Исследование специфики применения алгоритмов тематической сегментации для научных текстов

© К. К. Боярский
boyarin9@yandex.ru

© Н. Ф. Гусарова
Университет ИТМО
Санкт-Петербург
natfed@list.ru

©Н. В. Добренко
graziokisa@gmail.com

© Е. А. Каневский
Санкт-Петербург
kanev@emi.nw.ru

Assoul@yandex.ru

Аннотация

В статье рассматривается специфика применения алгоритмов тематической сегментации к реальным научным текстам. В качестве экспериментальной базы использованы монографии на трех языках по единой тематике, причем в исследуемый корпус включены параллельные выровненные фрагменты монографий на языках оригинала, а также их профессиональных переводов. В качестве реперного алгоритма выбран *TextTiling*, использующий локальную информацию о связности между соседними частями текста. Исследовано влияние на качество сегментации текста таких параметров, как размер скользящего окна, величина перекрытия между окнами, пороговый уровень. Определены оптимальные комбинации параметров сегментации для различных языков. На примере русского языка подтверждено, что подключение внешних лексических ресурсов (стоп-слова, классификатор, тезаурус) существенно повышает качество сегментации.

1 Введение

Эффективность исследовательской работы во многом предопределяется полнотой охвата релевантных информационных источников. Большую долю таких источников составляют тексты, относящиеся к жанру научной прозы, такие как монографии, учебники, научные статьи и т. д. Это, как правило, большие, информационно-насыщенные документы на языке оригинала, обладающие характерной для научного текста внутренней структурой [16, 26, 30]. При этом совокупный объем текстов по конкретной тематике невелик (1 – 2 монографии), что не позволяет корректно применять для их анализа такие статистические методы, как латентное размещение Дирихле, нейронные сети и т. п.

Во всем мире ведется активная работа по предоставлению доступа к первичным научным текстам через Интернет, однако эффективность информационного поиска в массиве таких документов остается невысокой. Дело в том, что их внутренняя структура далеко не всегда представима традиционными для Интернета поисковыми атрибутами – например, метатегами или ключевыми словами. В результате в ответ на свой поисковый запрос пользователь получает либо документ целиком, в котором он вынужден вручную осуществлять линейный поиск, либо вырванные из контекста страницы с максимальной частотой встречаемости ключевых слов, по которым сложно составить цельное представление о конкретной тематике, обсуждаемой в документе.

Таким образом, необходима организация информационного поиска внутри научного текста, что, в свою очередь, требует решения задачи автоматизированного деления документа на подтемы с учетом внутренних особенностей текста и целей пользователя. Эта задача особенно

Труды XVII Международной конференции DAMDID/RCDL'2015 «Аналитика и управление данными в областях с интенсивным использованием данных», Обнинск, 13-16 октября 2015

актуальна при первичном отборе источников, при работе с источниками на иностранных языках, при быстром погружении в проблематику исследования и т. п.

К настоящему времени в литературе описано достаточно большое число алгоритмов тематической сегментации текстов (их краткий обзор приводится ниже). Однако, как правило, они демонстрируют достаточную эффективность на искусственно созданных текстах, представляющих собой конкатенацию отдельных предложений или коротких текстовых фрагментов из газетных или интернет-публикаций (см., например, [3, 5, 9], или на больших текстовых корпусах [4, 7, 21]). В то же время результаты применения этих алгоритмов к реальным научным текстам представлены крайне скудно и часто противоречат друг другу (ср., например, [5, 9, 11]).

В статье рассматривается специфика применения алгоритмов тематической сегментации к реальным научным текстам. В качестве экспериментальной базы использованы монографии на трех языках по единой тематике, причем в исследуемый корпус включены параллельные выровненные по предложениям фрагменты монографий на языках оригинала, а также их профессиональных переводов. Исследуется зависимость качества сегментации текста от таких параметров, как размер скользящего окна, величина перекрытия между окнами, пороговый уровень, язык текста, отбор лексических единиц для анализа, а также от подключения внешних лексических ресурсов. Такая сегментация важна для выделения внутри авторской разметки (глав, параграфов) фрагментов для дословного перевода или другой детальной работы с текстом.

2 Смежные исследования

Практически все алгоритмы тематической сегментации текстов базируются на связности последних. Как показано в исследованиях [27, 31], в научных текстах преобладают такие типы связности субтекстов, как лексический (абсолютный, синонимический и заместительный) и лексико-синтаксический (параллельный):

- абсолютный тип связи – точный повтор терминов в смежных предложениях субтекста;
- синонимический тип связи – использование терминов, близких по смыслу, в одном контексте;
- заместительный тип связи – использование местоимения или термина с указательным местоимением для замещения смысла предшествующего предложения;
- параллельный тип связи – наличие серии предложений, раскрывающих тезис контекста, синтаксическими признаками которых являются параллелизм структуры, а также единство форм выражения сказуемых. Большинство существующих алгоритмов основано на выявлении в тексте связей типа (1). Алгоритмы этого типа можно разделить на две группы. Алгоритмы первой группы используют локальную информацию о связности между соседними частями текста. Один из наиболее известных алгоритмов этой группы – *TextTiling* [11, 12,] – состоит из следующих шагов:

- выполняется лемматизация текста и удаление стоп-слов; таким образом, текст преобразуется в последовательность из N токенов;
- последовательные токены группируются в псевдопредложения длиной W токенов каждое; из них формируется блок размером k псевдопредложений, который в дальнейшем используется как скользящее окно с шагом s псевдопредложений. В базовом алгоритме s=1, т.е. сравниваются сначала группа токенов от 0-го до (W*k)-го с группой токенов от W-го до (W*(k +1))-го, затем группа от W-го до (W*(k + 1))-го с группой от (W*2)-го до (W*(k+2))-го и т. д. до попадания последнего псевдопредложения на вторую границу;
- оценивается лексическая схожесть соседних блоков как косинус угла φ между (W*k)-мерными векторами:
- $$\cos\varphi_i = \frac{\sum_n w_{n,i-1} w_{n,i}}{\sqrt{\sum_n w_{n,i-1}^2}\sqrt{\sum_n w_{n,i}^2}}, \quad 0 \le \cos\varphi \le 1, \quad (1)$$
- где $w_{n,i}$ – вес n-го токена в i-м блоке.
- локальные минимумы функции (1) рассматриваются как границы между сегментами текста (с округлением до ближайшего предложения или абзаца).

Алгоритмы второй группы оценивают распределение повторяющихся токенов по всему тексту. Так, алгоритм *DotPlotting* [20] использует для анализа лексической связности двумерный график, на котором по осям X и Y откладывается положение токенов в тексте; если на позициях x и y в тексте находятся идентичные токены, то на графике ставятся точки в позициях (x, y) и (y, x). При этом связанные сегменты текста визуально соответствуют квадратам, расположенным вдоль диагонали, с высокой плотностью точек. Полученное распределение исследуется на экстремум с применением одной из двух стратегий: либо минимизация плотности точек на границах, либо максимизация плотности точек внутри сегмента. Развитием этой идеи является алгоритм *C99* [4], в котором аналогичным образом визуализируется мера лексической схожести между токенами соседних фрагментов, а затем методами динамического программирования находятся зоны с максимальной плотностью этой меры.

Таблица 1. Характеристики текстовых источников

Обозначение источника	Библиографическое описание источника	Язык источника	Размер фрагмента (печатн. знаков)
1. Romme	L' Art de la Marine, ou Principes et Préceptes Generaux de l'Art de Construire, d'Armer , de Manœuvrer et de Conduire des Vasseaux, par M. Romme. La Rochelle, 1787. Chapitres VII, VIII.	французский	110261
2. Ромм	Искусство, или главные начала и правила, научающие искусству строения, вооружения, правления и вождения кораблей, сочиненные господином Роммом, корреспондентом Парижской Академии наук и профессором навигации морского училища. Часть 1, 2. Перевел с французского флота капитан Александр Шишков. Санкт-Петербург, 1793. Главы 7, 8.	русский	161152
3. U-boat	Williamson G., Johnson L.: U-Boat crew 1914-45/ ed. Osprey Publishing, Great Britain, 1995.	английский	60563
4. Подводники	Подводный флот Германии. 1914–1945 / Г. Уильямсон; Пер. с англ. М.А. Мальцевой. – М.: ООО «Издательство АСТ», 2003.	русский	60560
5. Новости	Новостные источники – интернет	русский	23800

Предложен целый ряд модификаций базовых алгоритмов (см., например, [5, 6, 8, 13, 17]), позволяющих, помимо абсолютной связи, в той или иной мере учесть другие типы связи, присутствующие в тексте. Например, для учета связи типа (2) предлагается подключать внешние словари (например, *WordNet*) или интегрировать в косинусную меру коэффициент, отражающий частоту встречаемости слова во внешней коллекции документов (например, в Интернете) [5]. Для учета связи типа (4) предлагается одновременно использовать в оптимизируемом функционале *DotPlotting* меры лексического сходства внутри сегмента и между сегментами текста [25].

Анализируя рассмотренные работы, можно выделить ряд проблем, препятствующих их внедрению в реальную практику поисковых систем. Во-первых, большинство методов, представленных в литературе, разработано для английского языка. Другие языковые группы применительно к задаче тематической сегментации изучены слабее [3, 24]. Во-вторых, для многих языков, включая русский, в свободном доступе отсутствуют достаточно полные словари синонимов и другие сетевые лингвистические ресурсы. И, в-третьих, эффективность всех предложенных алгоритмов критически зависит от неизвестной заранее статистики параметров связности обрабатываемого текста.

В последнее время получили развитие алгоритмы иерархической сегментации текстов (см., например, [7, 14, 21]). Базой большинства из них служит модель лексических связей слов текста в виде многомерного распределения слов по темам, причем текст рассматривается как смесь небольшого количества тем, а появление каждого слова связывается с одной из тем. Математическую основу для такого подхода предоставляет латентное размещение Дирихле (*LDA*) [2], широко используемое в машинном обучении. Например, в [7] на основе *LDA* реализован иерархический байесовский алгоритм, позволяющий выявить два уровня линейной сегментации текста. Алгоритм *TopicTiling* [21], аналогично *TextTiling*, основан на вычислении косинусной меры сходства между соседними сегментами текста, однако в ней используются не частоты встречаемости слов, а частоты идентификаторов тем, предварительно рассчитанных для каждого слова текста с помощью *LDA*.

Алгоритмы этого типа показывают тем лучшую эффективность, чем больше обучающий корпус текстов и чем ближе его статистические распределения к анализируемому тексту. В идеале весь обучающий корпус и анализируемый текст должны принадлежать к одному домену [21], чего трудно ожидать для реального научного текста, ценность которого во многом связывается с его уникальностью.

В настоящей статье проводятся выбор и экспериментальная оценка параметров сегментации, важных с точки зрения организации тематической сегментации реальных научных текстов. Рассматривается базовый вариант – линейная сегментация и косинусная мера сходства между сегментами.

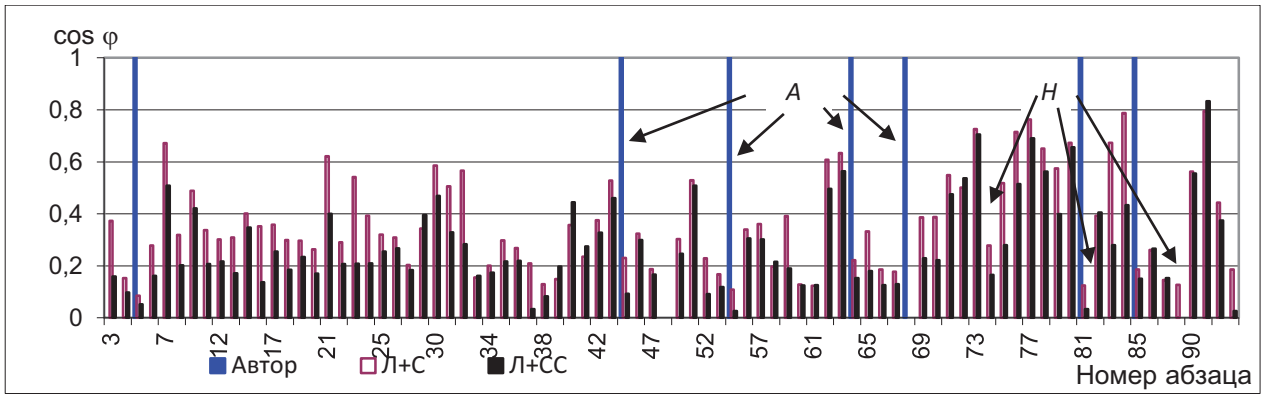

Рис. 1. Авторская разметка (*A*) и автоматически обнаруженные границы сегментов (*H*) при двух типах предобработки для текста «Romme»

3 Организация эксперимента

3.1 Отбор и предварительная обработка текстов

Тексты извлекались из монографий технической тематики на трех языках (русский, английский, французский) по единой тематике. Характеристики источников представлены в табл. 1. В исследуемый корпус вошли параллельные фрагменты монографий на языке оригинала (поз. 1, 3), а также их профессиональных переводов (поз. 2, 4). Такой подход снимает проблему идентичности домена для разных текстов и позволяет в чистом виде исследовать языковые особенности параметров сегментации. Частеречная разметка текстов на английском и французском языках выполнялась с помощью сетевого сервиса OpenXerox [18], на русском языке – с помощью синтаксического анализатора *SemSin* [29]. Список стоп-слов для каждого текста был сформирован вручную на основе частотного анализа. В качестве реперного был сформирован текст, состоящий из 20 новостных фрагментов, взятых случайным образом из Интернета (поз. 5 в табл. 1). До проведения сегментации осуществлялась предобработка текстов различными методами (табл. 2).

Таблица 2. Варианты предобработки текстов

Обозначение	Описание
Л	
	лемматизация + удаление стоп-слов
Л+С	лемматизация + частеречная разметка + отбор существительных
Л+С+П+Г	лемматизация + частеречная разметка + отбор существительных, прилагательных и глаголов
Л+С+ВнКл	лемматизация + частеречная разметка + подключение внешнего классификатора

3.2 Используемый алгоритм обработки текста

В качестве базового алгоритма обработки был выбран *TextTiling*, так как он наиболее прозрачно реализует механизм связности и косинусную меру. Последовательность шагов алгоритма и обозначения параметров подробно описаны выше (см. п. 2). В ходе экспериментов варьировались следующие параметры алгоритма: размер блоков W^*k [токенов]; размер перекрытия между блоками s^*k [токенов]. Кроме того, в ряде экспериментов использованы блоки переменной длины, соответствующей длине абзаца.

3.3 Метрики оценки качества сегментации

Для оценки качества сегментации текстов предложен ряд метрик, в том числе соотношение recall-precision [12], расстояние редактирования [19], меры $P\mu$ и P_k [1], метрика WindowDiff. Каждая из них не свободна от недостатков.

Например, метрика WindowDiff сравнивает положение референтных границ сегментов, сформированных в соответствии с принятой базовой линией, и границ, полученных оцениваемым алгоритмом, в пределах скользящего окна. Затем число окон, содержащих ошибочно определенные границы, нормируется на общее число окон. Однако последующие исследования [10, 15, 22] выявили ряд недостатков метрики WindowDiff. Она одинаково оценивает ложно определенные (*false positive, FP*) и пропущенные (*false negative, FN*) границы, которые должны иметь разную значимость в зависимости от конкретной задачи сегментации. Кроме того, метрика WindowDiff не позволяет учесть величину ошибки определения границы, а также придает завышенное значение ошибкам в начале и конце текста. В связи с этим предложен ряд модификаций метрики WindowDiff (см., например, [22]).

В нашем исследовании для оценки качества сегментации использовалась сбалансированная F-мера [28]:

$$F = \frac{2 * P * R}{P + R}, \qquad (2)$$

где

$$P = \frac{TP}{H}$$ – точность, $$R = \frac{TP}{TP + FN}$$ –

полнота, *FP* – количество ложных значений, *FN* – количество пропущенных значений, H – число найденных границ сегментов.

В качестве эталона для сравнения мы взяли экспертную разметку границ смены тем (*A*), см. рис. 1. В рассмотренных фрагментах текста таких границ обнаруживается от семи до десяти.

Подбор параметров сегментации проводился следующим образом. Анализировалось значение косинусной меры (1) в зависимости от варьируемого порогового уровня *z*. Значения cos φ, меньшие или равные *z*, рассматривались как смена темы, т. е. граница сегментов (*H*). Затем проводилось сопоставление полученных последовательностей *A* и *H*. Как правильное совпадение (*TP*) рассматривались случаи, когда в обеих последовательностях (*A* и *H*) были отмечены локальные минимумы («провалы») для одного и того же абзаца или для соседних абзацев. Такое округление представляется обоснованным, поскольку, как показал детальный анализ, часто в первом абзаце новой темы автор пытается плавно сменить тему, и, в сущности, переход происходит только в следующем абзаце. Остальные «провалы» *H* отмечались как ложные (*FP*), а «провалы» *A* – как пропущенные (*FN*). Для определения наилучшего значения F-меры уровень порога *z* варьировался от 0 до 1.

Отметим, что рассмотренные выше метрики оценивают качество сегментации только постфактум и не обеспечивают возможность тонкой настройки (оптимизации) параметров алгоритма сегментации, в то время как для разных задач требуется разное соотношение точности и полноты, что особенно важно для научных текстов. Это послужило еще одним аргументом в пользу выбора в качестве метрики F-меры (2), так как ее удобно рассматривать как критерий оптимизации, а ее максимум – как наилучшее соотношение *P* и *R*. Она легко модернизируется: в ней, например, легко ставить весовые коэффициенты между *P* и *R*.

4 Экспериментальные результаты и обсуждение

4.1 Влияние размера перекрытия между блоками

В качестве оптимальных параметров сегментации в базовом варианте алгоритма *TextTiling* [11] рекомендуется скользящее окно (блок) размером (*W***k*) токенов из *k* псевдопредложений длиной *W* токенов каждое, при перекрытии блоков *s*=(*W***k*)/2. Однако наши эксперименты с использованием разных комбинаций *s, W* и *k* показали, что любые значения перекрытия, отличные от нуля, снижают контрастность графика, и провалы, индицирующие границы смысловых фрагментов, практически исчезают (характерный пример приведен на рис. 2). Отметим, что этот результат совпадает с выводом, сделанным в экспериментальном исследовании [5]. Поэтому в наших дальнейших экспериментах перекрытие блоков не использовалось.

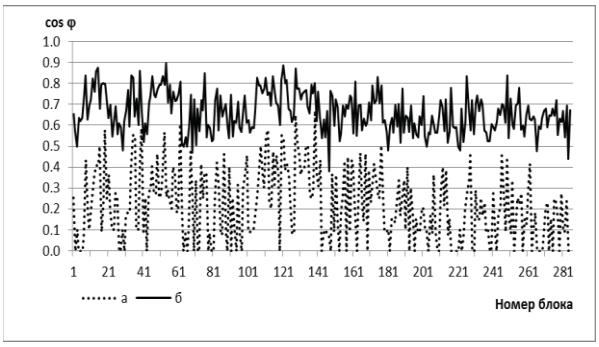

Рис. 2. Значения косинусной меры для текста «U-boat» с предобработкой типа Л+С: кривая а – *W***k*=10, *s*=0; кривая б – (*W***k*)=20, *s*=(*W***k*)/2;

4.2 Влияние размера блока

Как отмечалось выше, в базовом алгоритме *TextTiling* [11, 12] в качестве единицы анализа использованы окна фиксированной длины, а положение получающихся границ округляется до ближайшего предложения или абзаца. Наши эксперименты показали, что применительно к реальным научным текстам качество анализа в этом случае неудовлетворительно (табл. 3). Это связано с тем, что при коротком окне (10 существительных) случайные изменения лексики приводят к появлению большого числа «ложных тревог», уменьшая точность, а при большом окне (40 сущ.) велика вероятность того, что окно перекроет реальную границу, что приводит к падению полноты. Пример результатов для источника «U-boat» с предобработкой типа Л+С приведен в таблице 3, варианты 1–3.

Таблица 3. Влияние размера блока

№ варианта	1	2	3	4
Размер блока [токенов]	10	25	40	абзац
F-мера	0.06	0.04	0.03	0.17

В своей работе мы также исследовали предположение, что смысловые границы следует искать на границах абзацев. Нужно заметить, что в литературе представлены противоречивые мнения по этому поводу. Например, в [32] констатируется: «...можно сказать, что абзац в научном тексте – самостоятельный, графически выделенный элемент текста, содержащий одну развернутую мысль или ее фрагмент». С другой стороны, согласно [23], деление текста по формальным составляющим (абзацам или разделам) не позволяет выделить

подтемы документа: границы абзацев и тематических блоков могут не совпадать, деление на абзацы зависит от типа и назначения документа (например, новостной текст и художественный), большие абзацы могут содержать в себе несколько подтем.

Для экспериментов мы использовали окна переменной длины, совпадающие с абзацем. Слишком короткие абзацы присоединялись к следующему, так как с достаточно большой вероятностью такие абзацы представляют собой заголовки или авторские отступления, в которых вообще нет слов, совпадающих со словами окружающего контекста.

Результаты анализа демонстрируют резкое увеличение значения *F* в этом случае по сравнению с фиксированной длиной блока (табл. 3, вариант 4).

Поэтому в дальнейшем для всех текстов использовалось членение по абзацам.

4.3 Влияние порогового уровня

Определяющим параметром для принятия решения о проведении границы сегмента в данном месте является уровень порога *z*. По мере его роста уменьшается точность *P*, т. к. появляются пропуски границ, но растет полнота *R*, т. к. все большее число границ обнаруживается. Таким обратом, существует оптимальный уровень порога с точки зрения F-меры (рис. 2). Как показали наши эксперименты, оптимальное значение *z* сильно зависит от типа предобработки текста. Для всех вариантов предобработки, кроме Л+С+ВнКл (табл. 2), максимальные значения F-меры достигаются при *z*=0.1…0.15. При обработке с подключением дополнительных ресурсов (Л+С+ВнКл), более подробно описанной в разделе 4.5, максимальное значение F-меры достигается при бóльших значениях *z* (рис. 3).

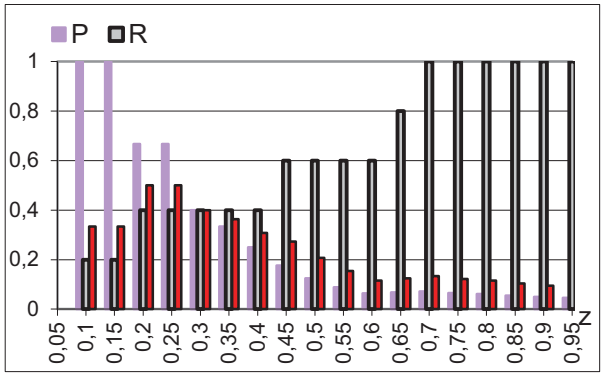

Рис. 3. Зависимость P, R и F-меры от уровня порога z

4.4 Влияние языка источника на выбор типа предобработки

Нами были исследованы возможности оптимизации предобработки в зависимости от языка текста. Очевидно, что тип предобработки фактически определяет способ отбора слов для анализа текстов на разных языках, и его можно соотнести с представленными в языке типами связности. В частности, Л+С и Л+СС (с исключением наиболее частотных слов) фиксируют абсолютный тип связи. Л+С+П+Г (наиболее стандартный способ, в качестве стоп-слов рассматриваются только служебные части речи), наряду с абсолютным, частично фиксирует параллельный тип связи.

В анализируемых текстах, помимо стандартных стоп-слов, исключались наиболее частотные содержательные слова: *лодка* и *Германия* – в русских текстах, *U-boat*, *jacket*, *war* – в английских, *voile* (парус), *poulie* (блок), *mat* (мачта), *fig* (фиг, ссылка на рисунок), *vergue* (рей) – во французских.

Значения F-меры для оптимального порогового уровня, полученные для различных текстов и предобработок «по словам» (т.е. для всех вариантов предобработки, кроме Л+С+ВнКл), приведены в табл. 4. Максимальные значения выделены полужирным шрифтом. Отметим, что полученные значения вполне соответствуют результатам таких работ, как [3, 5].

Таблица 4.

Текст			
		Л+С	Л+СС
Подводники	0.19	**0.21**	0.11
U-boat	**0.24**	0.17	0.19
Romme	0.46	0.44	**0.53**
Ромм	0.21	**0.22**	0.18
Новостные		**0.60**	

Анализ таблицы показывает, что в научных текстах авторское деление заголовками и подзаголовками не означает смены лексикона и плохо улавливается при автоматическом анализе. Текст оказывается «низкоконтрастным» с точки зрения используемой лексики. В то же время по набору сравнительно небольших новостных текстов даже без применения каких-либо дополнительных мер качество деления значительно выше.

Эти результаты легко проинтерпретировать с помощью диаграмм косинусной меры сходства между абзацами на примере источника «Romme» (рис. 1).

Сплошными вертикальными линиями показано авторское членение текста на подразделы. Видно, что удаление стоп-слов, естественно, понижает меру сходства, однако «провалы», интерпретируемые как разрыв между близкими по смыслу фрагментами, остаются на прежних местах. Поэтому качество деления, определяемое по F-мере, меняется незначительно: 0.46 по всем существительным при

отсечке по уровню $z = 0.15$ против 0.53 с удаленными стоп-словами при отсечке по уровню $z = 0.1$.

Таким образом, обнаружено, что в русском тексте предпочтительно использование только существительных, причем без исключения частотных (по типу Л+С). В английском тексте существенное повышение качества сегментации дает учет прилагательных и глаголов. Во французском тексте небольшой эффект дает удаление стоп-слов.

4.5 Влияние внешних лексических ресурсов

Из литературы известно, что подключение внешних лексических ресурсов существенно повышает устойчивость (робастность) сегментации, причем с этой целью может использоваться не только словарь синонимов [5], но и тематический классификатор слов текста [21]. Для русского языка словарь синонимов необходимого качества в открытом доступе отсутствует, поэтому в наших экспериментах мы использовали семантический классификатор [33], содержащий 1680 классов, по которым распределены более 190 тыс. лексем [29]. Классы этого классификатора в какой-то степени соответствуют гиперонимам WordNet. При обработке предполагалось, что слова, принадлежащие одному классу, совпадают, например, слова *судно*, *корабль*, *фрегат*,… в тексте «Ромм» или *каска*, *фуражка*, *бескозырка* в тексте «Подводники».

Рассмотрим характерный пример. Возьмем два предложения из смежных абзацев текста «Подводники»:

Ленточки бескозырок матросов кайзеровского флота имели надпись прописными печатными буквами, вышитыми золотой или серебряной канителью

и

Пилотка кроилась из темно-синего плотного сукна, обычно с черной или темно-синей подкладкой из искусственного шелка.

Эти предложения, несмотря на очевидную общность тематики, вообще не имеют одинаковых слов, поэтому при обработке «по словам» мера сходства равна нулю. Однако в соответствии с классификатором слова *бескозырка, пилотка, подкладка* относятся к классу одежды, а слова *ленточка, канитель, сукно, шелк* – к классу тканей. Соответственно, отбирая в качестве слов для сравнения существительные, при вычислении сходства «по классам» получаем $\cos \varphi = 0.71$.

На рис. 4 показана зависимость F-меры от порогового уровня для двух способов сегментации – «по словам» и «по классам» (т.е. с предобработкой типа Л+С+ВнКл) на примере текста «Ромм».

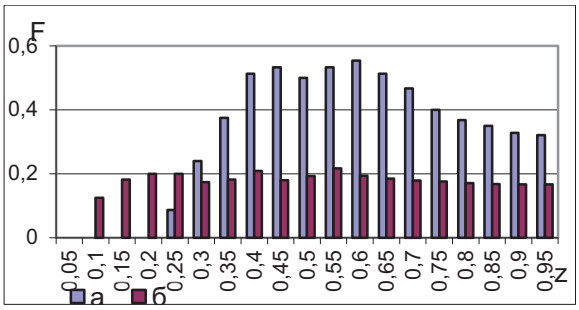

Рис. 4. Сравнение анализа «по классам» и «по словам»

Как видно, F-мера при переходе к обработке «по классам» повышается более чем в два раза.

Уровень обобщения по классам мог варьироваться для достижения наилучшего результата, например, головные уборы могли рассматриваться в качестве отдельного класса, а могли объединяться с другими в общий класс одежды. Сравнение максимальных значений F-меры представлено в таблице 5.

Таблица 5. F-мера при анализе «по словам» и «по классам»

Текст	«Подвод-ники»	«Ромм»	Новостные (репер)
По словам	0.21	0.22	0.60
По классам	0.50	0.7	0.78

Видно, что во всех вариантах сравнение «по классам» дает значительно лучшие результаты. Интересно, что для новостных текстов уже при пороговом уровне z=0.5 все способы дают идентичное сегментирование, т. е. позиции *FN* и *FP* совпадают.

Таким образом, очевидно, что для достижения приемлемых результатов при автоматической сегментации научных текстов нужно оценивать сходство и различие фрагментов текста по агрегатированной лексике, а не по отдельным словам, что позволит значительно эффективнее использовать все типы связей, присутствующие в тексте.

5 Заключение

Проведено исследование специфики применения алгоритмов тематической сегментации к реальным научным текстам на трех языках по единой тематике, причем в корпус включены параллельные фрагменты монографий на языках оригинала, а также их профессиональных переводов. В качестве реперного алгоритма выбран *TextTiling*, использующий локальную информацию о связности между соседними частями текста. Исследовано влияние на качество сегментации текста таких параметров, как размер скользящего окна, величина перекрытия между окнами, пороговый уровень. Определены оптимальные комбинации параметров сегментации для различных языков. На примере русского языка подтверждено, что подключение внешних лексических ресурсов существенно повышает качество сегментации.

Литература

[1] Beeferman, Douglas, Adam Berger, and John Lafferty. Statistical models of text segmentation. Machine Learning, 34(1-3), February 1999.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. Journal of Machine Learning Research, 3, p. 993–1022.

[3] Anja Habacha Chaibi, Marwa Naili, Samia Sammoud. Topic segmentation for textual document written in Arabic language. 18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems - KES2014. Procedia Computer Science 35 (2014), p. 437–446

[4] Choi, F.Y.Y. 2000. Advances in domain independent linear text segmentation. In Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics, p. 26–33.

[5] G. Dias, E. Alves, J.G.P.Lopes. Topic Segmentation Algorithms for Text Summarization and Passage Retrieval: An Exhaustive Evaluation. AAAI'07 Proceedings of the 22nd national conference on Artificial intelligence - Volume 2. p. 1334–1339

[6] Lan Du, Wray Buntine, and Mark Johnson. 2013. Topic Segmentation with a Structured Topic Model. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, p. 190–200, Atlanta, Georgia.

[7] Jacob Eisenstein/ Hierarchical Text Segmentation from Multi-Scale Lexical Cohesion. NAACL '09 Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, p. 353–361

[8] Jacob Eisenstein and Regina Barzilay. 2008. Bayesian Unsupervised Topic Segmentation. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, p. 334–343, Honolulu, Hawaii.

[9] Dominik Flejter, KarolWieloch, Witold Abramowicz. Unsupervised Methods of Topical Text Segmentation for Polish. Balto-Slavonic Natural Language Processing 2007, June 29, 2007, p. 51–58. Prague, June 2007. Association for Computational Linguistics.

[10] M Georgescul, A Clark, and S Armstrong. An analysis of quantitative aspects in the evaluation of thematic segmentation algorithms. SigDIAL'06 Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue, Jan. 2009.

[11] Marti A. Hearst. TextTiling: Segmenting text into multi-paragraph subtopic passages. Computational Linguistics, 23(1), p. 33–64, March 1997.

[12] Marti A. Hearst and Christian Plaunt. 1993. Subtopic structuring for full-length document access. In SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, p. 59–68, New York, NY, USA. ACM Press.

[13] Anna Kazantseva and Stan Szpakowicz. 2011. Linear Text Segmentation Using Affinity Propagation. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, p. 284–293, Edinburgh, Scotland.

[14] Anna Kazantseva, Stan Szpakowicz. Hierarchical Topical Segmentation with Affinity Propagation. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, p. 37–47, Dublin, Ireland.

[15] S. Lamprier, T. Amghar, and B. Levrat. 2008. On evaluation methodologies for text segmentation algorithms. 19th IEEE International Conference on Tools with Artificial Intelligence - Vol.2, Jan.

[16] David YW Lee. Genres, Registers, Text Types, Domains, and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle. Language Learning & Technology. Vol. 5, No. 3, p. 37-72, September 2001.

[17] Hemant Misra, Francëois Yvon, Olivier Cappй, and Joemon M. Jose. 2011. Text segmentation: A topic modeling perspective. Information Processing and Management, 47(4), p.528–544.

[18] POS Tagging Open Xerox. https://open.xerox.com/Services/fst-nlp-tools/Consume/Part%20of%20Speech%20Tagging%20%28Standard%29-178/ Дата обращения – 14.03.15

[19] Ponte, Jay and Bruce Croft. 1997. Text segmentation by topic. In Proceedings of the First European Conference on Research and Advanced Technology for Digitial Libraries.

[20] Jeffrey C. Reynar. 1994. An automatic method of finding topic boundaries. In Proceedings of the 32nd annual meeting on Association for Computational Linguistics, p. 331–333, Morristown, NJ, USA. Association for Computational Linguistics.

[21] Martin Riedl, Chris Biemann. Text Segmentation with Topic Models. JLCL 2012 – Band 27 (1), p. 47–69.

[22] M. Scaiano, D. Inkpen. Getting More from Segmentation Evaluation. 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, p. 362–366, Montreral, Canada, June 3-8, 2012.

[23] Stark, Heather A., 1988. What do paragraph markings do? Discourse Processes 11, p. 275–303.

[24] Xiaojun Wan. On the effectiveness of subwords for lexical cohesion based story segmentation of Chinese broadcast news. Information Sciences 177 (2007), p. 3718–3730.

[25] Na Ye, Jingbo Zhu, Huizhen Wang, Matthew Y. Ma, Bin Zhang. An Improved Model of Dotplotting for Text Segmentation. Journal of Chinese Language and Computing 17 (1), p. 27-40.
[26] Агаркова Н.В., Артемова Г.О., Гусарова Н.Ф. Система поддержки принятия проектных решений для документирования научно-технической информации // Научно-технический вестник информационных технологий, механики и оптики. 2012. № 1 (77). С. 128-134
[27] Аркуша А.Д., Артемова Г.О., Гусарова Н.Ф., Добренко Н.В.Методология моделирования бизнес-процессов и оценка альтернативных подходов с использованием метрик связности и сцепления // Научные труды SWorld. 2013. Т. 8. № 1. С. 81-93
[28] Е.И. Большакова и др. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика. М.: МИЭМ. 2011.
[29] Е.А. Каневский, К.К. Боярский. Семантико-синтаксический анализатор SemSin // Международная конференция по компьютерной лингвистике «Диалог-2012», Бекасово, 30 мая – 3 июня 2012 г. http://www.dialog-21.ru/digest/2012/?type=doc
[30] М П Котюрова. Стилистика научной речи. М.: Академия. 2010.
[31] Смольянина Е.А. Типы связности в научном тексте ((на материале научной статьи на английском языке M. Black “Metaphor”) // Вестник Пермского университета. Российская и зарубежная филология. – 2013. Вып. 4(24). – С. 140–150.
[32] Г.К. Трофимова. Русский язык и культура речи: Курс лекций.– М.: Флинта, Наука, 2004.
[33] В.А. Тузов. Компьютерная семантика русского языка. СПб: Изд-во С.-Петерб. ун-та, 2004

Specifics of Applying Topic Segmentation Algorithms to Scientific Texts

K. Boyarsky, N. Gusarova, N. Dobrenko, E. Kanevsky, N Avdeeva

This paper considers how to apply topic segmentation algorithms to real scientific texts. To study it we used monographs on the same subject written in three languages. The corpus includes several fragments both in the original and in professional translation.

The research is based on the TextTiling algorithm that analyses how tightly adjoining parts of a text cohere. We examined how some parameters (the cutoff rate, the size of moving window and of the shift from one block to the next one) influence the segmentation quality. The optimum combinations of these parameters are defined for several languages. The studies on the Russian language argue that external lexical resources (stop-lists, classifiers, ontologies) notably upgrade the quality of segmentation.