Тематическое моделирование русскоязычных текстов с опорой на леммы и лексические конструкции

А.Г. Седова, О.А. Митрофанова

Санкт-Петербургский государственный университет

agsedova@mail.ru, o.mitrofanova@spbu.ru

Аннотация

Исследование посвящено усовершенствованию методов вероятностного тематического моделирования, направленных на выявление скрытых взаимосвязей между словами, документами и темами в текстовых коллекциях. В большинстве тематических моделей темы представлены исключительно униграммами, что в некоторых случаях влечет за собой ухудшение точности и повышает сложность содержательной интерпретации выделяемых тем. Нами предложен новый алгоритм на основе метода LDA, позволяющий автоматически выделять в корпусе словосочетания, состоящие из двух слов, и добавлять их в тематические модели. В статье изложена работа алгоритма и приведены результаты его применения в автоматической обработке корпусов специальных текстов.

Ключевые слова: тематическое моделирование, LDA, биграммы.

1. Введение

В настоящее время в корпусной лингвистике активно развивается направление вероятностного тематического моделирования, опирающегося на статистическую обработку текстов. Тема (topic, latent topic), или скрытый паттерн (hidden pattern) определяется дискретным вероятностным распределением в пространстве слов заданного словаря [1]. Формально говоря, темы являются результатом бикластеризации, то есть одновременной кластеризации слов и документов с точки зрения их семантической близости. Тематическим моделированием при этом называется восстановление вероятностных распределений всех тем в тексте, рассматриваемом как случайная независимая выборка слов («мешок слов»), порожденная некоторыми темами. При этом достаточно небольшое число тем может породить документ, состоящий из большого количества слов. Порядок тем при различных запусках алгоритма может варьироваться, что обусловлено свойством неупорядоченности, или перестановочности (exchangeability) тем.

Тематическая модель описывает связи между словами и темами, документами и темами с помощью смеси дискретных распределений. Таким образом, тематическая модель выступает как средство обобщения и систематизации информации из больших текстовых коллекций, эта модель позволяет выявить скрытые структуры и неявные зависимости в данных. Тематическое моделирование находит широкое применение в решении задач информационного поиска, автоматического аннотирования и индексирования документов, пополнения тональных словарей, поиска классов переводных эквивалентов, определения сопоставимости текстов в многоязычных корпусах и т.д. [2].

Одним из наиболее распространенных методов построения тематических моделей является метод латентного размещения Дирихле (Latent Dirichlet Allocation, LDA) [3]. Данный алгоритм опирается на априорное распределение Дирихле и использует в своей работе модель «мешка слов» (bag-of-words) — модель для анализа текстов, которая учитывает только частоту слов, но не их порядок; данная модель хорошо подходит для тематического моделирования, поскольку она позволяет обнаруживать неявные взаимосвязи между словами с учетом полисемии. Метод LDA осуществляет мягкую кластеризацию и предполагает, что каждое слово в документе порождено некоторой латентной темой, определяющейся вероятностным распределением на множестве всех слов текста. В нашем исследовании мы пользовались именно этим алгоритмом.

На вход практически любой тематической модели поступает корпус текстов, каждый из которых является отдельным документом. Результатом работы модели является список тем, выявленных в корпусе и представленных списком первых, наиболее характерных *n* слов для каждой рассматриваемой темы. В базовых алгоритмах тематического моделирования темы представлены исключительно униграммами (например, *система, частота, время, измерение, объект, скорость, дальность, метод, параметр, обработка* и т.д.; *свойство, расстояние, отраженный, излучать, отражение, радиолокационный, земля, объект, мощность, отражать* и т.д). В первую очередь это происходит вследствие использования модели «мешка слов», не учитывающей линейной зависимости между словами внутри предложения. Зачастую это влечет за собой ухудшение точности и повышает сложность содержательной интерпретации выделяемых тем, особенно в случае некомпозиционных словосочетаний, значение которых не сводится к сумме значений входящих в них слов: например, «*железная дорога*» не сводится к значению слов «*железная*» и «*дорога*» соответственно [4]. Таким образом, добавление в темы расширение тем за счет *n*-грамм представляет собой актуальную исследовательскую задачу.

В последнее время было предложено несколько подходов к решению данной проблемы [5, 6]. Однако многие из них снижают качество модели или же излишне усложняют ее [4]. В данной работе была предпринята попытка предложить новый метод, который бы действительно упрощал интерпретацию тем и повышал их точность, оставаясь при этом понятным и простым для реализации.

2. Алгоритм добавления биграмм в тематические модели

2.1. Возможные решения задачи

В зависимости от используемого алгоритма предлагаемые для решения данной задачи алгоритмы делятся на две группы:

- алгоритмы, представляющие собой унифицированную тематическую модель, в рамках которой словосочетания выделяются в тексте одновременно с темами;
- алгоритмы, выделяющие многословные выражения на этапе предобработки текста.

Большинство существующих алгоритмов относятся к первой группе. Одним из них является, например, биграммная тематическая модель (bigram topic model). Данная модель является иерархической порождающей моделью, и при её работе в качестве основополагающего используется предположение о том, что появление слова w_i зависит исключительно от слова w_{i-1}, стоящего непосредственно перед интересующим нас словом:

$$P(w_i|w_{i-1}) = \frac{n_{ii-1}+\delta_{w_i}}{n_{i-1}+\delta_0}, \delta_0 = \sum_i \delta_{w_i},$$

где $\{\delta_{w_i}\}$ — гиперпараметры модели, n_{i-1} — частотность слова w_{i-1}, n_{ii-1} — частотность словосочетания $w_i w_{i-1}$ [5]. Однако недостатком данного алгоритма можно считать то, что он позволяет выделять в документах только темы, состоящие исключительно из биграмм, что может навредить точности и полноте тем.

Другим примером тематической модели, включающей в себя выделение словосочетаний в тексте одновременно с темами, является скрытая тематическая марковская модель с применением латентного размещения Дирихле (Hidden Markov Model with Latent Dirichlet Allocation, HMM-LDA). В данной модели строится совместное описание семантико-синтаксических особенностей текста с помощью разбиения каждого предложение на функциональные слова, которые порождаются с помощью скрытой марковской модели (таким образом, описываются локальные закономерности), и на термины, генерируемые тематической моделью LDA так дается глобальное тематическое описание документа) [7]. Модель состоит из последовательности переменных-слов $w = (w_1, \dots, w_n)$, тематических переменных $z = (z_1, \dots, z_n)$ и последовательности бинарных классификаций $c = (c_1, \dots, c_n)$, указывающих, образуют ли данное слово и предыдущее словосочетание. Значение c_n выбирается, основываясь на предыдущем слове w_{i-1}, исходя из распределения $P(x_i|w_{i-1})$. Если $c_i = 1$, то слова w_{i-1} и w_i образуют словосочетание и слово w_1 анализируется в семантическом аспекте, т. е. на основании тематического распределения Φ_z:

$$P(w_i|w_{i-1}, \mathrm{c}_{i=1)}.$$

Если же $c_i \neq 1$, то слово порождается из распределения Φ_C:

$$P(w_i|t, c_i = 0).$$

Несомненным достоинством унифицированных тематических моделей является их логическое теоретическое обоснование. Однако к их минусам можно отнести большое количество параметров, нуждающихся в настройке. Например, число параметров у биграммной тематической модели равно W^2T, в то время как у базовой модели LDA — WT, где W — размер словаря (т.е. число уникальных слов и словосочетаний корпуса), T — число выделенных тем.

Одним из наиболее распространенных алгоритмов, относящихся ко второму типу, то есть позволяющих выделить биграммы на этапе предобработки текста, является алгоритм, предложенный в работе [8]. В данном алгоритме коллокации выявляются на этапе предварительной обработки текста, упорядочиваются в соответствии с ассоциативной мерой *T-Score*:

$$T - Score(xy) = \frac{TF(xy) - \frac{TF(x) \times TF(y)}{|W|}}{\sqrt{TF(xy)}}$$

где $TF(xy)$ — частотность словосочетания xy, $TF(x)$ и $TF(y)$ – частотность слов x и y соответственно, $|W|$ — число различных слов в коллекции.

Далее наиболее удачные полученные словосочетания объединяются в один токен и добавляются в корпус, заменяя униграммы. Таким образом, в процессе собственно построения тематической модели (в данном случае, LDA) и построения модели «мешка слов», они рассматриваются наряду с другими униграммами как токены.

Другим алгоритмом такого типа являются алгоритмы PLSA-SIM и PLSA-ITER, предложенные в [9]. Оба данных алгоритма являются усовершенствованными версиями одной из базовых моделей PLSA — вероятностного латентного семантического анализа, основанной на введении слоя скрытых переменных для описания тематик документов из корпуса текстов [10] — и также используют в своей работе модель «мешка слов». Идея, которая легла в основу алгоритма PLSA-SIM, следующая: в любых текстах существует большое количество семантически и лексически близких слов и коллокаций: например, *бюджетный, бюджетные расходы, бюджетные средства, бюджетные доходы* [4]. В рамках данного алгоритма при выделении тем подобные словосочетания относятся к одной теме. Если же подобные слова и словосочетания никогда не встречаются в рамках одного документа, то для них выполняется стандартный алгоритм PLSA.

Дальнейшим усовершенствованием алгоритма PLSA-SIM является итеративный алгоритм PLSA-ITER, в рамках которого рассматриваются самые частотные униграммы, представляющие темы, и из них составляются биграммы. В качестве примера авторами приводится биграмма *ценный бумага*, которая может быть составлена, если в некой теме среди первых n слов окажутся униграммы *ценный* и *бумага*. В [4] рассматриваются первые 10 униграмм, из которых далее образуются биграммы и добавляются в тематические модели. Помимо этого, в данной модели, как и в предыдущей, принимается во внимание частеречная принадлежность слов: в выделении тем

участвуют только знаменательные слова, а в формировании биграмм для русского языка учитываются коллокации, построенные по фиксированным лексико-семантическим шаблонам: *существительное + существительное в родительном падеже, прилагательное + существительное.*

Подход, предполагающий предварительное выделение биграмм в текстовых коллекциях, возможно, не имеет такого изящного теоретического обоснования, однако позволяет строить алгоритмы, являющиеся гораздо более простыми в применении. В первую очередь это достигается за счет того, что количество настраиваемых параметров в данных моделях равен их количеству в исходных моделях (как правило, LDA или PLSA). Недостатком данного подхода можно назвать повышение перплексии, что ведет к ухудшению обобщающей способности выявленной модели.

Предлагаемый нами метод можно отнести ко второму типу, поскольку он предполагает автоматическое выделение биграмм на этапе предварительной обработки текста, а затем уже их последующее добавление в тематические модели.

2.2. Используемые корпусные данные

Для проведения эксперимента были выбраны два корпуса русскоязычных текстов:

- корпус специальных текстов по радиоэлектронике, ракетостроению и технике [11]; общий объем корпуса: 526 648 словоформ;
- корпус русскоязычных специальных текстов на лингвистическую тематику из лингвистического энциклопедического словаря (ЛЭС) под редакцией В.Н.Ярцевой и энциклопедии «Кругосвет» [12]; общий объем корпуса: 1 333 546 словоформ.

Причин выбора текстов именно научного стиля несколько. Во-первых, специфическими чертами подобных текстов являются а) значительная лаконичность, однозначность, б) малая экспрессивность, метафоричность и синонимия, что несомненно позволяет упростить построение и интерпретацию тематических моделей, повысить их точность и снизить долю шума. Во-вторых, их характерной чертой является повышенное содержание терминов. Термины (в рамках данного исследования: отдельные слова и двухсловные сочетания) всегда существуют в рамках терминосистемы, обладают фиксированным терминологическим значением и, как правило, обозначают какой-то предмет, явление или, в крайнем случае, процесс. С одной стороны, это облегчает процесс тематического моделирования, поскольку тематика документов из специального корпуса текстов предсказуема. С другой стороны, регулярность и воспроизводимость терминов в рамках одного текста упрощает выделение биграмм. Также не стоит упускать из виду то, что, как правило, термины — это номинативная лексика, терминосочетания формируются из существительных и прилагательных (реже глаголов и наречий), и коллокаты этих частей речи составляют наиболее частотные биграммы.

На этапе предварительной обработки текстов из них удаляются нетекстовые символы и сокращения (в рамках данного эксперимента было принято решение об удалении всех слов, длина которых составляет менее 3 символов). Помимо

этого, из текстов исключаются слова, входящие в список стоп-слов на основе словарей служебных слов и оборотов НКРЯ, а также 98 наиболее частотных глаголов и отвлеченных существительных (например, *использовать*, *позволять, наличие, отсутствие* и так далее).

Далее была произведена лемматизация текстов и автоматическое разрешение морфологической неоднозначности с помощью морфологического анализатора русского языка MyStem 3.0 [13].

После прохождения этапа предварительной обработки текстов объемы корпусов оказались следующими:

- корпус текстов по радиоэлектронике, ракетостроению и технике: 216 613 леммы;
- корпус текстов на лингвистическую тематику: 1 246 590 лемм.

2.3. Описание алгоритма

На первом этапе работы алгоритма в исследуемом корпусе были выявлены биграммы с использованием модуля Phrases, входящего в состав библиотеки Gensim[1]. Данный модуль, используя модель «мешка слов», автоматически определяет наиболее часто встречающиеся в документах многословные словосочетания.

Параметр, отвечающий за принятие решения о формировании биграммы (*threshold*) основан на совместной встречаемости слов. Слова a и b считаются биграммой, если:

$$\frac{(cnt(a,b) \; - \; min\_count) * N}{cnt(a) * cnt(b)} > threshold,$$

где N – общий размер словаря.

В наших экспериментах в качестве документов рассматривались предложения из корпуса, поскольку наибольший интерес вызывает совместная встречаемость слов именно в пределах синтагм. Обучение модуля проводилось непосредственно на корпусе текстов, с которым велась далее работа. Технически алгоритм был реализован на языке программирования Python.

В процессе работы алгоритма в корпусе текстов было образовано 14 542 биграмм для корпуса текстов по радиоэлектронике, ракетостроению и технике и 187 008 биграмм для корпуса текстов на лингвистическую тематику. Соотношение же с общим объемом корпусов составляет: 13.4% для корпуса текстов по радиоэлектронике, ракетостроению и технике и 30.0% для корпуса текстов на лингвистическую тематику. Возможных причин бóльшему количеству выделенных биграмм во втором корпусе несколько: во-первых, роль играет объем корпуса, поскольку от него напрямую зависит параметр *threshold*: при увеличении объема число биграмм также увеличивается. Во-вторых, стоит отметить разницу в лексической специфике собранных корпусов. Корпус текстов на лингвистическую тематику более однородный; лексическое наполнение входящих в него текстов более однообразно, наблюдается бóльшее

[1] https://radimrehurek.com/gensim/, дата последнего обращения: 27.04.2017

количество высокочастотных терминов. Напротив, корпус по радиоэлектронике, ракетостроению и технике представляет собой тематически более разнородную выборку текстов и изобилует низкочастотными терминами. Общая частота употребления терминов в корпусе снижается за счет лексического разнообразия, и поэтому меньшее количество униграмм проходит порог формирования биграммы. Помимо этого, в соответствии с установленными нами параметрами униграммы, встречающиеся в текстах менее двух раз, вовсе не рассматриваются; поэтому, вероятно, некоторое количество потенциальных биграмм, которые были образованы из низкочастотных терминов, не попали в конечный список.

Примеры фрагментов корпусов после первичной обработки, а также после работы алгоритма по выделению биграмм приведены в табл.1 и табл. 2.

Таблица 1. Примеры корпуса текстов по радиоэлектронике, ракетостроению
и технике на разных этапах обработки

Отрывок корпуса после предварительной обработки, перед запуском алгоритма по выделению биграмм	считаться линейно условие гауссовский величина связанный сила выясняться расчет точность потенциальный методика удовлетворительный точка полностью предел вправе шум производная дисперсия условие превышение укладываться смещать разброс наблюдатель настолько сигнал линейный участок омп большой побочный шумовой выброс необходимо обусловливать пик сигнал основный превосходить
Отрывок корпуса после работы алгоритма по выделению биграмм	линейно условие ***гауссовский_величина*** связанный сила выясняться ***расчет_точность*** потенциальный методика удовлетворительный точка полностью предел вправе шум ***производная_дисперсия*** ***условие_превышение*** укладываться смещать разброс наблюдатель ***настолько_сигнал*** линейный участок омп большой ***побочный_выброс*** шумовой необходимо обусловливать пик сигнал основный превосходить

Анализируя примеры из корпуса, можно заметить, что в результате работы алгоритма было образованно несколько корректных биграмм: например, *гауссовский_величина*, *побочный_выброс*, *расчет_точность* и *норма_произносительный*, *норма_орфоэпический*, *национальный_язык*, *норма_складываться*. Остальные же обобщенные униграммы, выделенные в данном отрывке, не соответствуют шаблонам, по которым должна производиться сборка биграмм.

На втором этапе работы алгоритма была построена тематическая модель экспериментального корпуса. Как уже было упомянуто ранее, нами было принято решение использовать для этого вероятностную тематическую модель латентного размещения Дирихле [3], включенную в пакет для анализа данных Scikit-Learn[2]. Эмпирическим путем были установлены параметры, позволяющие

[2] http://scikit-learn.org/stable, дата последнего обращения 27.04.2017

выделить темы наиболее точно: количество итераций алгоритма — 200, количество тем — 20, количество слов, представляющих каждую тему — 10 первых слов. Последний параметр неслучайно был выбран именно таким: исследования показали, что именно 10 первых слов содержат в себе 30% информации о теме, распределенной в других словах, что является достаточным для достаточно полного представления темы [14]. Также стоит отметить, что при построении тематической модели не учитывались слова, встретившиеся менее, чем в двух документах, а также высокочастотные слова — в данном случае, содержащиеся более чем в 80% документов.

Таблица 2. Примеры корпуса текстов на лингвистическую тематику на разных этапах обработки

Отрывок корпуса после предварительной обработки, перед запуском алгоритма по выделению биграмм	кодифицированный часто отставать реально норма складываться речь язык новый устный формирование национальный публичный одновременно действие развиваться форма складываться сфера орфоэпия расширяться язык разный национальный становление норма орфоэпический проходить разному процесс язык национальный норма орфоэпический проходить этап русский особенность
Отрывок корпуса после работы алгоритма по выделению биграмм	кодифицированный ***часто_отставать*** реально ***норма_складываться*** речь язык новый ***устный_формирование*** ***национальный_публичный*** одновременно действие развиваться форма складываться ***сфера_орфоэпия*** расширяться язык разный национальный становление ***норма_орфоэпический*** проходить разному процесс язык национальный ***норма_орфоэпический*** проходить этап русский особенность

На третьем этапе полученные темы были заново обработаны с помощью модуля Phrases, что позволило выделить еще некоторое количество биграмм. Полученные конечные результаты представлены в табл. 3 и табл. 4.

Таблица 3. Результаты тематического моделирования на корпусе текстов по радиоэлектронике, ракетостроению и технике с учетом выделенных биграмм

№	Список первых 10 слов из темы
1.	вектор множество элемент **пространство_линейный** оператор пример расстояние образовывать состоять
2.	оценка вероятность правило наблюдение гипотеза задача решение средний правдоподобие оптимальный
3.	система скорость рлс дальность цель измерение объект антенна точность координата
4.	исз потребитель точка доплеровский момент положение измерение шкала пересечение поверхность
5.	код суммарный канал измерение сдвиг система разностный частота устройство информация

№	Список первых 10 слов из темы
6.	последовательность код символ состояние сигнальный расстояние путь **скорость_код** сверточный пример
7.	активный **обзор_рлс** информация система эффективность **ширина_спектр** мощность дальность радиотехнический
8.	канал передача связь пользователь система характеристика частота цифровой скорость полоса
9.	фаза схема **огибать_омп** фильтр начальный амплитуда шум детектор частота
10.	вероятность условие величина случайный **случайный_величина** определение **фурье_преобразование** событие результат
11.	фильтр коэффициент алгоритм последовательность эквалайзер оценка линейный уравнение модель характеристика
12.	**суммарный_канал**, детектор, антенна, фаза, амплитуда, устранение, измеритель, характеристика
13.	мощность импульс потеря выходной энергия шум длительность входной отношение достигаться
14.	дальность, рлс, спектр, **радиолокационный_цель**, точность, антенна, частота, заданный, объект
15.	система измерение помеха измеритель вектор обработка радиотехнический фильтрация комплексный способ
16.	код источник бит вход кодирование кодовый **кодовый_слово** символ канал уровень
17.	различение цель дискриминатор **проверка_гипотеза** дальность условие оптимальный импульс характеристика рлс
18.	частота спектр импульс время частотный модуляция огибать амплитуда фаза полоса
19.	распределение процесс случайный дисперсия нормальный момент средний вероятность характеристика выражение
20.	генератор частота опорный потребитель исз шкала измерять скорость уравнение изменение

Таблица 4. Результаты тематического моделирования на корпусе текстов по лингвистике с учетом выделенных биграмм

№	Список первых 10 слов из тем
1.	система форма основа глагол гласный ряд согласный тип диалект группа
2.	местоимение лицо число человек числительный личный класс группа указательный **число_местоимение**
3.	имя **форма_падеж** число род система предлог русский морфологический прилагательный
4.	значение форма глагол тип русский выражение отношение грамматический функция вид
5.	китайский **латинский_письменность** слог романский диалект тон время часть французский
6.	предложение логический синтаксис вещь универсальный семантика анализ предмет событие психологический
7.	литературный диалект русский современный арабский форма языковой разный социальный национальный

№	Список первых 10 слов из тем
8.	английский немецкий французский современный новый германский греческий форма период изменение
9.	знак письмо буква система **алфавит_письменность** согласный гласный звук форма
10.	строй мышление **влияние_оказывать** след эпоха звук соответствие русский характер создавать
11.	морфема значение форма грамматический тип часть термин морфологический словоформа правило
12.	словарь значение лексика русский лингвистический семантический **словарный_толкование** лексический **словарный_статья**
13.	человек говорить языковой речевой случай мир система выражение речь отношение
14.	русский значение тип случай правило форма фонетический **ударение_позиция** разный
15.	семья группа время диалект говорить история языковой исследование современный число
16.	единица признак фонема морфема звук разный речь общий свойство уровень
17.	предложение глагол конструкция синтаксический подлежащее порядок дополнение сказуемое субъект тип
18.	текст анализ перевод дискурс год средство ряд термин автор структура
19.	языковой лингвистический исследование система теория лингвистика изучение языкознание развитие работа
20.	объект состояние предмет термин языкознание служить средство стиль сравнение изучение

Проанализировав результаты, полученные при работе с корпусом по радиоэлектронике, ракетостроению и технике, можно заметить, что большинство выделенных биграмм действительно образуют логичные словосочетания, такие как: *линейное пространство*, *ширина спектра*, *случайная величина*, *преобразование Фурье*, *суммарный канал*, *радиолокационная цель*, *кодовое слово*, *проверка гипотезы*. Появление остальных биграмм также вполне объяснимо: например, биграмма *потребитель_исз* была выделена, вероятно, вследствие того, что слова *исз* (искусственный спутник Земли) и *потребитель* (в значении *исследователь*, *наблюдатель*) часто встречаются в таких схожих контекстах, как *расстояние между потребителем и ИСЗ*, *скорость ИСЗ относительно потребителя* и так далее. Также слова *обзор* и *рлс* (радиолокационные станции), хоть и не встречаются в тексте стоящими рядом, во многих контекстах встречаются в непосредственной близости: например, *РЛС дальнего обзора*, *РЛС ближнего обзора*, *обзорные РЛС* и т.п.

Основное отличие результатов обработки корпуса текстов по лингвистике и данных, полученных из корпуса по радиоэлектронике, ракетостроению и технике, заключается в том, что общее количество выделенных в темах биграмм меньше, чем в предыдущем случае, однако значительная их часть является корректными сочетаниями: *латинская письменность*, *алфавитная письменность*, *падежная форма*, *словарное толкование*. Биграмму *ударение_позиция* также нельзя назвать случайной: составляющие его униграммы часто встречаются в одном предложении, например, «… фиксированное ударение ориентируется на крайние позиции в слове — либо на

его начало, либо на конец…» или даже в непосредственной близости: «…Особенность фонетики собственно алюторского диалекта — противопоставление по долготе в системе гласных, …, динамическое позиционное ударение…». Униграммы *число* и *местоимение*, по всей видимости, были объединены также на основании частой совместной встречаемости в одном предложении (несложно представить такие контексты, описывающие формы местоимений); однако нельзя утверждать, что они образуют корректную биграмму.

3. Заключение

В настоящей статье был предложен алгоритм для автоматического выделения биграмм на этапе предварительной обработки текста и их последующего добавления в тематические модели. За основу была взята вероятностная модель латентного размещения Дирихле. Разработанный алгоритм был проверен на двух корпусах специальных текстов. Достоинство предложенного алгоритма заключается в том, что он позволяет выявлять биграммы в корпусах текстов, существенно не усложняя модель и не требуя внедрения дополнительных параметров. Стоит также отметить, что, благодаря очевидному удобству применения данного метода, он может считаться универсальным и применимым для выделения словосочетаний в текстах разных стилей и типов (как научных, так и художественных, официально-деловых и публицистических).

В перспективе планируется усовершенствовать выделение биграмм, используя частеречную разметку корпуса. В большинстве своем правильно выделенные темы формируются именно из существительных и именных групп [15], поэтому в дальнейшем планируется формировать биграммы в корпусе текстов преимущественно в соответствии с моделями *существительное + существительное в родительном падеже*, *существительное + прилагательное*. Также планируется приведение биграмм из лемматизированной формы к согласованным словосочетаниям путем их повторного поиска в корпусе текстов и замены на исходные формы.

Исследование поддержано грантом РФФИ № 16–06–00529 «Разработка лингвистического комплекса для автоматического семантического анализа русскоязычных корпусов текстов с применением статистических методов» (2015–2018 гг.).

Литература

[1] Daud A., Li J., Zhou L., Muhammad F. Knowledge discovery through directed probabilistic topic models: a survey // Frontiers of Computer Science in China. 2010. Vol. 4. № 2. Pp. 280–301.

[2] Митрофанова О.А. Вероятностное моделирование тематики русскоязычных корпусов текстов с использованием компьютерного инструмента GenSim // Труды международной конференции «Корпусная лингвистика-2015». СПб., 2015. С. 332–343.

[3] Blei D.M, Ng A., Jordan M. Latent Dirichlet Allocation // Journal of Machine Learning Research. 2003. Vol. 3. Pp. 993–1022.
[4] Нокель М.А., Лукашевич Н.В. Тематические модели: добавление биграмм и учет сходства между униграммами и биграммами // Вычислительные методы и программирование. 2015. Т.6. С. 215–234.
[5] Wallach H. Topic Modeling: Beyond Bag-Of-Words // Proceedings of the 23rd International Conference on Machine Learning. 2006. Pp. 977–984.
[6] Wang X., McCallum A., Wei X. Topical N-Grams: Phrase and Topic Discovery, with an Application to Information Retrieval // Seventh IEEE International Conference on Data Mining (ICDM 2007). NY: IEEE, 2007. Pp. 697–702.
[7] Griffiths T.L. , Steyvers M., Blei D.M, Tenenbaum J.B. Integrating topics and syntax // Advances in Neural Information Processing Systems (NIPS) 17. Cambridge, MA, MIT Press. 2005. 18 p.
[8] Lau J.H., Baldwin T., Newman D. On Collocations and Topic Models // ACM 131 Transactions on Speech and Language Processing. ACM Press. Vol. 10, №3. 2013. Pp. 1–14.
[9] Нокель М.А. Методы улучшения вероятностных тематических моделей текстовых коллекций на основе лексико-терминологической информации: Дис. … канд. физ-мат. наук. М., 2016. 159 с.
[10] Hofmann T. Probabilistic latent semantic analysis // Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI), Stockholm, Sweden, 1999. Pp. 289–296.
[11] Дубовик А.Р. Автоматическое определение стилистической принадлежности текстов по их статистическим параметрам // Сборник научных статей XX Объединенной конференции «Интернет и современное общество» IMS-2017. – Санкт-Петербург, 21 – 23 июня 2017 года. СПб., 2017. [наст. изд.]
[12] Mirzagitova A., Mitrofanova O. Automatic assignment of labels in Topic Modelling for Russian Corpora // Proceedings of 7th Tutorial and Research Workshop on Experimental Linguistics, ExLing 2016 / ed. A. Botinis. – Saint Petersburg: International Speech Communication Association, 2016. Pp. 115–118.
[13] Segalovich I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine // Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications. MLMTA'03, June 23 –26, 2003. Las Vegas, Nevada, USA, 2003.
[14] Lau J.H., Newman D., Karimi S., Baldwin T. Best Topic Word Selection for Topic Labelling // COLING'10 Proceedings of the 23rd International Conference on Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2010. Pp. 605–613.
[15] Wang X., McCallum A., Wei X. Topical N-Grams: Phrase and Topic Discovery, with an Application to Information Retrieval // Seventh IEEE International Conference on Data Mining (ICDM 2007). NY: IEEE, 2007. Pp. 697–702.

Topic Modelling of Russian Texts based on Lemmata and Lexical Constructions

A. Sedova, O. Mitrofanova

Saint-Petersburg State University

The paper is devoted to the improvement of topic modelling algorithms aimed at extraction of latent relations between words, documents and topics in processed corpora. In the majority of cases topics generated by topic models contain only unigrams, so that the interpretation of extracted topics turns out to be a complicated task. This paper presents a new algorithm based on the classic LDA model which provides automatic extraction of bigrams in the given text collection and further incorporation of bigrams into the topic model. In the given paper we describe out algorithm in action and discuss results achieved in course of processing the Russian corpora on radioengineering and on linguistics.

Keywords: topic modelling, LDA, bigrams.