УДК 004.912

С.Н. Карпович

РУССКОЯЗЫЧНЫЙ КОРПУС ТЕКСТОВ SCTM-RU ДЛЯ ПОСТРОЕНИЯ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ

Карпович С.Н. **Русскоязычный корпус текстов SCTM-ru для построения тематических моделей.**

Аннотация. В статье рассматривается задача создания русскоязычного специального корпуса текстов для тестирования алгоритмов вероятностного тематического моделирования. В качестве наполнения корпуса предлагается использовать статьи международного новостного сайта «Русские Викиновости», распространяемого по свободной лицензии CC BY 2.5. Описан этап предварительной обработки и разметки корпуса текстов. Предложена разметка корпуса текстов, содержащая только необходимую в алгоритмах тематического моделирования информацию.

Ключевые слова: корпус текстов, обработка текста на естественном языке, тематическое моделирование, русский язык.

Karpovich S.N. **The Russian language text corpus for testing algorithms of topic model.**

Abstract. This paper describes the process of creating Russian language text corpus which is specialized for testing algorithms of probabilistic topic model. The articles of Wikinews licensed by Creative Commons Attribution 2.5 Generic (CC BY 2.5) were used as a source of texts for corpus. The stage of text's preprocessing and markup are described in the conclusion. We proposed an original markup of text corpus for testing algorithms of topic modeling.

Keywords: text corpora, topic model, natural language processing, Russian language.

1. Введение. В современном обществе главным продуктом и основным товаром становится информация. Активно развиваются: наука, экономика, политика, производственная сфера, во многих отраслях происходит создание и накопление цифровых данных. Для успешного извлечения и обработки информации из данных необходимо располагать подходящими инструментами, системами и алгоритмами. Растет потребность в системах обработки текстов на естественном языке. Обработка естественного языка (Natural Language Processing) уже применяется в привычных для пользователя программах и сервисах. Например, программы для чтения новостных лент умеют группировать новости по темам, поисковые системы находят документы с ценной для пользователя информацией, службы почтовых сообщений автоматически фильтруют спам. Используются различные алгоритмы кластеризации и классификации текстовых данных, наиболее популярные k-средних, SVM, нейронные сети. Перспективным направлением автоматической обработки текстов является разработка алгоритмов вероятностного тематического моделирования.

Тематическое моделирование – это способ построения тематической модели коллекции текстовых документов. Тематическая

модель задает отношение между темами и документами в корпусе текстов. Первое описание тематического моделирования появилось в работе Рагавана, Пападимитриу, Томаки и Вемполы 1998 году [1]. Томас Хофманн в 1999 году предложил вероятностное скрытое семантическое индексирование (PLSI) [2]. Одна из самых распространенных тематических моделей — это латентное размещение.

Дирихле (LDA), эта модель является обобщением вероятностного семантического индексирования и разработана Дэвидом Блейем, Эндрю Ыном и Майклом Джорданом в 2002 году [3]. Другие тематические модели, как правило, являются расширением LDA. В качестве примера на рисунке 1 представлен процесс построения тематической модели документа.

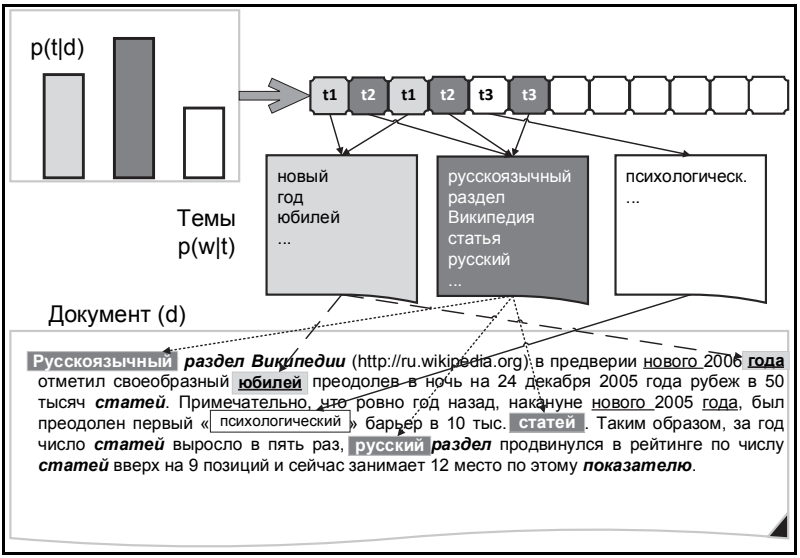

Рис. 1. Построение тематической модели документа: p(w|t) – матрица искомых условных распределений слов по темам, p(t|d) – матрица искомых условных распределений тем по документам: d — документ: w — слово: d, w — наблюдаемые переменные: t — тема (скрытая переменная)

Алгоритмы тематического моделирования ориентированы на работу с текстом на естественном языке. Первоначальные решения основывались на предположении, что текст — это «мешок слов», т.е. порядок слов в тексте не имеет значения. В последующих моделях успешно реализованы алгоритмы, учитывающие зависимости между словами с помощью скрытых Марковских моделей. В обзоре [4]

рассмотрены пять основных классов вероятностных тематических моделей: базовые, учитывающие отношения между документами, учитывающие отношения между словами, временные, обучаемые с учителем.

Наличие подготовленных текстовых корпусов позволит разрабатывать системы для автоматической обработки текстов на естественном языке, в том числе алгоритмы тематического моделирования. Создавая тематическую модель, необходимо учитывать языковые особенности текстов. Для развития методов тематического моделирования, работающих с русским языком, необходим русскоязычный корпус текстов, распространяемый по свободной лицензии.

Корпусная лингвистика – сложная лингвистическая дисциплина, которая сформировалась в последние десятилетия на базе электронной вычислительной техники. Она изучает построение лингвистических корпусов, способы обработки данных в них и собственно технологию их создания и использования. «Корпус – это информационно справочная система, основанная на собрании текстов на некотором языке в электронной форме», - такое определение текстового корпуса дано на сайте Национального Корпуса Русского Языка [5]. В научной работе [6] отмечено: «Корпусы, как правило, предназначены для неоднократного применения многими пользователями, поэтому их разметка и их лингвистическое обеспечение должны быть определенным образом унифицированы». Целесообразность создания и смысл использования корпуса определяется следующими предпосылками:

1. достаточно большой (репрезентативный) объем корпуса;
2. данные разного типа находятся в корпусе в своей естественной контекстной форме;
3. однажды созданный и подготовленный массив данных может использоваться многократно.

Корпусами первого порядка называют собрание текстов, объединенных общим признаком, например, источник, автор, место публикации. Специальный корпус текстов – это сбалансированный корпус, репрезентативный, как правило, небольшой по размеру подчиненный определенной исследовательской задаче и предназначенный для использования преимущественно в целях, соответствующих замыслу составителя. Текстовый корпус (text corpora, corpus), большая коллекция документов (large collection of documents), набор данных (dataset), как отмечено в работе [4] являются синонимичными понятиями.

Цель данной работы заключается в том, чтобы создать специальный русскоязычный корпус текстов СКТМ-ру (SCTM-ru), пригодный для исследования алгоритмов вероятностного тематического моделирования. Сформулируем требования к создаваемому корпусу. Корпус должен распространяться по свободной лицензии, количество документов должно быть достаточным для исследования, он должен содержать:

- оригинальный текст документов на естественном языке;
- даты описанных событий;
- информацию об авторстве;
- темы или тематические категории.

Рассмотрим возможность использования существующих текстовых корпусов для целей тестирования алгоритмов тематического моделирования.

2. Обзор текстовых корпусов и наборов данных. Национальный Корпус Русского Языка (НКРЯ) [5] – содержит более 335 тыс. документов на русском языке, разделенных на подкорпуса. Включает в себя 180 тыс. текстов Газетного корпуса. Для использования офлайновой версии основного корпуса (1 млн. словоупотреблений) необходимо подписать лицензионное соглашение. Статистика корпуса представлена на рисунке 2.

Подкорпус	Число текстов	Число предложений	Число словоупотреблений	% словоупотреблений
Основной корпус	76 882	17 574 752	209 198 275	57.3%
- в том числе со снятой омонимией	2 147	516 852	5 944 188	1.6%
Газетный корпус	181 175	8 553 495	113 292 003	31.0%
Диалектный корпус	197	20 273	194 283	0.1%
Обучающий корпус	229	65 666	664 751	0.2%
Параллельный корпус	370	1 609 609	24 022 437	6.6%
Поэтический корпус	41 448	638 861	6 738 474	1.8%
Устный корпус	3 034	1 604 626	10 122 579	2.8%
Мультимедийный корпус	31 741	148 619	648 576	0.2%
Всего:	**335 076**	**30 215 901**	**364 881 378**	**100%**

Рис. 2. Статистика НКРЯ, распределение текстов по подкорпусам [5]

Открытый корпус (OpenCorpora) [9, 10] – содержит порядка 3 тыс. документов на русском языке, 93 тыс. размеченных предложений,

10 источников данных, часть документов имеет информацию об авторе и дате описываемых событий. Частям текста приписана лингвистическая информация, такая как морфологическая, семантическая и синтаксическая. Для задач построения временных и автор-тематических моделей, корпус не подходит, так как не все документы корпуса содержат информацию об авторе и о дате событий.

Associated Press [11] – корпус содержит 2 тыс. документов на английском языке. Документы корпуса не имеют отметки о дате описанного события, авторе публикации, категории документа. Корпус применим для исследования ограниченного количества алгоритмов тематического моделирования.

The New York Times Annotated Corpus [12] – большой англоязычный текстовый корпус газетных заметок и новостей, распространяемый по закрытой лицензии.

20 Newsgroups [13] – коллекция новостей на английском языке, подготовленная для исследования алгоритмов автоматической обработки текстов. 20 новостных групп содержит порядка 20 тыс. документов. Важные для построения тематических моделей данные об авторе и дате публикации не размечены. Требуется предварительная обработка текста новостей для использования в тематическом моделировании.

Reuters Corpora [14] – большой англоязычный новостной корпус. В трех наборах данных более 3 млн. новостей. Распространяется по ограниченной лицензии только для научных исследований, предоставляется после подписания лицензионного соглашения. Существует более ранняя популярная для тестирования алгоритмов автоматической обработки текстов версия корпуса под названием Reuters-21578 [15], распространяемая по ограниченной лицензии, доступная для офлайн анализа.

Компьютерный корпус текстов русских газет конца XX-ого века [16, 17] – этот корпус был создан в 1999 году развивается и исследуется в настоящее время по грантам РФФИ. Корпус предназначен для анализа лингвистических особенностей (лексика, морфемика, морфология, словообразование, синтаксис, фразеология, стилистика) современного газетного языка. В корпусе 23 тыс. текстов по полным номерам 13-ти разных российских газет на русском языке. Размер корпуса 11 млн. словоупотреблений.

Корпус русского литературного языка [18, 19] – представлен в виде массива морфологически аннотированных текстов на русском литературном языке. Размер корпуса более 1 млн словоупотреблений со сбалансированным жанровым составом.

Хельсинкский аннотированный корпус русских текстов ХАНКО [20] – корпус содержит морфологическую, синтаксическую и функциональную информацию о текстах общим объемом 100 тыс. текстов, извлеченных из журнала «Итоги». Права на полные тексты статей журнала принадлежат правообладателям.

В таблице 1 представлено сравнение важных для исследования алгоритмов тематического моделирования характеристик корпусов: язык корпуса, лицензия распространения, доступность для скачивания и исследования на компьютерах без доступа в Интернет, информация об авторе, информация о дате описанных событий, тема текста.

Таблица 1. Сравнительная таблица характеристик текстовых корпусов

Корпус	Язык	Открытая лицензия	Доступен для скачивания	Инф. об авторе	Инф. о дате	Темы
НКРЯ	рус.	-	-	+	-	-
Открытый корпус	рус.	+	+	+	+	-
Associated Press	англ.	+	+	-	-	-
The New York Times Annotated Corpus	англ.	-	-	+	+	+
20 Newsgroups	англ.	+	+	-	-	-
Reuters Corpora	англ.	-	-	+	+	+
Компьютерный корпус текстов русских газет конца XX-ого века	рус.	-	-	-	-	-
Корпус русского литературного языка	рус.	-	-	-	-	-
ХАНКО	рус.	-	-	-	-	-
SCTM-ru	рус.	+	+	+	+	+

Рассмотренные текстовые корпуса в полной мере не соответствуют обозначенным в данной работе требованиям. Создаваемый специальный корпус для тематического моделирования СКТМ-ру (SCTM-ru) распространяется по свободной лицензии, язык корпуса русский, содержит информацию об авторстве, дате событий, тематической принадлежности документов, доступен для скачивания и проведения исследований на компьютерах без доступа в Интернет.

3. Технология создания корпуса SCTM-ru. Технологический процесс создания корпуса состоит из следующих шагов.

1. определение источника;
2. предварительная обработка текстов документов;
3. разметка параметров каждого документа в корпусе;
4. обеспечение доступа к корпусу.

В соответствии с обозначенным требованием к доступности данных корпуса, текста используемые в качестве наполнения должны

распространяться по свободной лицензии, должны быть доступны для скачивания и свободного использования.

В результате предварительной обработки текстов и разметки параметров каждого документа, в корпусе должна быть сохранена и специальным образом размечена информация необходимая для построения тематических моделей. Невостребованная в тематическом моделировании информация должна быть исключена из корпуса, за ненадобностью.

Различные задачи тематического моделирования могут требовать определенный порядок поступления данных в систему тематического моделирования, от последовательного для временных, до единовременного для обычных тематических моделей. Поэтому для обеспечения доступа к корпусу достаточно предоставить возможность для его скачивания, и последующего использования в соответствии с конкретными задачами, стоящими перед исследователем.

4. Источник данных для корпуса SCTM-ru. В качестве источника данных предлагаем использовать международный новостной сайт «Русские Викиновости» (Викиновости), тексты статей которого распространяются по свободной лицензии Creative Commons Attribution 2.5 Generic, доступны для скачивания и анализа на любых компьютерах, в том числе на компьютерах без доступа в Интернет. В работах [7, 8] отмечены преимущества вики-ресурсов, таких как Викисловарь и Википедия, для использования в качестве источника данных в исследовательских целях. Вики-ресурсы – это сайты второго поколения Интернет, характеризуемые тем, что к их наполнению привлечено огромное количество рядовых пользователей, с помощью которых происходит пополнение и актуализация информации. Большой объем, постоянное пополнение, нейтральность во взглядах, доступность относятся к преимуществам всех вики-ресурсов, в том числе к Викиновостям.

Викиновости – это братский проект большой Википедии, предназначен для написания новостных статей. Пример статьи Викиновостей представлен на рисунке 3. Отличительной особенностью сайта Викиновостей от любого другого новостного сайта является то, что каждый человек может принять участие в создании новости. Правила Викиновостей требуют писать новости с нейтральной точки зрения, в непредвзятом виде, выбирать существенные и актуальные темы, использовать достоверные источники.

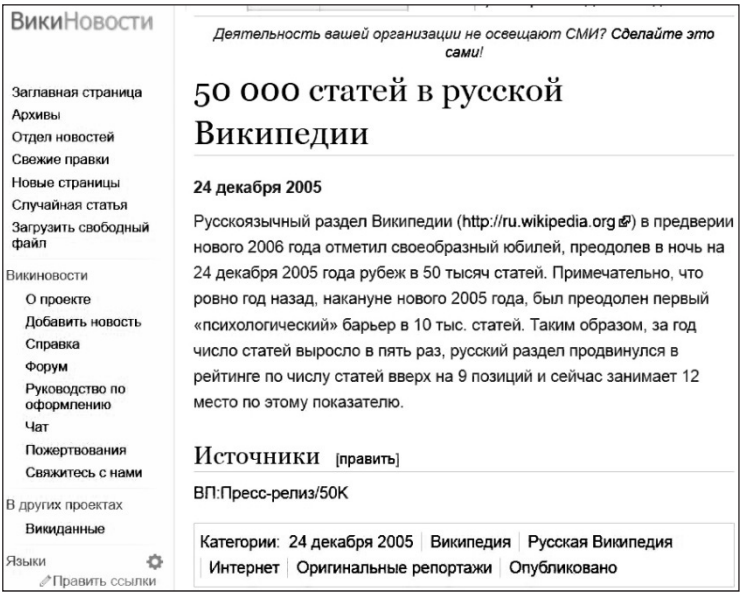

Рис. 3. Статья "50 000 статей в русской Википедии" на сайте русских Викиновостей

XML-файл экспорта базы данных Викиновостей состоит из следующих XML-элементов:

+ `<page>` – группа элементов новостной статьи;
+ `<title>` – название статьи;
– `<ns>` – идентификатор или имя пространства имен (namespace), элемент предназначен для отделения основных статей от служебных, ноль соответствует основному пространству имен;
+ `<id>` – уникальный идентификатор статьи;
+ `<revision>` – ревизия – это группа элементов актуальной версии статьи;
  – `<id>` – первичный ключ ревизии, используется для контроля изменений статьи;
  – `<parented>` – идентификатор родительской статьи;
  – `<timestamp>` – дата и время создания ревизии статьи;
  + `<contributor>` – группа элементов авторства статьи;
  – `<username>` – имя автора статьи;
  + `<id>` – уникальный идентификатор автора статьи;
  + `<text>` – текст статьи с элементами вики-разметки;

– `<sha1>` – хеш код статьи полученный алгоритмом криптографического хеширования SHA-1, используется для контроля версий;
– `<model>` – модель контента статьи, в данном случае wikitext;
– `<format>` – формат данных статьи, в данном случае text/x-wiki.

Для задач тематического моделирования необходима информация, которая содержится в элементах, отмеченных знаком плюс (+). Элементы, которые содержат информацию, неиспользуемую в алгоритмах тематического моделирования, отмечены знаком минус (-). Пример части XML-дерева экспортного файла базы данных Викиновостей представлен на рисунке 4.

```
<?xml version="1.0" encoding="utf-8"?>
<page>
  <title>50 000 статей в русской Википедии</title>
  <ns>0</ns>
  <id>1838</id>
  <revision>
    <id>75780</id>
    <parentid>39949</parentid>
    <timestamp>2011-10-01T18:45:01Z</timestamp>
    <contributor>
      <username>Schekinov Alexey Victorovich</username>
      <id>2156</id>
    </contributor>
    <text xml:space="preserve">{{:Дата |24 декабря 2005}}
{{ВикипедияН
 |Язык = Русская
 }}
Русскоязычный раздел [[Википедия|Википедии]] (http://ru.wikipedia.org)
в предверии нового 2006 года отметил своеобразный юбилей, преодолев в
ночь на 24 декабря 2005 года рубеж в 50 тысяч статей. Примечательно,
что ровно год назад, накануне нового 2005 года, был преодолен первый
«психологический» барьер в 10 тыс. статей. Таким образом, за год число
статей выросло в пять раз, русский раздел продвинулся в рейтинге по
числу статей вверх на 9 позиций и сейчас занимает 12 место по этому
показателю.
{{оригинальный репортаж 2}}
== Источники ==
{{w|ВП:Пресс-релиз/50K}}
{{публиковать}}
[[Категория:Русская Википедия]]</text>
    <sha1>ezckutzzznn6tioszytixr2atkv0v4p</sha1>
    <model>wikitext</model>
    <format>text/x-wiki</format>
  </revision>
</page>
```

Рис. 4. Пример XML статьи "50 000 статей в русской Википедии" на сайте русских Викиновостей

5. Предварительная обработка данных Викиновостей. В экспортном файле Викиновостей статьи отсортированы по дате создания ревизии `<timestamp>`, эта дата не связана с датой описанных событий. Авторам рекомендуется указывать с помощью вики-разметки дату событий в тексте статьи. Пример вики-разметки даты {{:Дата | 24 декабря 2005}} представлен на рисунке 4 внутри элемента `<text>`. Часть статей в экспортном файле Викиновостей не содержит дату событий в вики-разметке, но при этом она указана в тексте или в категории. Чтобы сохранить максимум востребованной в алгоритмах тематического моделирования информации, дата событий была по возможности восстановлена из текста и категорий. В 455 статьях не удалось восстановить дату событий, эти статьи являются подборками новостей, произошедших в один день, в разные годы и не представляют ценности для построения тематических моделей, они были исключены из корпуса. Документы корпуса SCTM-ru отсортированы по дате событий, от старых к новым.

В экспортном файле базы данных Викиновостей содержится информация об авторе последней ревизии статьи. Используем эту информацию как идентификатор авторства для построения автор-тематических моделей. Так как 58 Викиновостей не содержат информацию об авторе, а статьи ценны, то было принято техническое решение присвоить этим статьям уникальный идентификатор автора – 2, и включить их в корпус SCTM-ru.

Текст статьи Викиновостей содержит оформленные специальным образом ссылки. Ссылки делятся на три группы: внутренние – инструмент связывания страниц внутри языкового раздела Википедии, межязыковые ссылки (интервики) – средство для организации связей между различными вики-системами в сети Интернет и ссылки на страницы братских вики-проектов (например, на Википедию). Текст статьи, заключенный в двойные квадратные скобки является внутренней ссылкой, пример [[Википедия|Википедии]] представлен на рисунке 4. Если падеж ссылающегося слова или словосочетания не совпадает с именительным падежом, то в двойных квадратных скобках стоит черта, слева от которой указан именительный падеж текста ссылки, а справа текст, соответствующий грамматике предложения. Алгоритмы тематического моделирования учитывают количество вхождений каждой леммы слова в текст, во внутренних ссылках каждое слово имеет два вхождения в разных словоформах и будет дважды учтено в тематической модели, тем самым исказив частотные характеристики модели. В документах

корпуса SCTM-ru оставлена только та часть ссылки, которая соответствует грамматике предложения.

Новости должны сопровождаться ссылками на документальный источник. Они обычно делятся на четыре вида: другие статьи Викиновостей, внешние ссылки на онлайн-источники, цитаты печатных изданий и веб-сайты со справочной или связанной информацией. Для раздела статьи «Источники» используют вики-разметку == *Источники* == (см. пример на рисунке 4). Для целей тематического моделирования ссылки на источники не представляют большой ценности, поэтому было принято решение об их исключении из корпуса SCTM-ru.

Важным элементом разметки Викиновостей и важными данными для построения тематических моделей является информация о категориях, к которым статья имеет отношение. Категории статьи определяет ее автор.

Для предварительной обработки текстов была разработана программа на языке C#, среда разработки Visual Studio Express 2013. Для поиска по экспортному файлу Викиновостей использовались регулярные выражения. Пример задействованных регулярных выражений представлен в таблице 2. Программа многомодульная, каждый модуль выполняет одну определенную операцию. Программа получает на вход исходный XML-файл, специально подготовленные регулярные выражения последовательно обходят файл в поисках совпадения по шаблону, на выходе создается XML-файл с внесенными за одну итерацию изменениями. Для сохранения целостности первоначальных данных, каждый проход по исходному XML-файлу вносит лишь часть изменений, которые внимательно проверяет администратор системы, после чего программу запускают с другим модулем обработки.

Таблица 2. Примеры регулярных выражений для предварительной обработки текста Викиновостей

Регулярное выражение	Назначение поиска
^(=)?=(\s+)?Источник(и)?(\s+)?(=)?=\n(^([^\n]+)\n)+^\n	блок источники
\{\{Категории|([^|]+)|([^|]+)\}\}	блок категории
\[\[([^|]+)|([^\]]+)\]\]	Ссылки

Для подсчета статистики корпуса SCTM-ru была разработана многомодульная программа. Модуль подсчета документов осуществляет разбор XML-дерева корпуса, извлекает уникальные

идентификаторы каждого документа и считает их общее количество. Модуль подсчета авторов извлекает список уникальных идентификаторов авторов статей Викиновостей и подсчитывает их количество. Модуль подсчета категорий извлекает уникальные категории из XML-дерева корпуса и считает их количество. Модуль обработки дат описанных в статьях событий осуществляет разбор XML-дерева корпуса, извлекает информацию о дате события каждого документа, подсчитывает уникальные значения, находит самую раннюю и самую позднюю дату документа.

Для подсчета словарного состава корпуса SCTM-ru был разработан модуль с использованием регулярных выражений и программы MyStem. Модуль берет текст из заданных элементов XML-дерева (title, text), регулярные выражения из текста извлекают все последовательности букв русского алфавита. При подсчете слов последовательность букв русского алфавита, отделенная от других букв не буквами (например, знаки препинания, пробел), считается словом. Для определения лемм слов использовалась программа MyStem. Программа MyStem производит морфологический анализ текста на русском языке. Для слов, отсутствующих в словаре, порождаются гипотезы [21].

6. Разметка корпуса SCTM-ru. В качестве формата хранения корпуса SCTM-ru выбран XML (eXtensible Markup Language — расширяемый язык разметки), как один из наиболее удобных форматов для использования в программной среде и конвертации данных в другие форматы. Возможности XML позволяют сохранить текст исходной статьи Викиновости и выделить дополнительные параметры документа.

XML-файл корпуса (SCTM-ru) состоит из следующих элементов:

- `<page>` - группа элементов документа;
- `<title>` - название документа;
- `<id>` - уникальный идентификатор документа;
- `<userid>` - уникальный идентификатор автора;
- `<category>` - категория документа;
- `<date>` - дата событий документа;
- `<text>` - текст документа;

Пример разметки одного документа в корпусе SCTM-ru представлен на рисунке 5.

```
<?xml version="1.0" encoding="utf-8"?>
<page>
  <title>50 000 статей в русской Википедии</title>
  <id>1838</id>
  <userid>2156</userid>
  <category>Русская Википедия</category>
  <data>24 декабря 2005</data>
  <text>
    Русскоязычный раздел Википедии (http://ru.wikipedia.org) в предверии
    нового 2006 года отметил своеобразный юбилей, преодолев в ночь на 24
    декабря 2005 года рубеж в 50 тысяч статей. Примечательно, что ровно
    год назад, накануне нового 2005 года, был преодолен первый
    «психологический» барьер в 10 тыс. статей. Таким образом, за год число
    статей выросло в пять раз, русский раздел продвинулся в рейтинге по
    числу статей вверх на 9 позиций и сейчас занимает 12 место по этому
    показателю.
  </text>
</page>
```

Рис. 5. Пример XML-документа "50 000 статей в русской Википедии" в корпусе SCTM-ru

Заголовок документа (*title*) отделен от текста документа, т.к. словам заголовка может придаваться большее значение при построении тематической модели.

Уникальный идентификатор автора статьи (*userid*) – это параметр, который необходим в автор-тематических моделях. Автор-тематическая модель во времени (Author-Topic over Time) [22] представляет собой расширение LDA при построении модели оценивается распределение авторов, тем и документов по времени.

Категории документа (*category*) – это указанные автором статьи категории. Например, на рис. 4 в статье "50 000 статей в русской Википедии" указана категория «Русская Википедия». Информация о категории важна для тематического моделирования, поэтому сохранена в корпусе SCTM-ru см. рис. 5. Наличие информации о принадлежности документов к категориям позволит автоматически проверять точность, полноту, аккуратность тестируемых алгоритмов тематического моделирования. Информация о категориях документа может быть использована в моделях Labeled LDA, описанных в [23].

Дата описанных в статье событий (*date*) используется при построении временных (temporal) тематических моделей. Пример модели, использующей дату под названием «Тематики во времени» (Topic over Time - TOT) представлен в работе [24]. При построении временной модели наряду со стандартными распределениями слов по темам и тем по документам оцениваются

распределения каждой темы по времени, что позволяет проследить и отобразить динамику изменения тем во времени.

Текст документа (*text*) соответствует тексту исходной статьи. Мы целенаправленно оставляем исходный текст без изменения, без преобразования его в модель «мешка слов», без лингвистической обработки для возможности исследования уникальных особенностей русского языка. Информация о последовательности слов в тексте документа используется в моделях, учитывающих взаимную встречаемость слов. Например, модель под названием «Скрытая тематическая Марковская модель» (Hidden Topic Markov's Model - HTMM), описанная в работе [25], основана на предположениях, что слова в составе предложения, а также сами предложения связаны одной общей темой и темы слов в документе образуют цепь Маркова. В результате работы HTMM уменьшает неоднозначность слов, расширяет понимание темы.

7. Заключение. В результате проделанной работы был подготовлен специальный русскоязычный корпус текстов (SCTM-ru), подходящий для тестирования различных алгоритмов вероятностного тематического моделирования. Поставленные в работе цели были достигнуты: корпус SCTM-ru содержит оригиналы текстов документов на русском языке, информацию о дате описанных в документе событиях, информацию об авторе и категориях, к которым относится документ, доступен для скачивания и использования на устройствах без доступа в Интернет.

Источником данных корпуса является международный новостной сайт «Русские Викиновости». Корпус SCTM-ru состоит из 7 тыс. документов, 185 авторов, почти 12 тыс. уникальных категорий. События, описанные в документах, распределены по более чем 2 тыс. уникальным датам, с ноября 2005 года по июнь 2014 года. В корпусе SCTM-ru 2,4 млн словоупотреблений, состоящих только из русских букв. Словарный состав корпуса – 150,6 тыс. уникальных словоформ, 59 тыс. уникальных лемм.

Объем созданного корпуса дает основания предположить его репрезентативность для различных задач автоматической обработки текстов на естественном языке. Как отмечено в работе [26] «Неразумно ждать пока кто-то по-научному сбалансирует корпус, перед тем как его использовать, и неосмотрительно было бы оценивать результаты анализа корпуса как «малодостоверные» или «неуместные» просто потому, что нельзя доказать, что используемый корпус «сбалансирован»». Разнообразие описанных в корпусе SCTM-ru событий и огромный коллектив авторов статей (21 тыс. участников)

обосновывают предположение о его сбалансированности. Убедиться в сбалансированности корпуса можно после проведения анализа его внутренних признаков и построения тематических моделей.

Предложенная технология создания корпуса текстов для задач тематического моделирования позволяет расширять корпус SCTM-ru за счет новых статей. Аналогичным образом могут быть созданы языковые корпуса на любом из 33-х представленных в Викиновостях языках. В предложенном формате могут быть созданы коллекции и корпуса, из различных источников данных, при этом должна быть сохранена только востребованная в алгоритмах тематического моделирования информация.

Далее на базе созданного корпуса будут исследованы особенности существующих вариаций алгоритмов тематического моделирования, будут разработаны новые алгоритмы, учитывающие лингвистические особенности русского языка. Корпус SCTM-ru распространяется по открытой лицензии, доступен для скачивания на сайте www.cims.ru.

Литература

1. *Papadimitriou C.H., Raghavan P., Tamaki H., Vempala S.* Latent semantic indexing: A probabilistic analysis. 1998. pp. 159–168.
2. *Hoffman T.* Probabilistic Latent Semantic Indexing // Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval. 1999. pp. 50–57.
3. *Blei D.M., Ng A.Y., Jordan M.I.* Latent Dirichlet Allocation // Journal of Machine Learning Research. 2003. pp. 993–1022.
4. *Daud A., Li J., Zhou L., Muhammad F.* Knowledge discovery through directed probabilistic topic models: a survey // In Proceedings of Frontiers of Computer Science in China. 2010. pp. 280–301.
5. Сайт Национального корпуса русского языка НКРЯ. URL: www.ruscorpora.ru. (дата обращения: 12.01.2015).
6. *Захаров В.П.* Международные стандарты в области корпусной лингвистики // Структурная и прикладная лингвистика. 2012. № 9. С. 201–221.
7. *Крижановский А.А., Смирнов А.В.* Подход к автоматизированному построению общецелевой лексической онтологии на основе данных викисловаря // Известия РАН. Теория и системы управления. 2013. № 2. С. 53–63.
8. *Смирнов А.В., Круглов В.М., Крижановский А.А., Луговая Н.Б., Карпов А.А., Кипяткова И.С.* Количественный анализ лексики русского WordNet и викисловарей // Труды СПИИРАН. 2012. Вып. 23. С. 231–253.
9. *Грановский Д.В., Бочаров В.В., Бичинева С.В.* Открытый корпус: принципы работы и перспективы // Компьютерная лингвистика и развитие семантического поиска в Интернете: Труды научного семинара XIII Всероссийской объединенной конференции «Интернет и современное общество». Санкт-Петербург. 2010 г. СПб. 2010. 94 с.
10. Сайт Открытого корпуса. URL: opencorpora.org (дата обращения: 10.01.2015).
11. Small corpus of Associated Press. URL: www.cs.princeton.edu/~blei/lda-c/ (дата обращения: 06.01.2015).

12. The New York Times Annotated Corpus. URL: catalog.ldc.upenn.edu/LDC2008T19 (дата обращения: 14.01.2015).
13. The 20 Newsgroups data set. URL: qwone.com/~jason/20Newsgroups/ (дата обращения: 24.01.2015).
14. Reuters Corpora. URL: trec.nist.gov/data/reuters/reuters.html (дата обращения: 24.01.2015).
15. Reuters-21578 Text Categorization Collection Data Set. URL: archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection (дата обращения: 24.01.2015).
16. *Виноградова В.Б., Кукушкина О.В., Поликарпов А.А., Савчук С.О.* Компьютерный корпус текстов русских газет конца 20-го века: создание, категоризация, автоматизированный анализ языковых особенностей // "Русский язык: исторические судьбы и современность" Международный конгресс русистов-исследователей. Труды и материалы. М.: Изд-во Моск. ун-та. 2001. С. 114–115.
17. Компьютерный корпус текстов русских газет конца XX-ого века. URL: www.philol.msu.ru/~lex/corpus/corp_descr.html (дата обращения: 24.01.2015).
18. *Венцов А.В., Грудева Е.В.* О корпусе русского литературного языка (narusco.ru) // Русская Лингвистика. 2009. Том 33. № 2. С. 195–209.
19. Корпус русского литературного языка. URL: www.narusco.ru (дата обращения: 24.01.2015).
20. Хельсинкский аннотированный корпус русских текстов ХАНКО. URL: www.helsinki.fi/venaja/russian/e-material/hanco/index.htm (дата обращения: 24.01.2015).
21. Официальный сайт программы морфологического анализа текстов на русском языке MyStem. URL: api.yandex.ru/mystem/ (дата обращения: 12.12.2014).
22. *Xu S., Shi Q., Qiao X., et al*. Author-Topic over Time (AToT): a dynamic users' interest model, in Mobile, Ubiquitous, and Intelligent Computing // Springer. Berlin. 2014. pp. 239–245.
23. *Ramage D., Hall D., Nallapati R., Manning C.D*. Labeled LDA. A supervised topic model for credit attribution in multi-labeled corpora // In Empirical Methods in Natural Language Processing. 2009. pp. 248–256.
24. *Wang X., McCallum A*. Topics over Time: A Non-Markov Continuous Time Model of Topical Trends // In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia. USA. 2006.
25. *Gruber A., Rosen-Zvi M., Weiss Y.* Hidden Topic Markov Models. In: Proceedings of Artificial Intelligence and Statistics (AISTATS) // San Juan. Puerto Rico. USA. 2007.
26. *Захаров В.П., Азарова И.В.* Параметризация специальных корпусов текстов // Структурная и прикладная лингвистика: Межвузовский сборник. СПб: СПбГУ. 2012. Вып. 9. С. 176–184.

References

1. Papadimitriou C.H., Raghavan P., Tamaki H., Vempala S. Latent semantic indexing: A probabilistic analysis. 1998. pp. 159–168.
2. Hoffman T. Probabilistic Latent Semantic Indexing. Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval. 1999. pp. 50–57.
3. Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet Allocation. Journal of Machine Learning Research. 2003. pp. 993–1022.
4. Daud A., Li J., Zhou L., Muhammad F. Knowledge discovery through directed probabilistic topic models: a survey. In Proceedings of Frontiers of Computer Science in China. 2010. pp. 280–301.

5. Sajt Nacional'nogo korpusa russkogo jazyka NKRJa. [Website of Russian National Corpus]. Available at: www.ruscorpora.ru (accessed: 12.01.2015). (In Russ.).
6. Zakharov V.P. [International standards in corpora linguistics]. *Strukturnaja i prikladnaja lingvistika – Structural and Applied Linguistics*. 2012. vol. 9. pp. 201–221. (In Russ.).
7. Smirnov A.V., Krizhanovsky A.A. [An approach to automated construction of a general-purpose lexical ontology based on Wiktionary]. *Izvestija RAN. Teorija i sistemy upravlenija – Journal of Computer and Systems Sciences International.* 2013. vol. 52. no. 2. pp. 215–225. (In Russ.).
8. Smirnov A.V., Kruglov V. M., Krizhanovsky A.A., Lugovaya N.B., Karpov A.A., Kipyatkova I.S. [A quantitative analysis of the lexicon in Russian WordNet and Wiktionaries]. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2012. vol. 23. pp. 231–253. (In Russ.).
9. Granovsky D.V., Bocharov V.V., Bichineva S.V. [Opencorpora: how it work and perspectives]. *Kompyuternaya lingvistika i razvitie semanticheskogo poiska v internete: Trudy nauchnogo seminara XIII vserossiyskoy obedinennoy konferencii «internet i sovremennoe obschestvo»* [Computer linguistics and development of semantic search on Internet: Proceedings of the 13th All-Russian integrated conference «Internet and Modern Society»]. St. Petersburg. 2010. 94 p. (In Russ.).
10. Sajt Otkrytogo korpusa [OpenCorpora Website]. Available at: opencorpora.org (accessed: 15.01.2015). (In Russ.).
11. Small corpus of Associated Press. Available at: www.cs.princeton.edu/~blei/lda-c/ (accessed: 6.01.2015).
12. The New York Times Annotated Corpus. Available at: catalog.ldc.upenn.edu/LDC2008T19 (accessed: 14.01.2015).
13. The 20 Newsgroups data set. Available at: qwone.com/~jason/20Newsgroups/ (accessed: 24.01.2015).
14. Reuters Corpora. Available at: trec.nist.gov/data/reuters/reuters.html (accessed: 24.01.2015).
15. Reuters-21578 Text Categorization Collection Data Set. Available at: archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection (accessed: 24.01.2015).
16. Vinogradova V.B., Kukushkina O.V., Polikarpov A.A., Savchuk S.O. [The computer corpus of Russian newspapers of the XX th century end: the creation, categorization, automated analysis of linguistic features]. *"Russkiy yazyk: istoricheskie sudby I sovremennost." Mezhdunarodnyy congress rusistov issledovateley. Moskva, filologicheskiy f t MGU im. M.V. Lomonosova* ["Russian Language: its Historical Destiny and Present State" International Congress of Russian Language Researchers]. M.: University Pressю 2001. pp. 114–115. (In Russ.).
17. Komp'juternyj korpus tekstov russkih gazet konca XX-ogo veka [The computer corpus of Russian newspapers of the XXth century end]. Available at: www.philol.msu.ru/~lex/corpus/corp_descr.html (accessed: 24.01.2015). (In Russ.).
18. Vencov A.V., Grudeva E.V. [About Corpus of Standard Written Russian (narusco.ru)]. *Russkaja Lingvistika – Russian Linguistics*. 2009. vol. 33. no. 2. pp. 195–209. (In Russ.).
19. Korpus russkogo literaturnogo jazyka [Corpus of Standard Written Russian]. Available at: www.narusco.ru (accessed: 24.01.2015). (In Russ.).
20. Hel'sinkskij annotirovannyj korpus russkih tekstov HANKO [HANCO Corpus]. Available at: www.helsinki.fi/venaja/russian/e-material/hanco/index.htm (accessed: 24.01.2015).

21. Oficial'nyj sajt programmy morfologicheskogo analiza tekstov na russkom jazyke MyStem [System for automatic morphological analysis of Russian MyStem]. Available at: api.yandex.ru/mystem/ (accessed: 12.12.2014). (In Russ).
22. Xu S., Shi Q., Qiao X., et al. Author-Topic over Time (AToT): a dynamic users' interest model, in Mobile, Ubiquitous, and Intelligent Computing. Springer. Berlin. 2014. pp. 239–245.
23. Ramage D., Hall D., Nallapati R., Manning C.D. Labeled LDA. A supervised topic model for credit attribution in multi-labeled corpora. In Empirical Methods in Natural Language Processing. 2009. pp. 248–256.
24. Wang X., McCallum A. Topics over Time: A Non-Markov Continuous Time Model of Topical Trends. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia. USA. 2006.
25. Gruber A., Rosen-Zvi M., Weiss Y. Hidden Topic Markov Models. In Proceedings of Artificial Intelligence and Statistics (AISTATS). San Juan. Puerto Rico. USA. 2007.
26. Zakharov V.P., Azarova I.V. [Special text corpora parametrization]. *Strukturnaya i prikladnaya lingvistika: mezhvuzovskiy sbornik* – Structural and Applied Linguistics: Interuniversity collection. SPb. St. Petersburg State University. 2012. vol. 9. pp 176–184. (In Russ.).

Карпович Сергей Николаевич — руководитель отдела поисковой оптимизации ООО "Рамблер Интернет Холдинг". Область научных интересов: тематическое моделирование, обработка текстов на естественном языке, кластеризация, классификация, обработка данных, машинное обучение. Число научных публикаций — 1. cims@yandex.ru, http://www.cims.ru/; 117105, Москва, Варшавское ш., 9, стр. 1, БЦ «Даниловская мануфактура», корпус «Ряды Солдатенкова»; р.т. +7(495)7851700.

Karpovich Sergey Nikolaevich — head of Search Engine Optimization Rambler Internet Holding LLC. Research interests: topic model, natural language processing, classification, clustering, data mining. The number of publications — 1. cims@yandex.ru, http://www.cims.ru/; Varshavskoe sh., 9, str. 1, BC «Danilovskaja manufaktura», korpus «Rjady Soldatenkova», 117105, Moscow; office phone +7(495)7851700.

РЕФЕРАТ

Карпович С.Н. **Русскоязычный корпус текстов SCTM-ru для построения тематических моделей**

В статье предложен специальный корпус текстов для тестирования алгоритмов тематического моделирования SCTM-ru. В условиях стремительного роста количества информационных данных, остро проявляется проблема разработки инструментов и систем для их автоматической обработки. Для создания систем и тестирования алгоритмов должны существовать подходящие наборы данных. Необходимо наличие свободных коллекций документов, текстовых корпусов на русском языке, для исследований методов автоматической обработки текстов на естественном языке, с учетом лингвистических особенностей языка. Обозначены требования к специальному корпусу: он должен распространяться по свободной лицензии, количество документов должно быть достаточным для исследования, должен содержать текста документов на естественном языке, должен содержать востребованную в алгоритмах тематического моделирования информацию. Проведен сравнительный анализ корпусов на русском и иностранных языках, выявлено несоответствие характеристик существующих корпусов с обозначенным требованиям.

Описана технология создания корпуса, выбор подходящего источника данных, этап предварительной обработки текстов документов, разметка корпуса и обеспечение доступа. Источником данных корпуса является международный новостной сайт «Русские Викиновости». Корпус SCTM-ru состоит из 7009 документов, 185 авторов, 11 895 уникальных категорий. События, описанные в документах, распределены по 2 236 уникальным датам, с ноября 2005 года по июнь 2014 года. В корпусе SCTM-ru 2,4 млн словоупотреблений, состоящих только из русских букв. Словарный состав корпуса – 150,6 тыс. уникальных словоформ, 59 тыс. уникальных лемм. Корпус репрезентативен. Убедиться в сбалансированности корпуса предлагается в ходе его исследования.

Разработанный подход создания корпуса позволяет постоянно расширять корпус SCTM-ru за счет новых статей. Аналогичным образом может быть подготовлен языковой корпус на любом из 33-х представленных в Викиновостях языках. Предложенная технология подготовки корпуса текстов для задач тематического моделирования позволяет создавать коллекции и корпуса из данных, полученных из различных источников, при этом будет сохранена только востребованная в алгоритмах тематического моделирования информация. Далее на базе созданного корпуса будут исследованы особенности существующих вариаций алгоритмов тематического моделирования, будут разработаны новые алгоритмы, учитывающие лингвистические особенности русского языка.

SUMMARY

Karpovich S.N. **The Russian language text corpus for testing algorithms of topic model.**

This paper proposes a special text corpus SCTM-ru to be used for testing algorithms of probabilistic topic model. With rapidly increasing amounts of information, there is a critical need for tools and systems to be able to automatically process them. To create the systems and to test the algorithms require suitable data sets. We need free document libraries, text corpora in Russian to research into the methods of natural language processing taking into account the linguistic features of the language. The requirements for a special corpus have been defined: it should be distributed under a free license, contain enough documents for the research, with the text in the documents being in natural languages, and contain relevant information for the topic modelling. A comparative analysis of corpora in Russian and foreign languages has been done revealing a non-compliance of the existing corpora with the above requirements.

The article describes a technology for the creation of such a corpus, how to choose a suitable data source, a stage for the preprocessing of the texts, and how to format and provide access to the corpus. The data source for the corpus is the international news website Russian Wikinews. The SCTN-ru corpus contains 7,009 documents written by 185 authors and split into 11,895 unique categories. The events described in the documents cover 2,236 unique dates, from November 2005 to June 2014. There are 2.4 million tokens consisting only in Russian letters in SCTM-ru corpus. The corpus contains 150.6 thousand unique word forms and 59 thousand unique lemmas. The corpus is representative. The readers are invited to see that the corpus is balanced during its analysis.

The developed approach to the creation of the corpus allows SCTM-ru to be constantly expanded with new articles. In a similar way, a corpus in any of 33 languages presented in Wikinews can be created. The proposed technology of the creation of text corpora for the topic modelling makes it possible to create collections and corpora using data obtained from different sources, with only relevant for the topic model information being saved. Next, the created corpus will be used as a basis for the research into the features of the existing variations of the topic modelling, and new algorithms taking into account the linguistic features of Russian language will be developed.