. An algorithm is proposed for automatically dividing text into paragraphs as thematically homogeneous fragments by using a ratio that takes into account the frequency of occurrence of a word and the length of the text segment where it occurs. This approach does not require training on large amounts of data and special linguistic knowledge, except for a grammatical dictionary and simple rules that take into account anaphoric references.

Results: This algorithm is implemented in experimental text segmentation software. Using this software, experiments were conducted and the proposed algorithm was compared with other well–known text segmentation methods using two popular evaluation characteristics - WinDiff and Pk.

Conclusion: A comparison was made between automatic division into paragraphs, and the manual division. As a result, it was found that the proposed text segmentation algorithm demonstrates slightly better performance than other compared approaches.

РЕЗЮМЕ

А. В. Ниценко, В. Ю. Шелепов, С.А. Большакова
Автоматическое разбиение текста на семантически однородные фрагменты (абзацы)

Предпосылки: В настоящее время проблема семантической сегментации текста становится все более актуальной благодаря экспоненциальному росту объемов данных, зачастую представленных в виде текстовых документов. В том случае, когда в тексте присутствует определенная семантическая разметка, сегментация не представляет проблем. Сложнее, когда такой информации нет или сегментацию нужно выполнить более детально. В таком случае появляется необходимость в алгоритмах, которые позволяют осуществить это автоматически.

Материалы и методы: В работе выполнен обзор ряда существующих подходов к этой задаче. Предложен алгоритм автоматического разбиения текста на абзацы как тематически однородные фрагменты за счет использования отношения, учитывающего частоту встречаемости слова и длину отрезка текста, где оно встречается. Этот подход не требует обучения на больших объемах данных и специальных лингвистических знаний, кроме грамматического словаря и простых правил, учитывающих анафорические ссылки.

Результаты: Данный алгоритм реализован в экспериментальном программном обеспечении для сегментации текста. С помощью данного программного обеспечения проведены эксперименты и сравнение предложенного алгоритма с другими известными методами сегментации текста с использованием двух популярных оценочных характеристик – WinDiff и Pk.

Заключение: Проведено сравнение текста художественного произведения, автоматически разбитого на абзацы, с разбиением, выполненным человеком вручную. В результате установлено, что предложенный алгоритм сегментации текста демонстрирует показатели несколько лучшие, нежели другие сравниваемые подходы.

Ниценко Артём Владимирович – к.т.н., старший научный сотрудник отдела распознавания речевых образов, Федеральное государственное бюджетное научное учреждение «Институт проблем искусственного интеллекта», г. Донецк. *Область научных интересов*: искусственный интеллект, обработка естественного языка, компьютерная лингвистика. Эл. почта: nav_box@mail.ru, адрес: 283048, г. Донецк, ул. Артема, д. 118 б, телефон: +7 (856) 311-34-24.

Шелепов Владислав Юрьевич – д.ф.-м.н., профессор, главный научный сотрудник отдела распознавания речевых образов, Федеральное государственное бюджетное научное учреждение «Институт проблем искусственного интеллекта», г. Донецк. *Область научных интересов*: искусственный интеллект, обработка естественного языка, компьютерная лингвистика. Эл. почта: vladislav.shelepov2012@yandex.ru, адрес: 283048, г. Донецк, ул. Артема, д. 118 б, телефон: +7 (856) 311-34-24.

Большакова Светлана Анатольевна – младший научный сотрудник отдела распознавания речевых образов, Федеральное государственное бюджетное научное учреждение «Институт проблем искусственного интеллекта». *Область научных интересов*: искусственный интеллект, компьютерная лингвистика, интеллектуальный анализ информации. Эл. почта: svetlako@yandex.com, адрес: 283048, г. Донецк, ул. Артема, д. 118 б, телефон: +7 (856) 311-34-24.

Статья поступила в редакцию 15.05.2025.