The Russian Language Text Corpus for Testing Algorithms of Topic Model

Karpovich S. N.
JSC “Olimp”
Moscow, Russia
cims@yandex.ru

***Abstract.* This paper proposes a special corpus for testing algorithms Topic model SCTM-ru. In the conditions of the prompt growth of quantity of data, the problem of development of tools and systems for their automatic processing. To create systems and testing algorithms should be suitable datasets. Existence of free collections of documents, text corpora in Russian, is necessary for researches methods of natural language processing, considering linguistic features of language. Designated special housing requirements: must be distributed under a free license, the number of documents should be sufficient for the study, must include the text of documents in natural language should contain demanded algorithms Topic model information. The comparative analysis of corpus in Russian and foreign languages is carried out, discrepancy of characteristics of the existing corpus with the designated requirements is revealed.**

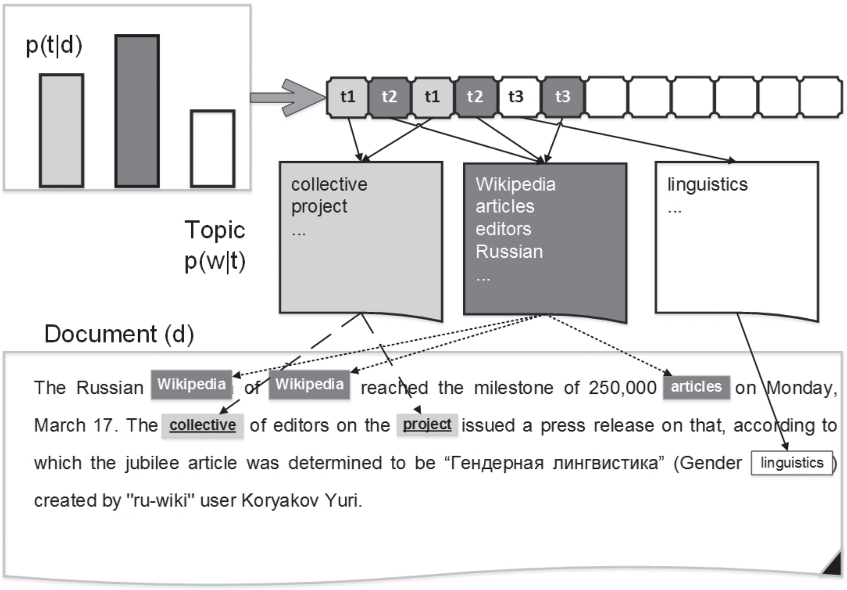
***Keywords:* text corpora, topic model, natural language processing, Russian language.**

Introduction

Information is becoming the main product and commodity of the contemporary society. The following areas are in the process of active development: science, economics, politics, and manufacturing; digital data is being generated and accumulated in many fields. To successfully retrieve and process information from data, one shall have proper tools, systems and algorithms. A demand for Natural Language Processing systems is growing. Natural Language Processing is already used in common software and services. For instance, software for reading news feeds is capable of grouping news by topics, search engines find documents with information of value for the user and mailing services filter spam messages automatically. Various algorithms are used for clustering and classification of text data; most popular are k-means, SVM, neural networks. An upcoming trend in automatic processing of texts is development of probabilistic topic modelling algorithms.

Topic modelling is a method of building of a topic model of a text documents collection. The topic model sets the ratio between topics and documents in the corpus of texts. For the first-time topic modelling was described in the paper by C. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala in 1998 [1]. Thomas Hoffman in 1999 proposed probabilistic latent semantic indexing (PLSI) [2]. One of most popular topic models is Latent Dirichlet Allocation (LDA), this model is the generalization of probabilistic semantic indexing and was developed by David Blei, Andrew Ng and Michael I. Jordan in 2002 [3]. Other topic models are usually an extension to LDA. Fig. 1 is an example of building a topic model of a document.

Algorithms of topic modelling are oriented at the work with a natural language text. Initial solutions were based on a suggestion that text is a “bag of words”, i. e. word order in the text is of no value. Further models have successfully implemented algorithms that consider dependencies between words with the help of latent Markovian models. The review [4] considers five primary



The Russian Wikipedia of Wikipedia reached the milestone of 250,000 articles on Monday, March 17. The collective of editors on the project issued a press release on that, according to which the jubilee article was determined to be “Гендерная лингвистика” (Gender linguistics) created by "ru-wiki" user Koryakov Yuri.

Fig. 1. Building a topic model of a document: p(w|t) matrix of sought for conditional distributions of words by topics; p(t|d) matrix of sought for conditional distributions of topics by documents; d document; w word; d, w observable variables; t topic (latent variable)

classes of probabilistic topic models: basic, taking into account relations between documents, taking into account relations between words, temporary, teacher-taught.

Availability of text corpora will make it possible to develop systems for automatic Natural Language Processing, as well as topic modelling algorithms. When developing a topic model, one shall consider language features of texts. A Russian text corpus, which is distributed through a free license, is necessary for development of topic modelling methods operating with the Russian language.

Corpus linguistics is a complicated linguistic discipline, which has formed in the last decades based on computer equipment. It studies the construction of linguistic corpora, methods of data processing in them and their generation and usage technology proper. "A corpus is a reference system based on an electronic collection of texts composed in a certain language" – such definition of the text corpus is available on the website of the Russian National Corpus [5]. The study [6] notes the following: "Corpora, as a rule, are designed for repeated use by many users; therefore, their markup and their linguistic support shall be unified in a certain manner". Feasibility of establishment and sense behind using the corpus depend on the following premises:

1) sufficient (representative) volume of the corpus;

2) data of different type is contained within the corpus in their natural context form;

3) once developed and prepared, data store may be used repeatedly.

First-order corpora are a collection of texts with a common feature, for instance, source, author, place of publication. A special text corpus is a balanced corpus, representative, as a rule, small, identifiable to a certain research task and designed for use mainly for purposes that are in line with composer's ideas. Text corpora or corpus, large collection of documents, dataset, as specified in the paper [4], are synonymic concepts.

The purpose of this paper is to create a special Russian text corpus SCTM-ru, suitable for study of algorithms of probabilistic topic modelling. Let us specify requirements to the created corpus. The corpus shall be distributed through a free license, the quantity of documents shall be sufficient for research, and it shall contain the following:

- original text of documents;
- dates of described events;
- authorship information;
- topics.

Let us consider the opportunity of using existing text corpora for purposes of testing topic modelling algorithms.

Review of text corpora and datasets

The Russian National Corpus (RNC) [5] contains more than 335 K documents in Russian, which divided into sub corpora. It includes 180 K texts of the Newspaper Corpus. A license agreement shall be signed to use the offline version of the main corpus (1 M tokens).

The OpenCorpora [7, 8] contains around 3 K documents in Russian, 93 K marked sentences, 10 data sources, some documents contain information about the author and date of described events. Linguistic information, such as morphological, semantic and syntactic, is assigned to text parts. The corpus is not suitable for objectives of building temporal and author-based models, since not all documents of the corpus contain information about the author and the date of the events.

The Associated Press [9] corpus contains 2 K documents in English. Corpus documents are not labelled with a date of a described event, publication author, and document category. The corpus is applicable to research a limited quantity of topic modelling algorithms.

The New York Times Annotated Corpus [10] is a large English text corpus of newspaper articles and news distributed through a closed license.

20 Newsgroups [11] is a collection of news in English, prepared to research algorithms of automatic text processing. 20 Newsgroups contains around 20 K documents. Data about authors and date of publication that is important for building topic models is not marked up. Preliminary processing of news text is required for use in topic modelling.

Reuters Corpora [12] is a large English news corpus. Three datasets contain more than 3 M news. It is distributed through a limited license only for scientific research and is provided only upon signature of the license agreement. There is an earlier version of the corpus named Reuters-21578 [13], which is popular for testing algorithms of automatic text processing and is distributed through a limited license, being available for offline analysis.

The computer corpus of texts from Russian newspapers of the end of the 20th century [14, 15] was established in 1999; it is currently developed and researched using grants of the Russian Foundation for Basic Research. The corpus is designed to analyses linguistic features (vocabulary, morphology, syntax, phraseology, stylistics) of the contemporary newspaper language. The corpus contains 23 K texts of full issues of 13 different Russian newspapers in Russian. The corpus includes 11 M tokens.

The corpus of the Russian literary language [16, 17] is represented in the form of an array of morphologically annotated texts in the Russian literary language. The corpus includes more than 1 M tokens with a balanced genre composition.

The Helsinki Annotated Corpus of Russian texts HANCO [18] – the corpus contains morphological, syntactic and functional information about texts with total volume of 100 K texts retrieved from the magazine "Itogi". Rights to full texts of magazine articles belong to owners.

Table represents comparison of algorithms of topic modelling of corpus characteristics that are important for research: corpus language, distribution license, availability for download and research on computers with no access to Internet, data on author, data on date of described events, text topic. Considered text corpora do not fully meet the requirements specified in this paper. The developed special corpus for topic modelling SCTM-ru is distributed through a free license, the language of the corpus is Russian, it contains data on authorship, the date of the events, topic affiliation of documents, is available for download and research on computers with no access to Internet.

Technology of SCTM-ru corpus development

The technological process of corpus development includes the following steps:

1) source detection;

2) preliminary processing of document texts;

3) markup of parameters of each document in the corpus;

4) provision of access to the corpus.

Table. Comparative table of text corpus characteristics

Corpus	Language	Open license	Available for download	Data on author	Data on date	Topics
RNC	Russian			+		
Open Corpus	Russian	+	+	+	+	
Associated Press	English	+	+			
The New York Times Annotated Corpus	English			+	+	+
20 News groups	English	+	+			
Reuters Corpora	English			+	+	+
Russian Literary Language Corpus	Russian					
HANCO	Russian					
SCTM ru	Russian	+	+	+	+	+

In accordance with the indicated requirement to availability of corpus data, texts used as content shall be distributed through a free license, available for download and free use.

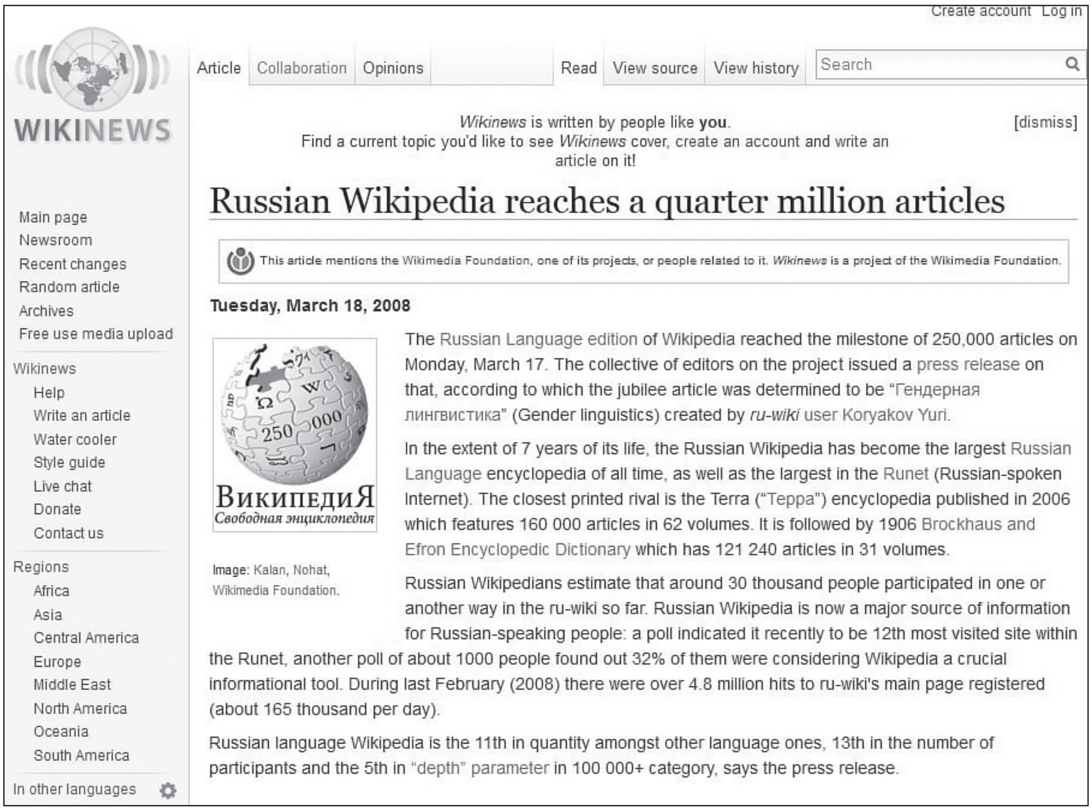
As result of preliminary processing of texts and markup of parameters of each document, the corpus shall store and mark up in a certain way the information necessary to construct topic models. Information that was unclaimed in topic modelling shall be excluded from the corpus, as useless.

Various objectives of topic modelling may require a certain procedure of data arrival into the topic modelling system, from successive for temporal models, to one-off for regular topic models. Therefore, to provide access to the corpus, it is sufficient to provide an opportunity for its download and subsequent use in accordance with specific objectives of the researcher.

Source of data for SCTM-ru corpus

We suggest using the international news website “Russian Wikinews” (Wikinews) as a data source, where texts of articles are distributed through a free license of Creative Commons Attribution 2.5 Generic, are available for download and analysis on any computers, including computers with no access to Internet. Papers [19, 20] specify advantages of wiki-resources, such as Wikidictionary and Wikipedia, for use as a source of data for research purposes. Wiki-resources are websites of the second generation of Internet characterized by the fact that many ordinary users were involved to develop their content, and those users help to expand them and update information. Large volume, continuous expansion, neutrality of opinions, and availability are among the advantages of all wiki-resources, including Wikinews.

Wikinews is a brother project of large Wikipedia designed to write news articles. The example of a Wikinews article is shown in fig. 2. A signature feature of the Wikinews website compared to any other news website is the fact that any person may take part in creating a piece of news. Rules of Wikinews require writing news from a neutral point of view, in unbiased form, selecting material and relevant topics, using valid sources.



Wikinews is written by people like **you**.
Find a current topic you'd like to see *Wikinews* cover, create an account and write an article on it! [dismiss]

Russian Wikipedia reaches a quarter million articles

This article mentions the Wikimedia Foundation, one of its projects, or people related to it. *Wikinews* is a project of the Wikimedia Foundation.

Tuesday, March 18, 2008

Image: Kalan, Nohat, Wikimedia Foundation.

The Russian Language edition of Wikipedia reached the milestone of 250,000 articles on Monday, March 17. The collective of editors on the project issued a press release on that, according to which the jubilee article was determined to be “Гендерная лингвистика” (Gender linguistics) created by *ru-wiki* user Koryakov Yuri.

In the extent of 7 years of its life, the Russian Wikipedia has become the largest Russian Language encyclopedia of all time, as well as the largest in the Runet (Russian-spoken Internet). The closest printed rival is the Terra (“Терра”) encyclopedia published in 2006 which features 160 000 articles in 62 volumes. It is followed by 1906 Brockhaus and Efron Encyclopedic Dictionary which has 121 240 articles in 31 volumes.

Russian Wikipedians estimate that around 30 thousand people participated in one or another way in the ru-wiki so far. Russian Wikipedia is now a major source of information for Russian-speaking people: a poll indicated it recently to be 12th most visited site within the Runet, another poll of about 1000 people found out 32% of them were considering Wikipedia a crucial informational tool. During last February (2008) there were over 4.8 million hits to ru-wiki's main page registered (about 165 thousand per day).

Russian language Wikipedia is the 11th in quantity amongst other language ones, 13th in the number of participants and the 5th in “depth” parameter in 100 000+ category, says the press release.

Fig. 2. Article „50,000 Articles in Russian Wikipedia“ on website of Russian Wikinews

XML-file of Wikinews data base export includes the following XML-elements:

+ <page> – group of news article elements;

+ <title> – article title;

– <ns> – identifier or name of a namespace, the element is intended to separate primary articles from internal ones, zero corresponds to the primary namespace;

+ <id> – unique article ID;

+ <revision> – group of elements of the relevant article version;

– <id> – primary revision key used to monitor article changes;

– <parented> – parent article ID;

– <timestamp> – date and time of article revision creation;

+ <contributor> – group of article authorship elements;

– <username> – article author name;

+ <id> – unique article author ID;

+ <text> – article text with elements of wiki-markup;

– <sha1> – article hash code produced by Secure Hash Algorithm SHA-1, used to monitor versions;

– <model> – model of article content, in this case wikitext;

– <format> – format of article data, in this case text/x-wiki.

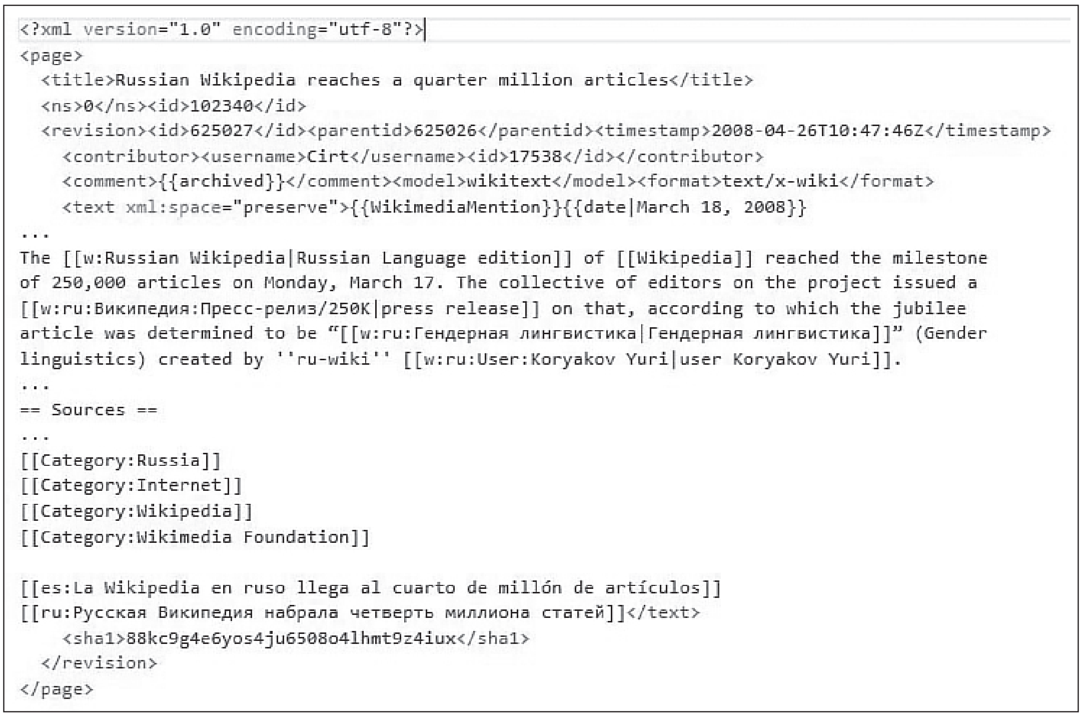
For topic modelling objectives the information, which is contained in elements marked with (+), is necessary. Elements that contain information, which is not used in topic modelling algorithms are marked with (–). Example of a part of XML-tree of Wikinews data base export file is shown in fig. 3.

Preliminary processing of Wikinews data

In the export file of Wikinews the articles are sorted according to the revision creation date <timestamp>, this date is not related to the date of the described events. Authors are recommended to specify the date of the events in the article text, using wiki-markup. The example of wiki-markup of the date {{: Date | December 24, 2005}} is shown in fig. 4 inside the element <text>. Some articles in the export file of Wikinews does not contain the date of the events in the wiki-markup, but at the same time it is specified in the text or in the category. To save maximum information required for topic modelling algorithms, the date of the events was, where possible, restored from the text and categories. In 455 articles it was not possible to restore the date of the events, these articles are selections of news that occurred on the same day, in different years, and present no value for constructing topic models, and they were excluded from the corpus. Documents of the SCTM-ru corpus are sorted by the date of the events, from old to new ones.

The export file of the Wikinews database contains information about the author of the last article revision. We use such information as authorship ID for constructing author-topic models. Since 58 pieces of Wikinews do not contain the author data, and articles are valuable, the technical decision was made to assign a unique author ID – 2 – to these articles and include the latter into the SCTM-ru corpus.

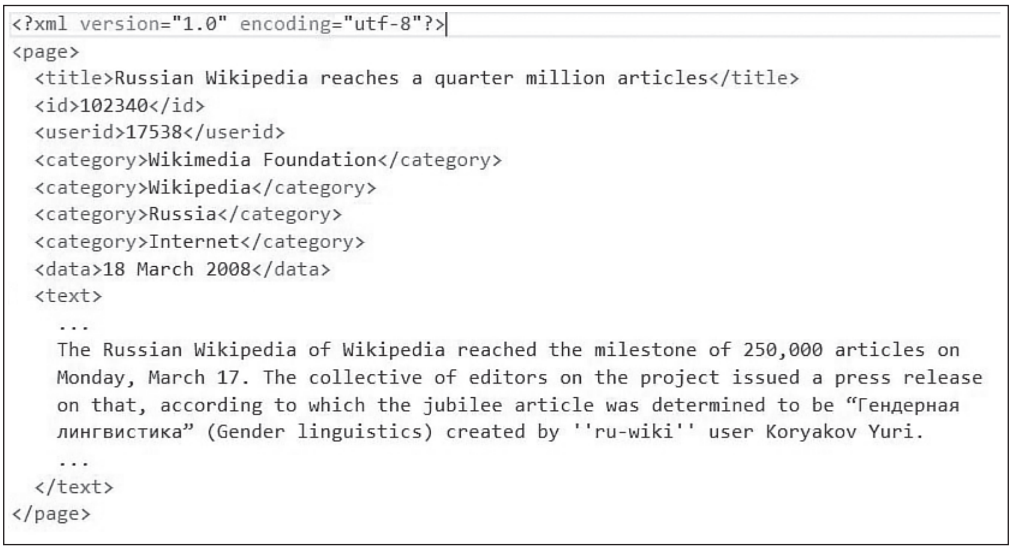
The text of the Wikinews article contains references that are arranged in a certain way. References are divided into three groups: internal – a tool to relate pages inside the language section of Wikipedia, interlingual links (interwiki) – means to organize connections between various wiki-systems in Internet and references to pages of brotherly wiki-projects (for instance, to Wikipedia). The article text enclosed within double square brackets is an internal reference. If the case of the referring word or token does not match with the nominative case, there is a line within double square brackets, to the left of which there is the nominative case of the reference text, and to the right – the text that corresponds to the sentence grammar. Topic modelling algorithms consider the number of each word lemma entries into the text, in internal references each word has two entries in different wordforms and will be taken into account twice in the topic model, thus having perverted frequency characteristics of the model. Documents of the SCTM-ru corpus contain



```
<?xml version="1.0" encoding="utf-8"?>
<page>
  <title>Russian Wikipedia reaches a quarter million articles</title>
  <ns>0</ns><id>102340</id>
  <revision><id>625027</id><parentid>625026</parentid><timestamp>2008-04-26T10:47:46Z</timestamp>
    <contributor><username>Cirt</username><id>17538</id></contributor>
    <comment>{{archived}}</comment><model>wikitext</model><format>text/x-wiki</format>
    <text xml:space="preserve">{{WikimediaMention}}{{date|March 18, 2008}}
...
The [[w:Russian Wikipedia|Russian Language edition]] of [[Wikipedia]] reached the milestone
of 250,000 articles on Monday, March 17. The collective of editors on the project issued a
[[w:ru:Википедия:Пресс-релиз/250K|press release]] on that, according to which the jubilee
article was determined to be “[[w:ru:Гендерная лингвистика|Гендерная лингвистика]]” (Gender
linguistics) created by ''ru-wiki'' [[w:ru:User:Koryakov Yuri|user Koryakov Yuri]].
...
== Sources ==
...
[[Category:Russia]]
[[Category:Internet]]
[[Category:Wikipedia]]
[[Category:Wikimedia Foundation]]

[[es:La Wikipedia en ruso llega al cuarto de millón de artículos]]
[[ru:Русская Википедия набрала четверть миллиона статей]]</text>
    <sha1>88kc9g4e6yos4ju6508o4lhmt9z4iux</sha1>
  </revision>
</page>
```

Fig. 3. Example of XML article „50,000 Articles in Russian Wikipedia“ on website of Russian Wikinews



```
<?xml version="1.0" encoding="utf-8"?>
<page>
  <title>Russian Wikipedia reaches a quarter million articles</title>
  <id>102340</id>
  <userid>17538</userid>
  <category>Wikimedia Foundation</category>
  <category>Wikipedia</category>
  <category>Russia</category>
  <category>Internet</category>
  <data>18 March 2008</data>
  <text>
    ...
    The Russian Wikipedia of Wikipedia reached the milestone of 250,000 articles on
    Monday, March 17. The collective of editors on the project issued a press release
    on that, according to which the jubilee article was determined to be “Гендерная
    лингвистика” (Gender linguistics) created by ''ru-wiki'' user Koryakov Yuri.
    ...
  </text>
</page>
```

Fig. 4. Example of XML document „50,000 Articles in Russian Wikipedia“ in SCTM ru corpus

only that part of the reference that corresponds to the sentence grammar.

News shall be accompanied with references to a documentary source. They are usually divided into four types: other articles of Wikinews, external links to online-sources, citations of printed media and websites with reference or related information. For the section of the article “Sources” they use wiki-markup == Sources == (see example in fig. 4). For purposes of topic modelling, links to sources are of no big value, therefore the decision was made to exclude them from the SCTM-ru corpus.

The important element of Wikinews markup and important data for constructing topic models is information on categories, to which the article is related. Categories of the article are defined by its author.

Software was developed in C# language, the development environment is Visual Studio Express 2013, for preliminary processing of texts. Regular expressions were used to search within the export file of Wikinews. The software is multi-modular, each module performs one certain operation. The software receives the initial XML-file as an input, specially prepared regular expressions sequentially search through the file looking for a match by the template, and an XML-file is created with changes made within one iteration at the output. To preserve integrity of initial data, each search through the initial XML-file makes only a few changes that are thoroughly checked by the system administrator, afterwards the software is started with another processing module.

Multi-modular software was developed to calculate statistics of the SCTM-ru corpus. The document count module analyses the XML-tree of the corpus, retrieves unique IDs for each document and counts their total. The author count module retrieves a list of unique IDs of Wikinews article authors and counts their total. The category count module retrieves unique categories from the XML-tree of the corpus and counts their total. The module for processing of the dates of the events described in articles analyses the XML-tree of the corpus, retrieves information on the date of the event of each document, counts unique values, finds the earliest and latest dates of the document.

To count vocabulary of the SCTM-ru corpus, a module was developed using regular expressions and MyStem software. The module takes the text from specified elements of the XML-tree (title, text), regular expressions from the text retrieve all sequences of Russian alphabet letters. In the process of word count the sequence of Russian alphabet letters separated from other letters with something other than letters (punctuation marks, blanks) is a word. MyStem software was used to determine the word lemmas. MyStem software performs morphological analysis of the text in Russian. Hypotheses are generated for the words that are absent in the dictionary [21].

SCTM-ru corpus markup

As SCTM-ru corpus storage format, XML (eXtensible Markup Language) was chosen, as one of most convenient formats for use in software environment and conversion of data into other formats. XML features make it possible to save the text of the initial Wikinews article and highlight additional parameters of the document.

XML-file of the corpus (SCTM-ru) consists of the following elements:

- <page> – group of document elements;
- <title> – document title;
- <id> – unique document ID;
- <userid> – unique author ID;
- <category> – document category;
- <date> – date of document events;
- <text> – document text;

Example of one document markup in SCTM-ru corpus is shown in fig. 4.

The document title (title) is separated from the document text, since title words may be given higher priority in construction of a topic model.

The unique article author ID (userid) is a parameter necessary in author-topic models. The Author-Topic over Time model [22] is an extension to LDA, where distribution of authors, topics and documents in time is evaluated in process of model construction.

Document categories (category) are categories specified by the article author. For instance, in fig. 4 in the article „50,000 Articles in Russian Wikipedia“ the „Russian Wikipedia“ category is specified. Information on the category is important for topic modelling, therefore saved in the SCTM-ru corpus, see fig. 4.

Availability of information on documents belonging to categories will make it possible to automatically check accuracy, completeness, exactness of tested topic modelling algorithms. Information on the document categories may be used in Labeled LDA models described in [23].

The date of the events described in the article (date) is used to build temporal topic models. Example of the model using the date under the title „Topic over Time – TOT“ is shown in the paper [24]. When a temporal model is constructed, apart from standard distributions of words among topics and topics by documents, one shall assess distribution of each topic over time, which makes it possible to track and display dynamics of topics variation over time.

The document text (text) corresponds to the initial article text. We purposefully leave the initial text without any change, without its conversion into a model of a „bag of words“, and without linguistic processing, to make it possible to study unique features of the Russian language. Information on the sequence of words in the document text is used in the models that consider mutual occurrence of words. For instance, the model titled Hidden Topic Markov’s Model – HTMM described in the paper [25] is based on suggestions that words in the sentence structure, as well as sentences themselves are related to one common topic, and the topics of words in the document produce a Markov's chain. As a result of work the HTMM reduces ambiguity of words, widens topic understanding.

Conclusion

As a result of the work done, a special Russian corpus of texts was prepared (SCTM-ru), which is suitable to test various algorithms of probabilistic topic modelling. The objectives set for the work were achieved: SCTM-ru corpus contains original texts of documents in Russian, information on date of events described in the document, information about author and categories, to which the document is related, is available for download and use on devices with no access to Internet.

The source of corpus data is the international news website “Russian Wikinews”. The SCTM-ru corpus contains 7 K documents, 185 authors, almost 12 K unique categories. Events described in the documents are distributed among more than 2 K unique dates, from November 2005 to June 2014. The SCTM-ru corpus contains 2.4 M tokens that consist of Russian letters only. The corpus vocabulary includes 150.6 K unique wordforms, 59 K unique lemmas.

The volume of the developed corpus gives ground to suggest its representativeness for various tasks of automatic processing of natural language texts. As noted in the paper [26], “It is not reasonable to wait until someone balances the corpus scientifically before using it, and it would not be prudent to assess results of corpus analysis as “not well-founded” or “irrelevant” just because one cannot prove that the used corpus is “balanced”. Variety of events described in the SCTM-ru corpus and large team of article authors (21 K members) justify the suggestion on its balance. One may be convinced of corpus balance after analysis of internal features and construction of topic models.

The suggested technology of text corpus development for objectives of topic modelling makes it possible to expand the SCTM-ru corpus due to new articles. Similarly, language corpora may be established in any language from 33 languages presented in Wikinews. Within the proposed format collections and corpora may be created from various sources of data, at the same time only information required for topic modelling algorithms shall be kept.

Then on the basis of the developed corpus the features of existing variations of topic modelling algorithms will be studied, new algorithms will be developed, which take into account linguistic features of the Russian language. The SCTM-ru corpus is distributed through an open license and is available for download at <www.cims.ru>.

References

1. Papadimitriou Ch. H., Raghavan P., Hisao Tamaki, Vempala S. Latent semantic indexing: A probabilistic analysis. – 1998.

2. Hoffman Th. Probabilistic Latent Semantic Indexing. *Proc. 22 Annual Int. SIGIR Conf. Res. Dev. Inform. Retrieval*, 1999.

3. Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 2003.

4. Daud A., Li J., Zhou L., Muhammad F. Knowledge discovery through directed probabilistic topic models: a survey. *Proc. Front. Comput. Sci. Chin.*, 2010, pp. 280-301.

5. Russian National Corpus. Available at: www.ruscorpora.ru (accessed 12.01.2015). (In Russ.)

6. Zakharov V. P. International standards in corpora linguistics. *Struct. Appl. Ling. [Strukturnaya i prikldnaya lingvistika]*, 2012, no. 9, pp. 201-221. (In Russ.)

7. Granovsky D. V., Bocharov V. V., Bichineva S. V. Opencorpora: how it work and perspectives [Otkrytyy korpus: printsipy raboty i perspektivy]. *Computer linguistics and development of semantic search on Internet: Proc. 13th All Russian integrated conf. “Internet and Modern Society” [Kompyuternaya lingvistika i razvitie semanticheskogo poiska v internete: Trudy nauchnogo seminara XIII vserossiyskoy obedinennoy konferencii “Internet i sovremennoe obschestvo]*. St. Petersburg, Oct. 19-22, 2010 / ed. V. S. Rubashkin. St. Petersburg, 2010. 94 p. (In Russ.)

8. OpenCorpora. Available at: opencorpora.org (accessed 15.01.2015). (In Russ.)

9. Small corpus of Associated Press. Available at: www.cs.princeton.edu/~blei/lda-c (accessed 6.01.2015).

10. The New York Times Annotated Corpus. Available at: catalog.ldc.upenn.edu/LDC2008T19 (accessed 14.01.2015).

11. The 20 Newsgroups data set. Available at: qwone.com/~jason/20Newsgroups (accessed 24.01.2015).

12. Reuters Corpora. Available at: trec.nist.gov/data/reuters/reuters.html (accessed 24.01.2015).

13. Reuters-21578 Text Categorization Collection Data Set. Available at: archive.ics.uci.edu/ml/datasets/Reuters21578-+Text+Categorization+Collection (accessed 24.01.2015).

14. 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. *Russian Language: its Histori cal Destiny and Present State: Int. Congress of Rus. Language Res. Philological Faculty of the Lomonosov Moscow State Univ. (MSU) [Russkiy yazyk: istoricheskie sudby i sovremennost: Me zhdunarodnyy congress rusistov issledovateley. Moskva, filo logicheskiy f t MGU im. M. V. Lomonosova]* 13-16 March 2001. Moscow: Moscow State Univ. Press, 2001. P. 398. (In Russ.)

15. The computer corpus of Russian newspapers of the XX century end. Available at: www.philol.msu.ru/~lex/corpus/corp descr.html (accessed 24.01.2015). (In Russ.)

16. Ventsov A. V., Grudeva E. V. About Corpus of Standard Written Russian (narusco.ru). *Rus. Ling.*, 2009, Vol. 33, no. 2, pp. 195-209. (In Russ.)

17. Corpus of Standard Written Russian. Available at: www.narusco.ru (accessed 24.01.2015). (In Russ.)

18. HANCO Corpus. Available at: www.helsinki.fi/venaja/russian/e-material/hanco/index.htm (accessed 24.01.2015).

19. Krizhanovsky A. A., Smirnov A. V. An approach to automated construction of a general-purpose lexical ontology based on Wiktionary.*J. Comput. Syst. Sci. Int.*, 2013, Vol. 52, no. 2, pp. 215-225.

20. 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. *SPIIRAS Proc*. [*Trudy SPIIRAN*], 2012, Is. 23, pp. 231-253. (In Russ.)

21. 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. Berlin, Germany; Springer, 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. *Empirical Methods in Nat. Lang. Proc.*, 2009. Pp. 248-256.

24. Xuerui Wang, McCallum A. Topics over Time: A Non-Markov ContinuousTime Model of Topical Trends. *Proc. 12th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Min ing*, Philadelphia, USA, Aug. 20-23, 2006.

25. Gruber A., Rosen-Zvi M., Weiss Ya. Hidden Topic Markov Models. *Proc. Artificial Intel. Statistics (AISTATS)*, San Juan, Puerto Rico, USA, March 21-24, 2007.

26. Zakharov V. P., Azarova I. V. Special text corpora parametrization. *Structural and Applied Linguistics: Interuniv. collec tion* [*Strukturnaya i prikladnaya lingvistika: mezhvuzovskiy sbornik*]. Vol. 9. St. Petersburg; St. Petersburg State Univ., 2012. Pp. 176-184. (In Russ.)

Корпус текстов русского языка для тестирования алгоритмов тематического моделирования

Карпович С. Н.
АО «Олимп»
Москва, Россия
cims@yandex.ru

Аннотация. **Предложен специальный корпус текстов SCTM-ru для тестирования алгоритмов тематического моделирования. 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 тексты документов на естественном языке, а также востребованную в алгоритмах тематического моделирования информацию. Проведен сравнительный анализ корпусов на русском и иностранных языках, выявлено несоответствие характеристик существующих корпусов обозначенным требованиям.**

Ключевые слова: **текстовый корпус, тематическая модель, обработка естественного языка, русский язык.**

Литература

1. Papadimitriou Ch. H., Raghavan P., Hisao Tamaki, Vempala S. Latent semantic indexing: A probabilistic analysis. – 1998.

2. Hoffman Th. Probabilistic Latent Semantic Indexing // Proc. 22 Annual Int. SIGIR Conf. Res. Dev. Inform. Retrieval, 1999.

3. Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet Allocation // J. Mach. Learn. Res. 2003.

4. Daud A., Li J., Zhou L., Muhammad F. Knowledge discovery through directed probabilistic topic models: a survey // Proc. Front. Comput. Sci. Chin. 2010. P. 280-301.

5. Национальный корпус 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. Грановский Д. В., Бочаров В. В., Бичинева С. В. Открытый корпус: принципы работы и перспективы // Компьютерная лингвистика и развитие семантического поиска в Интернете: тр. науч. семинара XIII Всерос. Объединен. конф. «Интернет и современное общество». Санкт-Петербург, 19-22 окт. 2010 г. /под ред. В. Ш. Рубашкина. – СПб., 2010. 94 с.

8. Открытый корпус. URL: opencorpora.org (дата обращения 10.01.2015).

9. Small corpus of Associated Press. URL: www.cs.princeton.edu/~blei/lda-c (дата обращения 6.01.2015).

10. The New York Times Annotated Corpus. URL: catalog.ldc.upenn.edu/LDC2008T19 (дата обращения 14.01.2015).

11. The 20 Newsgroups data set. URL: qwone.com/~jason/20Newsgroups (дата обращения 24.01.2015).

12. Reuters Corpora. URL: trec.nist.gov/data/reuters/reuters.html (дата обращения 24.01.2015).

13. Reuters-21578 Text Categorization Collection Data Set. URL: archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection (дата обращения 24.01.2015).

14. 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: Междунар. конгресс русистов-исследователей. Москва, филологический ф-т МГУ им. М. В. Ломоносова 13-16 марта 2001 г. Труды и материалы. – М.: Изд-во Москов. ун-та, 2001. С. 398.

15. Компьютерный корпус текстов русских газет конца XX века. URL: www.philol.msu.ru/~lex/corpus/corp descr.html (дата обращения 24.01.2015)

16. Венцов А. В., Грудева Е. В. О корпусе русского литературного языка (narusco.ru) // Рус. лингвистика. 2009. Т. 33, № 2. С. 195-209.

17. Корпус русского литературного языка. URL: www.narusco.ru (дата обращения 24.01.2015).

18. Хельсинкский аннотированный корпус русских текстов ХАНКО. URL: www.helsinki.fi/venaja/russian/e-material/hanco/index.htm (дата обращения 24.01.2015).

19. Krizhanovsky A. A., Smirnov A. V. An approach to automated construction of a general-purpose lexical ontology based on Wiktionary // J. Comput. Syst. Sci. Int. 2013. Vol. 52, № 2. P. 215-225.

20. 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 // Тр. СПИИРАН. 2012. Вып. 23. С. 231-253.

21. Программа 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. – Berlin (Germany): Springer, 2014. P. 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 // Empirical Methods Nat. Lang. Proc. 2009. P. 248-256.

24. Xuerui Wang, McCallum A. Topics over Time: A Non-Markov ContinuousTime Model of Topical Trends // Proc. 12th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, Philadelphia, USA, Aug. 20-23, 2006.

25. Gruber A., Rosen-Zvi M., Weiss Ya. Hidden Topic Markov Models // Proc. Artificial Intel. Statistics (AISTATS), San Juan, Puerto Rico, USA, March 21-24, 2007.

26. Захаров В. П., Азарова И. В. Параметризация специальных корпусов текстов // Структурная и прикладная лингвистика: межвуз. сб. Вып. 9. – СПб.: СПбГУ, 2012. С. 176-184.