УДК 004.048:519.765
DOI 10.25205/1818-7900-2018-16-2-5-18

Т. В. Батура [1, 2], **С. Е. Стрекалова** [1]

[1] *Новосибирский государственный университет*
ул. Пирогова, 1, Новосибирск, 630090, Россия

[2] *Институт систем информатики им. А. П. Ершова СО РАН*
пр. Академика Лаврентьева, 6, Новосибирск, 630090, Россия

tatiana.v.batura@gmail.com, svetlana.strekalova@gmail.com

ПОДХОД К ПОСТРОЕНИЮ РАСШИРЕННЫХ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ

Представлен новый подход для получения расширенных тематических моделей текстов научных статей на русском языке. Под расширенной моделью понимается тематическая модель, содержащая кроме однословных терминов термины, состоящие из нескольких слов (также называемые многословные термины или ключевые фразы). Такие модели лучше интерпретируются пользователями и точнее описывают предметную область документа, чем модели, состоящие только из униграмм (отдельных слов).

На основе предложенного подхода была разработана система, в результате работы которой для каждого документа предоставляется набор содержащихся в нем тем с указанными вероятностями, ключевыми словами и фразами для каждой темы.

Предложенный в статье подход может быть полезен при построении рекомендательных систем и систем автореферирования.

Ключевые слова: тематические модели, обработка текста, извлечение ключевых слов, извлечение многословных терминов, определение темы текста.

Введение

В современном мире непрерывно производятся огромные объемы электронной информации. Значительную ее часть составляют тексты на естественном языке. В связи с этим становится все более актуальной задача автоматической обработки таких текстов с целью извлечения из них структурированных данных, пригодных для дальнейшего использования в машинном анализе.

Одним из современных инструментов обработки естественного языка являются тематические модели. Тематическое моделирование заключается в построении модели некоторой коллекции текстовых документов. В такой модели каждая тема представляется дискретным распределением вероятностей слов, а документы – дискретным распределением вероятностей тем [1].

Следует обратить внимание на то, что нельзя смешивать понятия тематического моделирования и тематической классификации. Основное отличие состоит в том, что при определении тем текстов отсутствует какая-либо информация о темах: неизвестно ни их количество, ни их содержание (что подразумевается под каждой темой). Для классификации же необходимы априорные знания о структуре классов. В этом смысле процесс тематического моделирования больше похож на кластеризацию, чем на классификацию. Однако ни классификация, ни кластеризация не справляются с синонимией и полисемией, в отличие от тематического

Батура Т. В., *Стрекалова С. Е.* Подход к построению расширенных тематических моделей текстов на русском языке // Вестн. НГУ. Серия: Информационные технологии. 2018. Т. 16, № 2. С. 5–18.

© Т. В. Батура, С. Е. Стрекалова, 2018

моделирования. А, как известно, важнейшим препятствием при создании систем автоматической обработки текстов является лексическая неоднозначность. Так, в тематической модели слова, являющиеся синонимами, с большой вероятностью попадут в одну и ту же тему, так как зачастую они используются в одинаковом контексте. В то же время омонимы (слова одинаковые по написанию, но имеющие разное значение) с большой вероятностью будут отнесены к разным темам, так как обычно контексты их использования не совпадают.

В данной статье описан новый подход для получения расширенных тематических моделей текстов научных статей на русском языке. Под расширенной моделью здесь понимается тематическая модель, содержащая помимо однословных терминов термины, состоящие из нескольких слов (также называемые многословными терминами или ключевыми фразами). Такие модели лучше интерпретируются пользователем и точнее описывают предметную область документа, чем модели, состоящие только из униграмм (отдельных слов).

Основные понятия и постановка задачи построения тематической модели

Тематическое моделирование – построение тематической модели некоторой коллекции текстовых документов. Тематическая модель представляет собой описание коллекции с помощью тематик, использующихся в документах этой коллекции, и определяет слова, относящиеся к каждой из тематик [1].

Вероятностная тематическая модель представляет каждую тему как дискретное распределение на множестве слов, а документ – как дискретное распределение на множестве тем [2].

Одной из разновидностей тематических моделей являются тематические модели, выявляющие ключевые фразы (термины) предметной области. Под ключевой фразой в данной работе подразумевается устойчивая последовательность слов (*n*-грамма), имеющая определенную семантику в контексте заданной предметной области, относящаяся к одной из выявленных в тексте тем и обладающая значительной частотой встречаемости по сравнению с другими *n*-граммами.

Задача построения тематической модели

Пусть задана некоторая коллекция документов D, тогда W – множество всех встречающихся в данной коллекции терминов (слов или *n*-грамм). Каждый документ $d \in D$ представляется в виде последовательности терминов $\left(w_1,...,w_{n_d}\right)$ длиной n_d, $w \in W$, при этом каждый термин может встретиться в документе несколько раз.

Предполагается, что существует некоторое множество тем T, причем каждое вхождение термина w связано с некоторой темой t. Коллекция документов рассматривается как множество троек (d, w, t), выбранных случайно и независимо из дискретного распределения $p(d, w, t)$, заданного на конечном множестве $D \times W \times T$. При этом документы $d \in D$ и термины $w \in W$ являются наблюдаемыми переменными, а тема $t \in T$ – скрытой переменной.

Гипотеза о том, что элементы выборки независимы, эквивалентна предположению «мешка слов»: порядок слов в тексте документа не имеет значения, и тематику можно выявить даже при произвольной перестановке терминов в тексте. В этом случае каждый документ можно представить как подмножество $d \subseteq W$, в котором в соответствие с каждым элементом w_d поставлено количество вхождений n_{d_w} термина w в документ d.

Согласно определению условной вероятности, формуле полной вероятности и гипотезе условной независимости

$$p\left(w|d\right) = \sum_{t \in T} p\left(t|d\right) \cdot p\left(w|t\right).$$

Тогда задача построения тематической коллекции документов заключается в нахождении для известной коллекции D множества всех использующихся в ней тем T, а также для каждого

$d \in D$ по распределению слов по документам $p(w|d)$ восстановить распределения тем в документе $p(t|d)$ и слов по темам $p(w|t)$.

Обзор существующих решений

В настоящее время тематические модели находят применение в самых различных областях. К примеру, в [3] авторы используют тематическое моделирование с помощью алгоритма Latent Dirichlet Allocation (LDA) на отзывах пользователей для создания персонализированных медицинских рекомендаций. В работе [4] авторы используют тематическую модель, включающую в себя авторов, тексты и цитирования, для библиографического анализа. Также тематическое моделирование применяется в обучении: в работе [5] авторы предлагают использовать тематическое моделирование для упрощения оценки учителем письменных работ учеников. Помимо этого, тематическое моделирование применяется для анализа данных социальных сетей [6–8], для многоязычного информационного поиска [9], выявления трендов в новостных потоках или научных публикациях [10], для автоматического присвоения тегов веб-страницам [11], в рекомендательных системах, учитывающих контекст [12], в анализе террористической активности в сети Интернет [13] и мн. др.

Современные требования к тематическим моделям довольно разнообразны. Основное из них заключается в том, что тематические модели должны хорошо поддаваться интерпретации, конечному пользователю должны быть понятны причины выделения определенных тем в тексте и структура самих тем. Эта особенность является главным преимуществом тематических моделей перед набирающими популярность нейронными сетями. Кроме того, часто требуется, чтобы тематические модели учитывали разнородные данные, выявляли динамику тем во времени, автоматически разделяли темы на подтемы, использовали не только отдельные ключевые слова, но и многословные термины и т. д.

Основными подходами к тематическому моделированию являются алгоритмы PLSA (Probabilistic Latent Semantic Analysis, вероятностный латентный семантический анализ), LDA (Latent Dirichlet Allocation, латентное размещение Дирихле) и библиотека ARTM (Additive Regularization for Topic Modeling, аддитивная регуляризация тематических моделей).

PLSA – вероятностная тематическая модель представления текста на естественном языке. Модель называется латентной, так как предполагает введение скрытого (латентного) параметра, являющегося темой. Впервые описана Томасом Хофманном в 1999 г. [14].

LDA – модель, позволяющая объяснять результаты наблюдений с помощью неявных групп, благодаря чему возможно выявление причин сходства некоторых частей данных. Например, если наблюдениями являются слова, собранные в документы, утверждается, что каждый документ представляет собой смесь небольшого количества тем и появление каждого слова связано с одной из тем документа [15].

ARTM является обобщением большого числа алгоритмов тематического моделирования, позволяет комбинировать регуляризаторы, тем самым комбинируя тематические модели. При таком подходе PLSA представляет собой тематическую модель без регуляризаторов, а LDA – тематическую модель, в которой каждая тема сглажена одним и тем же регуляризатором Дирихле. Модель ARTM в предложена 2014 г. [16]. В настоящее время ARTM приобретает все большую популярность благодаря своей универсальности и гибкости настройки параметров моделей.

Многословные термины

Проблема извлечения многословных терминов

Как уже говорилось, основным требованием к тематическим моделям является их интерпретируемость. При этом в большинстве алгоритмов тематического моделирования в качестве терминов используются только слова, а не *n*-граммы. В то же время для человека использование ключевых фраз для обозначения тем может упростить интерпретацию выявленной

темы и разрешить возможную неоднозначность. При этом стоит заметить, что в русском языке задача извлечения ключевых фраз является гораздо более сложной, чем, например, в английском. Это связано с тем, что русский язык флективный, т. е. каждое слово в речи может быть представлено множеством различных словоформ. Обычные алгоритмы извлечения ключевых фраз, основанные на относительной частоте встречаемости *n*-грамм в документах, показывают низкий уровень точности извлечения. Каждую словоформу такие алгоритмы воспринимают как различные термины, и из-за этого частота встречаемости снижается в несколько раз.

Существует несколько основных подходов к решению данной проблемы. Во-первых, для распознавания словоформ можно использовать словари, содержащие все возможные формы слова [17]. Очевидно, что в этом случае точность определения будет высокой для имеющихся в словаре слов. Однако очевидно, что применимость словарных алгоритмов ограничена предметной областью словаря.

Другой подход к этой задаче – использование лексико-синтаксических шаблонов [18; 19]. В [18] описана стратегия распознавания в заданном тексте фрагментов, соответствующих заданному лексико-синтаксическому шаблону, предложен язык записи шаблонов, позволяющий задавать лексические и грамматические свойства входящих в него элементов. В статье [19] приводится описание системы с возможностью ручной настройки видов шаблонов для извлечения словосочетаний с помощью набора морфологических признаков. К сожалению, основными недостатками методов, основанных на шаблонах, является их большая трудоемкость.

Проблему многословных терминов можно обойти, если использовать стемминг (нахождение основы слова) или лемматизацию (приведение слова к его начальной форме). Однако тогда возникает проблема с восстановлением изначальных словосочетаний: так, биграмма будет после стемминга выглядеть как «тематическ моделировании», а после лемматизации – как «тематический моделирование». Очевидно, такие биграммы не могут быть использованы в качестве ключевых фраз в научной статье или на веб-странице, и для дальнейшего использования нужно преобразовать их в изначальное словосочетание.

Предложенное решение проблемы многословных терминов

Для решения проблемы согласования словосочетаний применялись лексико-синтаксические шаблоны. Исследование многословных ключевых терминов, выбранных для статей авторами, позволило составить базовый набор шаблонов. Мы не можем утверждать, что этот набор является полным, так как для составления полного набора шаблонов понадобилось бы привлечь экспертов-лингвистов с целью проведения дополнительного исследования. По этой причине вопрос о полноте набора шаблонов терминов пока остается открытым. Однако предусмотрено возможное расширение набора шаблонов, и в случае увеличения их количества потребуются лишь минимальные изменения в модуле согласования словосочетаний. Выделенные шаблоны удобно записать при помощи логики предикатов первого порядка.

Рассмотрим словарь V – множество слов коллекции документов. Пусть $x, x_1, x_2, ..., x_i, ..., x_n$ – множество прилагательных из V; $y, y_1, y_2, ..., y_i, ..., y_m$ – множество существительных из V. Для морфологических признаков введем следующие обозначения: $z_1 = \{mal, fem, neu\}$ содержит информацию о категории рода (мужской, женский, средний); $z_2 = \{sin, plu\}$ – о категории числа (единственное, множественное); $z_3 = \{nom, gen, dat, acc, ins, pre\}$ – о категории падежа (именительный, родительный, дательный, винительный, творительный, предложный). Далее введем четырехместные предикаты $A(x, z_1, z_2, z_3)$ для прилагательных и $N(y, z_1, z_2, z_3)$ для существительных. Теперь шаблоны многословных терминов можно записать в виде формул исчисления предикатов, т. е. в случае согласованных словосочетаний будут истинны следующие шаблоны.

1. $MWE_1(x, y): A(x, z_1, z_2, nom) \wedge N(y, z_1, z_2, nom)$.

Например, «линейное уравнение».

2. $MWE_2(y_1, y_2): N(y_1, z_1^1, z_2^1, nom) \wedge N(y_2, z_1^2, z_2^2, gen)$.

Например, «разработка системы».

3. $MWE_3(y_1, x, y_2): N(y_1, z_1^1, z_2^1, nom) \wedge A(x, z_1^2, z_2^2, gen) \wedge N(y_2, z_1^2, z_2^2, gen)$.

Например, «гипотеза условной независимости».

4. $MWE_4(x_1, x_2, y): A(x_1, z_1, z_2, nom) \wedge A(x_2, z_1, z_2, nom) \wedge N(y, z_1, z_2, nom)$.

Например, «вероятностная тематическая модель».

5. $MWE_5(y_1, y_2, y_3): N(y_1, z_1^1, z_2^1, nom) \wedge N(y_2, z_1^2, z_2^2, gen) \wedge N(y_3, z_1^3, z_2^3, gen)$.

Например, «определение тематики документа».

6. $MWE_6(x, y_1, y_2): A(x, z_1^1, z_2^1, nom) \wedge N(y_1, z_1^1, z_2^1, nom) \wedge N(y_2, z_1^2, z_2^2, gen)$.

Например, «общая теория относительности».

7. $MWE_7(y_1, y_2): N(y_1, z_1^1, z_2^1, nom) \wedge N(y_2, z_1^2, z_2^2, ins)$.

Например, «умножение столбиком».

8. $MWE_8(y_1, y_2, y_3): N(y_1, z_1^1, z_2^1, nom) \wedge N(y_2, z_1^2, z_2^2, ins) \wedge N(y_3, z_1^3, z_2^3, gen)$.

Например, «решение методом прогонки».

Обобщение шаблонов 1 и 4 можно переписать в виде

$$\bigwedge_{i=1}^{n} A(x_i, z_1^i, z_2^i, nom) \wedge N(y, z_1, z_2, nom).$$

Обобщение шаблонов 2 и 5 запишем в виде

$$N(y_1, z_1^1, z_2^1, nom) \wedge \bigwedge_{j=2}^{m} N(y_j, z_1^j, z_2^j, gen).$$

Был разработан модуль согласования словосочетаний на основе вышеперечисленных шаблонов, использующий для извлечения морфологической информации программу Mystem [1]. На вход модулю подаются лемматизированные словосочетания, которые сопоставляются с каждым шаблоном из набора. После определения требуемого шаблона словосочетание приводится в согласованный вид путем преобразования зависимых слов в форму, обусловленную формой главного слова и видом связи в словосочетании.

Данный модуль показывает приемлемые результаты, а набор модулей покрывает значительную часть используемых в качестве ключевых фраз многословных терминов. Для улучшения результатов работы можно использовать как расширение набора шаблонов, так и дополнительные способы согласования.

Основным недостатком текущей версии модуля является невозможность построения словосочетаний, в которых существительные находятся во множественном числе. Для решения данной проблемы в дальнейшем планируется использовать модуль поиска начальной формы из базового подхода, модифицировав его для поиска всех вариантов заданного лемматизированного словосочетания, а затем применить морфологический анализатор для определения нужного числа существительного.

Также к недостаткам модуля можно отнести несовершенство изменения формы слов с точки зрения лингвистики. В русском языке множество исключений, например, слова, оканчивающиеся на -мя (время, пламя и др.) не относятся к первому, второму или третьему склонению, а склоняются смешанным способом, причем при склонении к корню добавляется -ен (времени, пламени). Этот вид исключений был учтен в разработанной программе, однако, чтобы учесть все варианты исключений, встречающихся в русском языке, потребуется участие эксперта-лингвиста.

[1] MyStem – Технологии Яндекса. URL: https://tech.yandex.ru/mystem/

Разработка системы

Предобработка текста

Для лемматизации текста и построения морфологического словаря коллекции документов используется программа Mystem. Программа лемматизирует слова, используя анализ контекста для снятия лексической неоднозначности, а также предоставляет морфологическую информацию (часть речи, род, число, падеж, склонение и др.) для каждого слова. Программа распространяется бесплатно для некоммерческого использования.

Тематическое моделирование

Выбор методов тематического моделирования объясняется наличием определенных особенностей. Для сравнения некоторые из них приведены в табл. 1.

Таблица 1

Сравнение методов тематического моделирования

Название метода	Увеличение количества параметров модели с ростом числа документов	Применимость к большим наборам данных	Использование многословных терминов	Единственность и устойчивость решения
PLSA	да, есть линейная зависимость	нет	нет	нет
LDA	нет	да	нет	нет
ARTM	нет	да	нет	да
ARTM + Turbotopic (предлагаемый)	**нет**	**да**	**да**	**да**

Также для выбора базового алгоритма построения униграммных тематических моделей был проведен ряд экспериментов. Была подготовлена коллекция текстов научных статей на русском языке на основе выложенных в открытом доступе архивов журналов «Программные продукты и системы» [2], «Сибирский психологический журнал» [3] и «Cloud of Science» [4]. Статьи очищены от формул, таблиц, рисунков и библиографических ссылок, аннотация и ключевые слова были удалены. Размер коллекции составляет более двухсот шестидесяти текстов.

Для оценки результатов были выбраны следующие метрики, реализованные в библиотеке BigARTM и описанные в работе [20]: перплексия, разреженность матриц Φ и Θ, доля фоновых слов, мощность ядер тем, чистота ядер тем, контрастность ядер тем.

Первоначальные эксперименты выявили, что LDA показывает значительно худшие результаты перплексии по сравнению с PLSA и ARTM. В связи с этим дальнейшее сравнение проводилось только для двух последних алгоритмов при числе проходов по коллекции 100. Результаты представлены в табл. 2.

[2] http://www.swsys.ru/
[3] http://journals.tsu.ru/psychology/
[4] https://cloudofscience.ru/

Таблица 2

Сравнение алгоритмов PLSA и ARTM

Метрика	PLSA	ARTM
Перплексия	754.784	751.888
Разреженность матрицы Φ	0.769	0.769
Разреженность матрицы Θ	0.000	0.635
Доля фоновых слов	0.059	0.050
Средняя чистота ядер тем	0.370	0.364
Средняя контрастность ядер тем	0.787	0.788
Средняя мощность ядер тем	2085.000	2085.600

По результатам эксперимента, приведенным в табл. 2, можно увидеть, что ARTM показывает аналогичные либо лучшие результаты по сравнению с PLSA для всех метрик, за исключением средней чистоты ядер, где ухудшение незначительно. В совокупности с особенностями алгоритмов, приведенными в табл. 1, было принято решение использовать в качестве алгоритма построения униграммных тематических моделей алгоритм ARTM в реализации библиотеки BigARTM [16].

Извлечение ключевых фраз

Для извлечения многословных терминов из текстов используется адаптированный алгоритм извлечения ключевых слов Turbotopics. Суть оригинального алгоритма Turbotopics, описанного в работе [21], обобщенно состоит в следующем.

Первоначально строится униграммная модель текста при помощи алгоритма LDA. Затем производится расширение модели многословными терминами. Для каждого отдельного ключевого слова, полученного при помощи LDA, или уже добавленной фразы w осуществляется проверка в исходном тексте на наличие соседних слов u, которые с высокой вероятностью будут предшествовать w в тексте или следовать за ним. Пара таких найденных слов (u,v) или (v,u) считается многословным термином и добавляется к списку ключевых фраз. Данный алгоритм был разработан для применения в текстах на английском языке на основе алгоритма построения тематических моделей LDA и показал довольно хорошие результаты. Поэтому в данной работе он был адаптирован для работы с русскими текстами с использованием алгоритма ARTM библиотеки BigARTM.

Для определения списка ключевых слов для каждого документа изначально предполагалось использовать список наиболее часто встречающихся терминов (одно- и многословных) для каждой темы, к которой относится данный документ. Однако этот подход привел к тому, что из документа извлекались ключевые слова темы, а не самой статьи: для различных документов списки ключевых слов были очень похожи, а термины, которые должны быть ключевыми исходя из текста статьи, не попадали в список из-за низкой частоты встречаемости. Для решения данной проблемы было предложено использовать TF-IDF – статистическую меру, оценивающую важность каждого слова для документа, в котором оно встречается [22]. Наибольшее значение TF-IDF будут иметь слова, которые часто встречаются в данном документе, но редко встречаются в остальных документах коллекции.

Общий вид системы

В рамках исследования была разработана система, позволяющая строить расширенные тематические модели, включающие многословные термины, для коллекций научных статей на русском языке. Система написана на языке Python 3 с использованием библиотеки BigARTM. Используемые в системе алгоритмы из этой библиотеки были настроены таким

образом, чтобы получить оптимальные результаты относительно различных метрик (перплексия, разреженность и др.) при использовании текстов научных статей на русском языке. Обобщенная схема работы системы представлена ниже. Далее приведено подробное описание процесса построения расширенной тематической модели и извлечения ключевых фраз разработанной системой.

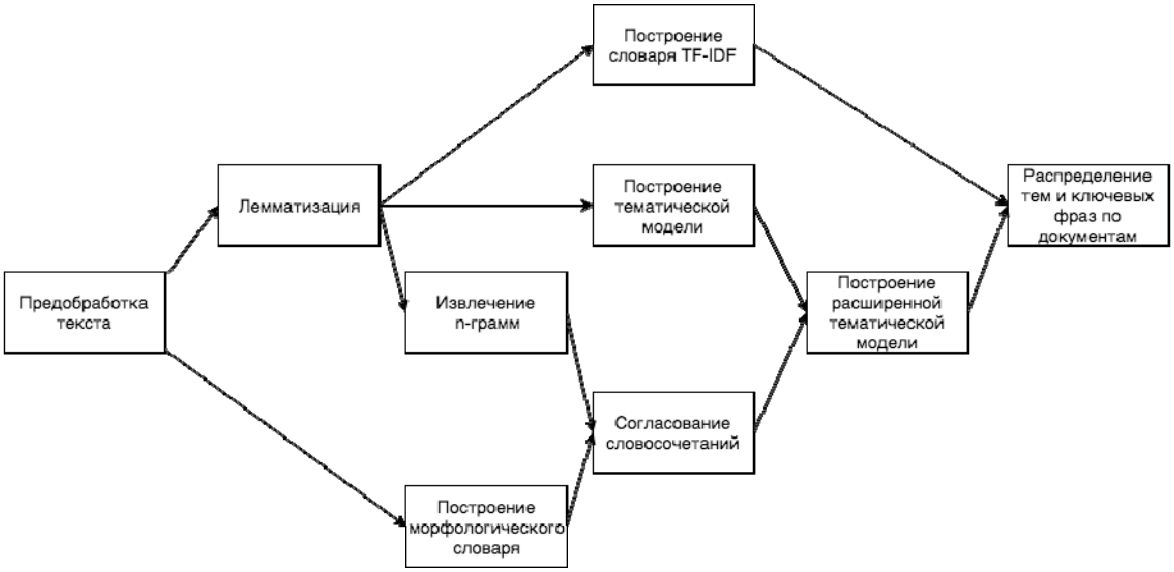

Схема работы системы

Опишем схему работы системы как последовательность шагов.

Шаг 0. На вход системе подается коллекция документов в формате .txt. Каждый документ должен быть представлен одним файлом, все документы помещены в одну директорию, путь к которой передается программе в качестве параметра.

Шаг 1. В модуле предобработки текста каждый документ очищается от специальных символов (отличных от кириллических и латинских букв), из документа удаляются стоп-слова, все слова приводятся к нижнему регистру. Далее строится корпус коллекции в формате последовательного Vowpal Wabbit.

Шаг 2. Производится вызов программы Mystem, на вход которой подается файл с построенным на предыдущем этапе работы корпусом. Результатом работы является файл лемматизированного корпуса (формат, аналогичный полученному ранее корпусу, только каждое слово заменено его начальной формой), а также файл морфологического словаря, где каждой строке соответствует слово и описывающая его морфологическая информация.

Шаг 3. На лемматизированном корпусе производится поиск ключевых слов и *n*-грамм с помощью алгоритма Turbotopics.

Шаг 4. Найденные алгоритмом Turbotopics *n*-граммы преобразуются из лемматизированного вида в согласованный с использованием шаблонов, описанных выше, и морфологического словаря, полученного на шаге 2.

Шаг 5. Для лемматизированного корпуса строится тематическая модель коллекции документов с использованием алгоритма ARTM. Параметры алгоритма можно подобрать автоматически или использовать заранее вычисленные (так как подбор параметров – задача весьма трудоемкая и занимает значительное время).

Шаг 6. Полученная на шаге 5 тематическая модель расширяется с помощью многословных терминов, извлеченных из коллекции на шаге 3 и согласованных на шаге 4.

Шаг 7. Для каждого документа строится словарь TF-IDF: с каждым словом в лемматизированном документе сопоставляется значение меры TF-IDF. Слова в словаре сортируются по убыванию значения меры.

Шаг 8. На основе матрицы распределения тем по документам с каждым документом сопоставляется набор присутствующих в нем тем и их вероятностей (учитываются только темы, вероятность появления которых в данном документе превышает порог $\delta = \frac{1}{N_t}$, где N_t – количество тем в модели).

После этого сравниваются два множества: первые N_1 слов из отсортированного словаря TF-IDF и первые N_2 слов и словосочетаний для каждой темы, отсортированных по вероятности встретить этот термин в документе. Итоговыми ключевыми словами для темы документа будет пересечение этих множеств. N_1 и N_2 могут настраиваться; по умолчанию эти значения равны 100 и 300 соответственно. Такие значения параметров были подобраны эмпирическим путем, чтобы каждому документу в среднем соответствовало порядка 5–10 ключевых слов и фраз.

Результатом работы программы является текстовый файл, содержащий следующую информацию:

- название исходного документа;
- список тем, для каждой из которых указана вероятность содержания ее в тексте как десятичная дробь от 0 до 1;
- список ключевых слов и фраз для каждой темы.

Также для пользователя доступен файл с описанием тем, где с каждой темой сопоставлено множество слов и словосочетаний с наибольшей вероятностью для этой темы.

Полученные результаты

Поскольку невозможно автоматически оценить интерпретируемость тем и приемлемость извлеченных ключевых фраз, результаты были оценены вручную. Далее приведены несколько примеров работы алгоритма для различных публикаций разной направленности. Некоторые из наиболее частотных слов и фраз для первых пяти тем расширенной тематической модели коллекции, представлены в табл. 3.

Таблица 3

Расширенная тематическая модель коллекции научных статей

Тема	Расширенная тематическая модель
Тема 1	'алгоритм', 'решение', 'задача', 'значение', 'вершина', 'значение параметра', 'время распознавания', 'класс объекта', 'обработка информации', 'алгоритм поиска', 'вершина графа', 'изображение объекта', 'граница решения', 'задача поиска', 'граф решения'
Тема 2	'метод', 'данные', 'алгоритм', 'классификация', 'текст', 'слово', 'классификатор', 'обучение', 'значение параметра', 'класс объекта', 'множество признака', 'представление текста', 'процесс обучения', 'метод классификации', 'построение модели', 'задача классификации', 'качество классификации', 'обучение классификатора', 'классификация текста'
Тема 3	'человек', 'ребенок', 'психологический', 'группа', 'отношение', 'стратегия восприятия', 'процесс формирования', 'образ мира', 'группа испытуемая', 'уровень развития', 'респондент группы', 'развитие ребенка'
Тема 4	'система', 'управление', 'процесс', 'модель', 'требование', 'разработка', 'система управления', 'орган управления', 'процесс разработки', 'модель прогнозирования', 'критерий эффективности проекта', 'этап прогнозирования', 'критерий эффективности', 'эффективность проекта'
Тема 5	'исследование', 'отношение', 'испытуемый', 'элемент', 'диагностический', 'результат исследования', 'значение параметра', 'удовлетворенность отношения', 'процесс формирования', 'поиск решения', 'вид деятельности', 'группа испытуемая', 'удовлетворенность брака', 'формирование религиозности’

По представленным в табл. 3 результатам можно отметить, что темы из разных предметных областей (технические науки и психология) очень хорошо различимы в тематической модели. При этом граница между более узкими темами не настолько четкая: если тема 4 довольно хорошо интерпретируется как отдельная предметная область, связанная с управлением проектами и процессом разработки, темы 1 и 2 связаны с классификацией и распознаванием, а темы 3 и 5 – с психологической диагностикой. При этом важно заметить, что в теме 5 многословные термины («удовлетворенность отношения», «формирование религиозности» и т. д.) улучшают интерпретируемость темы как относящуюся к психологии, тогда как термины «исследование», «испытуемый» являются более общими.

В табл. 4 представлены извлеченные программой ключевые слова и фразы для нескольких научных публикаций.

Таблица 4

Ключевые слова и фразы

№	Название статьи	Извлеченные ключевые слова и фразы
1	Алгоритм детектирования объектов на фотоснимках с низким качеством изображения	объект, класс, изображение, набор, автокодировщик, обучение, объект, класс, набор, изображение, слой, пиксел
2	Проектирование интерфейса программного обеспечения с использованием элементов искусственного интеллекта	программный, пользователь, система управления, уровень развития, нечеткий, интерфейс, характеристика, эксперт, система управления
3	Родительское отношение как фактор формирования религиозности личности	ребенок, отношение, родитель, формирование, религиозность, религиозный, религия, семья, родительский, решение задачи
4	Прогнозирование платежеспособности клиентов банка на основе методов машинного обучения и марковских цепей	прогнозирование, состояние, клиент, классификатор, ак, заемщик, решение задачи, дерево решения
5	Разработка системы хранения ансамблей нейросетевых моделей	данные, модель, набор, ансамбль, ряд, преобразование, хранение, нейросетевой, оценка качества, процесс формирования, классификация текста

Можно утверждать, что извлеченные ключевые слова и фразы соответствуют содержанию статей и хорошо определяют предметную область исследований. При этом можно заметить, что в некоторых случаях они дают большее представление о содержании публикации, чем ее название: например, ключевая фраза «дерево решения» дает понять, что в качестве алгоритма машинного обучения в четвертой статье использовались деревья решений, а ключевая фраза «классификация текста» в статье 5 указывает, что ансамбли нейросетевых моделей здесь использовались для классификации текста (а не только изображений, например).

Заключение

Тематические модели позволяют автоматически систематизировать большие коллекции текстовых документов на естественном языке, повышают эффективность информационного поиска. В ходе данного исследования была разработана система построения тематических моделей и извлечения ключевых слов и фраз для текстов научных статей на русском языке. Для проведения экспериментов была подготовлена коллекция «очищенных» текстов научных статей на русском языке из размещенных в открытом доступе журналов [5].

[5] Коллекция текстов доступна по ссылке: https://github.com/Serenitas/topic-modeller/.

Разработанная система способна строить расширенные тематические модели, включающие, помимо униграмм, словосочетания в согласованном виде. Для каждого документа предоставляется набор содержащихся в нем тем с указанными вероятностями и ключевыми словами и фразами для каждой темы.

Благодаря расширению тематической модели многословными терминами темы хорошо интерпретируются. Извлекаемые ключевые слова и фразы соответствуют содержанию документа.

Предложенный в статье подход может быть полезен при построении рекомендательных систем и систем автореферирования.

Список литературы

1. *Коршунов А., Гомзин А.* Тематическое моделирование текстов на естественном языке // Тр. Ин-та системного программирования РАН. 2012. С. 215–242.

2. *Воронцов К. В.* Вероятностное тематическое моделирование. 2013. URL: http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf

3. *Yin Zhang, Min Chen, Dijiang Huang, Di Wu, Yong Li.* iDoctor: Personalized and professionalized medical recommendations based on hybrid matrix factorization // Future Generation Computer Systems. 2017. Vol. 66. P. 30–35.

4. *Kar Wai Lim, Wray Buntine.* Bibliographic Analysis with the Citation Network Topic Model // JMLR: Workshop and Conference Proceedings. 2014. Vol. 39. P. 142–158.

5. *Ye Chen, Bei Yu, Xuewei Zhang, Yihan Yu.* Topic modeling for evaluating students' reflective writing: a case study of pre-service teachers' journals // LAK '16 Proceedings of the Sixth International Conference on Learning Analytics & Knowledge. 2016. P. 1–5.

6. *Zhao X. W., Wang J., He Y., Nie J.-Y., Li X.* Originator or propagator: Incorporating social role theory into topic models for Twitter content analysis // Proceedings of the 22nd ACM International Conference on Conference on Information and Knowledge Management. CIKM '13. New York, NY, USA: ACM, 2013. P. 1649–1654.

7. *Varshney D., Kumar S., Gupta V.* Modeling information diffusion in social networks using latent topic information // Intelligent Computing Theory / Ed. by D.-S. Huang, V. Bevilacqua, P. Premaratne. Springer International Publishing, 2014. Vol. 8588 of Lecture Notes in Computer Science. P. 137–148.

8. *Pinto J. C. L., Chahed T.* Modeling multi-topic information diffusion in social networks using latent Dirichlet allocation and Hawkes processes // Tenth International Conference on Signal-Image Technology & Internet-Based Systems. 2014. P. 339–346.

9. *Vulic I., De Smet W., Tang J., Moens M.-F.* Probabilistic topic modeling in multilingual settings: an overview of its methodology and applications // Information Processing & Management. 2015. Vol. 51, no. 1. P. 111–147.

10. *Cui W., Liu S., Tan L., Shi C., Song Y., Gao Z., Qu H., Tong X.* TextFlow: Towards better understanding of evolving topics in text // IEEE transactions on visualization and computer graphics. 2011. Vol. 17, no. 12. P. 2412–2421.

11. *Allahyari M., Kochut K. J.* Semantic Tagging Using Topic Models Exploiting Wikipedia Category Network // IEEE Tenth International Conference on Semantic Computing (ICSC). 2016.

12. *Allahyari M., Kochut K.* Semantic Context-Aware Recommendation via Topic Models Leveraging Linked Open Data // International Conference on Web Information Systems Engineering. WISE 2016. Lecture Notes in Computer Science. Vol. 10041. P. 263–277.

13. *Золотарев О. В., Шарнин М. М., Клименко С. В.* Семантический подход к анализу террористической активности в сети Интернет на основе методов тематического моделирования // Вестн. Российского нового университета. Серия: Сложные системы: модели, анализ и управление. 2016. № 3. С. 64–71.

14. *Hofmann T.* Probabilistic Latent Semantic Indexing // Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval (SIGIR-99). 1999. P. 289–296.

15. *Blei D. M., Ng A. Y., Jordan M. I.* Latent Dirichlet allocation // Journal of Machine Learning Research. 2003. No. 3. P. 993–1022.

16. *Воронцов К. В., Фрей А. И., Ромов П. А., Янина А. О., Суворова М. А., Апишев М. А.* BigARTM: библиотека с открытым кодом для тематического моделирования больших текстовых коллекций. 2014. URL: http://docplayer.ru/27022431-Bigartm-biblioteka-s-otkrytym-kodom-dlya-tematicheskogo-modelirovaniya-bolshih-tekstovyh-kollekciy.html

17. *Кипяткова И. С., Карпов А. А.* Аналитический обзор систем распознавания русской речи с большим словарем // Тр. СПИИРАН. 2010. Вып. 12. С. 7–20.

18. *Большакова Е. И., Баева Н. В., Бордаченкова Е. А., Васильева Н. Э., Морозов С. С.* Лексико-синтаксические шаблоны в задачах автоматической обработки текста // Компьютерная лингвистика и интеллектуальные технологии: Тр. Междунар. конф. «Диалог 2007». М.: Изд-во РГГУ, 2007. С. 70–75.

19. *Загорулько М. Ю., Сидорова Е. А.* Система извлечения предметной терминологии из текста на основе лексико-синтаксических шаблонов // Тр. XIII Междунар. конф. «Проблемы управления и моделирования в сложных системах» / Под ред. Е. А. Федосова, Н. А. Кузнецова, В. А. Виттиха. 2011. С. 506–511.

20. *Воронцов К. В., Потапенко А. А.* Регуляризация вероятностных тематических моделей для повышения интерпретируемости и определения числа тем // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Междунар. конф. «Диалог» М.: Изд-во РГГУ, 2014. Вып. 13 (20). С. 676–687.

21. *Blei D. M., Lafferty J. D.* Visualizing Topics with Multi-Word Expressions // Semantic Scholar. 2009. URL: https://arxiv.org/pdf/0907.1013.pdf

22. *Leskovec J., Rajaraman A., Ullman J. D.* Mining of Massive Datasets. 2014. 513 p.

Материал поступил в редколлегию 03.03.2018

T. V. Batura [1, 2], **S. E. Strekalova** [1]

[1] *Novosibirsk State University*
1 Pirogov Str., Novosibirsk, 630090, Russian Federation

[2] *A. P. Ershov Institute of Informatics Systems SB RAS*
6 Academician Lavrentiev Ave., Novosibirsk, 630090, Russian Federation

tatiana.v.batura@gmail.com, svetlana.strekalova@gmail.com

AN APPROACH TO BUILDING EXTENDED TOPIC MODELS OF RUSSIAN TEXTS

A new approach to building extended topic models of Russian scientific texts is described in this article. An extended topic model is a topic model containing not only one-word terms, but also multiword terms (key phrases). Such models are better interpreted for the user and more accurately describe the subject area of the document than models consisting only of unigrams (separate words).

On the basis of the proposed approach, a system was developed which, as a result of the work, provides for each document a set of topics with probabilities, key words and phrases for each topic.

The approach proposed in the article can be useful for development of recommendation systems and summarization systems.

Keywords: topic models, text processing, keyword extraction, multiword term extraction, topic detection.

References

1. Korshunov A., Gomzin A. Topic modelling of natural languge texts. *Proceedings of the Institute for System Programming of the RAS*, 2012, p. 215–242. (in Russ.)

2. Vorontsov K. V. Probabilistic topic modeling. 2013. URL: http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf (in Russ.)

3. Yin Zhang, Min Chen, Dijiang Huang, Di Wu, Yong Li iDoctor: Personalized and professionalized medical recommendations based on hybrid matrix factorization. *Future Generation Computer Systems*, 2017, vol. 66, p. 30–35.

4. Kar Wai Lim, Wray Buntine, Bibliographic Analysis with the Citation Network Topic Model. *JMLR: Workshop and Conference Proceedings*, 2014, vol. 39, p. 142–158.

5. Ye Chen, Bei Yu, Xuewei Zhang, Yihan Yu Topic modeling for evaluating students' reflective writing: a case study of pre-service teachers' journals. *LAK '16 Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, 2016, p. 1–5.

6. Zhao X. W., Wang J., He Y., Nie J.-Y., Li X. Originator or propagator?: Incorporating social role theory into topic models for Twitter content analysis. *Proceedings of the 22nd ACM International Conference on Conference on Information and Knowledge Management. CIKM '13*. New York, NY, USA, ACM, 2013, p. 1649–1654.

7. Varshney D., Kumar S., Gupta V. Modeling information diffusion in social networks using latent topic information. *Intelligent Computing Theory*. Ed. by D.-S. Huang, V. Bevilacqua, P. Premaratne. Springer International Publishing, 2014, vol. 8588 of Lecture Notes in Computer Science, p. 137–148.

8. Pinto J. C. L., Chahed T. Modeling multi-topic information diffusion in social networks using latent Dirichlet allocation and Hawkes processes. *Tenth International Conference on Signal-Image Technology & Internet-Based Systems*, 2014, p. 339–346.

9. Vulic I., De Smet W., Tang J., Moens M.-F. Probabilistic topic modeling in multilingual settings: an overview of its methodology and applications. *Information Processing & Management*, 2015, vol. 51, no. 1, p. 111–147.

10. Cui W., Liu S., Tan L., Shi C., Song Y., Gao Z., Qu H., Tong X. TextFlow: Towards better understanding of evolving topics in text. *IEEE transactions on visualization and computer graphics,* 2011, vol. 17, no. 12, p. 2412–2421.

11. Allahyari M., Kochut K.J. Semantic Tagging Using Topic Models Exploiting Wikipedia Category Network. *IEEE Tenth International Conference on Semantic Computing* (*ICSC*), 2016.

12. Allahyari M., Kochut K. Semantic Context-Aware Recommendation via Topic Models Leveraging Linked Open Data. *International Conference on Web Information Systems Engineering. WISE. Lecture Notes in Computer Science*, 2016, vol. 10041, p. 263–277.

13. Zolotarev O. V., Sharnin M. M., Klimenko S. V. Semantic approach for terroristic activity analysis in the Internet based on topic modelling methods. *Russian New University Bulletin. Series: Complex systems: models, analysis and control*, 2016, vol. 3, p. 64–71. (in Russ.)

14. Hofmann T. Probabilistic Latent Semantic Indexing. *Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR-99*), 1999, p. 289–296.

15. Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet allocation. *Journal of Machine Learning Research*. 2003, no. 3, p. 993–1022.

16. Vorontsov K. V., Frey A. I., Romov P. A., Yanina A. O., Suvorova M. A., Apishev M. A. BigARTM: Open Source Library for Topic Modeling of Large Text Collections. 2014. URL: http://docplayer.ru/27022431-Bigartm-biblioteka-s-otkrytym-kodom-dlya-tematicheskogo-modelirovaniya-bolshih-tekstovyh-kollekciy.html (in Russ.)

17. Kipyatkova I. S., Karpov A. A. Analytical review of recognition systems for Russian language with large dictionary. *Proceedings of SPIIRAS*, 2010, vol. 12, p. 7–20. (in Russ.)

18. Bolshakova E. I., Baeva N. V., Bordachenkova E. A., Vasilyeva N. E., Morozov S. S. Lexico-syntactic templates in natural language processing. *Computational linguistics and intellectual technologies: Proceedings of the international conference "Dialogue 2007"*. Moscow, RSUH, 2007, p. 70–75. (in Russ.)

19. Zagorulko M. Yu., Sidorova E. A. System of extraction of subject terminology from text based on lexico-syntactic templates. *Proceedings of XIII International conference "Problems of control and modelling in complex systems".* Eds. E. A. Fedosova, N. A. Kusnetsova, V. A. Vittikh. 2011, p. 506–511. (in Russ.)

20. Vorontsov K. V., Potapenko A. A. Regularization of Probabilistic Topic Models to Improve Interpretability and Determine the Number of Topics. *Computational linguistics and intellectual technologies: Proceedings of the annual international conference "Dialogue"*. Moscow, RSUH, 2014, vol. 13 (20), p. 676–687. (in Russ.)

21. Blei D. M., Lafferty J. D. Visualizing Topics with Multi-Word Expressions. *Semantic Scholar*, 2009. URL: https://arxiv.org/pdf/0907.1013.pdf

22. Leskovec J., Rajaraman A., Ullman J. D. Mining of Massive Datasets, 2014, 513 p.

For citation:

Batura T. V., Strekalova S. E. An Approach to Building Extended Topic Models of Russian Texts. *Vestnik NSU. Series*: *Information Technologies*, 2018, vol. 16, no. 2, p. 5–18. (in Russ.)

DOI 10.25205/1818-7900-2018-16-2-5-18