Rubric Extraction for Popular Science Articles in the Russian Language

Maria Khokhlova
and Natalia Safonova

St Petersburg State University, Universitetskaya emb. 7-9-11,
199034 St Petersburg, Russia
m.khokhlova@spbu.ru, st095325@student.spbu.ru

Abstract. The paper deals with topic analysis of a corpus compiled from popular science articles in the Russian language. The paper describes the procedures of corpus construction, data preprocessing, and model training on its basis. The purpose of our study is to compare the effectiveness of different methods in identifying topic words and specifying rubrics for the texts. We dwell on topic modeling using Latent Dirichlet Allocation model. The corpus was built on texts from the journal "Nauka i Zhizn" ("Science and Life"), containing articles of the news section from 2015 to 2024. We analyzed the topic diversity of the texts in the rubric and compared the topic words extracted automatically with the tags that were manually attributed by the authors of the articles. The results show that the algorithm can be used to analyze the content of texts as well as enable a more effective information retrieval.

Keywords: Natural language processing, Machine learning, Topic modeling, Latent Dirichlet Allocation

1 Introduction

Topic modeling is an effective tool for processing textual data, which allows combining documents that are close in their content. The method is widely used in as computational linguistics [1], sociology [2], bioinformatics [3] and other fields. Its results can be used in information retrieval, text structure analysis, automatic information extraction, as well as facilitates the analysis of large amounts of data and contributes to the improvement of models related to artificial intelligence.

In the study we evaluate the effectiveness of the probabilistic topic model LDA [4] applied on popular science texts in Russian. The results allowed us to analyze the diversity of the collected corpus, and also to compare the automatically identified words (and hence assigned topics) with the hashtags manually attributed by the authors of the articles.

The paper has the following structure. The Introduction explains the motivation of the study. Section 2 gives an overview of the approaches. The next

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2024, pp. 109–118, 2024.

section describes the design of the corpus compiled for the experiment. Section 4 discusses the results of the experiment. The last section concludes the paper and proposes plans for future work.

2 Topic modelling approaches

One of the crucial tasks in analyzing large amounts of data is to discover some common characteristics among the units within the collection. When it concerns textual data analysis, that task can often involve identifying the range of topics and concepts that appear in a document. It is not difficult for a human to determine what the text is about, but special procedures are required for automatic definition of topics by automatic systems.

Studies devoted to automatic natural language processing have demonstrated that topic modeling can enhance the performance in sentiment analysis [5], as well as automatic abstracting and summary generation [6]. The results of topic modeling techniques have potential for use in the analysis of large collections of documents, since they allow the extraction of the main information relevant for text comprehension. Originally, topic models were offered only as a tool for information extraction and document annotation [7], but at present, the possibilities of their use have expanded considerably. In addition to their wide range of applications in different scientific fields, topic modeling has a significant role for computer vision [8] and recommendation systems.

A topic model is a type of a statistical model, a useful tool for discovering the hidden semantic structure in a set of documents, namely a set of topic words [9]. The building of a topic model is based on the assumption that a text is a randomly selected collection of independent words ("bag of words"), which is produced by certain topics. In this respect, topic modeling refers to the restoration of the probability distributions of all the topics in the text. The generated topic model calculates the degree to which a document belongs to each topic, and also measures the level of accuracy of specific terms in representing a given topic. The model is based on fuzzy biclustering meaning that words and documents are simultaneously clustered into the same "topic clusters". Fuzzy clustering assumes that one word can be assigned with different probabilities to several clusters and, alternatively, one document can be assigned to several clusters with different probabilities [10]. The number of extracted topics can either be chosen by the researcher using the empirical data, or can be calculated by comparing the models automatically. To select the best model, calculating a coherence measure reflecting the level of semantic similarity between words in the same topic can be helpful. Perplexity, which shows how well the model predicts new data, is also an important characteristic for evaluating model performance.

As mentioned above, topic models have been used as information retrieval tools and for automatic indexing of documents in data mining. The models allowed solving the problem of inaccurate document retrieval. Topic models first appeared in the 1990s, as an increase in the processing power of computers

and the transition to machine learning enabled the use of statistical approaches for natural language processing. Among the first topic models introduced, one well-known model is LSI (Latent Semantic Indexing) [11]. Another model, pLSA (Probabilistic Latent Semantic Analysis) was proposed in 1999 [12]. Both models were based on the ideas of distributive semantics, which determines the degree of semantic proximity between text units by designing a vector space. The basic principle of these models is to map documents and terms into a representation in a latent semantic space.

Active development of topic models took place in the 2000s, when new efficient algorithms such as LDA (Latent Dirichlet Allocation) and NMF (Non-Negative Matrix Factorisation) were introduced by groups of researchers [13]. LDA is recognised as the most advanced and efficient for processing large text collections [14].

The models mentioned above can be classified into algebraic and probabilistic models. Recently, models that combine a probabilistic approach with the use of a distributed vector model have emerged. These include LDA2Vec, Top2Vec, and BERTtopic. Naturally, such models have an advantage over algorithms that represent the corpus as a bag-of-words. They reduce the losses associated with this representation, since the quality of semantic text structure representation they provide is higher [15].

Mainly, studies using topic models are carried out on the material of scientific texts, as well as on the texts from social media. For instance, algorithms are used to investigate the evolution of topic structure of journals [16] or to analyze the development of a particular topic within a single magazine [17]. Thematic structure studies are also conducted on the basis of fiction and poetic texts [18]. Another study has demonstrated that, in combination with other methods, the topics that are extracted using LDA can be used to predict the genre or subgenre of a work of literature [19].

3 Corpus construction

As the material for our study we used news articles published on the website "Nauka i Zhizn" ("Science and Life"). These texts are characterized by thematic diversity, while some of the materials are tagged with topic labels (which can be regarded as manually assigned rubrics). We aimed to examine the thematic range of articles, as well as to organize thematically similar texts into an appropriate number of groups. It is worth noting that the journal's website offers a search system, but this system does not provide an opportunity to select materials on a particular theme or field of science, and this fact additionally emphasizes the relevance and practical significance of our work. The search is supposed to be carried out by keywords, which the user has to define on their own, and this may entail certain difficulties. The result is a list of materials sorted by their relevance: the articles in which the key expression occurs most frequently are regarded as more relevant.

However, keywords, marking the topic of the article, do not always have to be the most frequent, and in many cases the user may not get the most relevant article on the output page. This is due to the fact that news items from October 2018 only, that is only 32% of all publications in the rubric, are tagged with topic tags. We arranged the authors' tags in a data frame in order to compare them with the automatically extracted topic words from the model after the results were collected.

The corpus of popular science articles was constructed automatically with the help of the BeautifulSoup library [20]. It included 4,205 materials in Russian in the period 2015-2024 from the "News" section. The corpus had to be preprocessed in the course of the work. This process included tokenization, lemmatization, as well as removal of stop words, punctuation marks and other non-textual symbols. To perform these tasks, we used NLTK library [21] and PyMorphy3 [22]. The corpus had a total of 1,078,538 word instances after processing. The resulting data was manually verified and errors in the lemmatization were corrected where possible. The preprocessing procedure was necessary because LDA works following the bag-of-words principle, which suggests that the model does not take into account grammatical and syntactic criteria, but rather relies on the frequency of the word occurrence in the document.

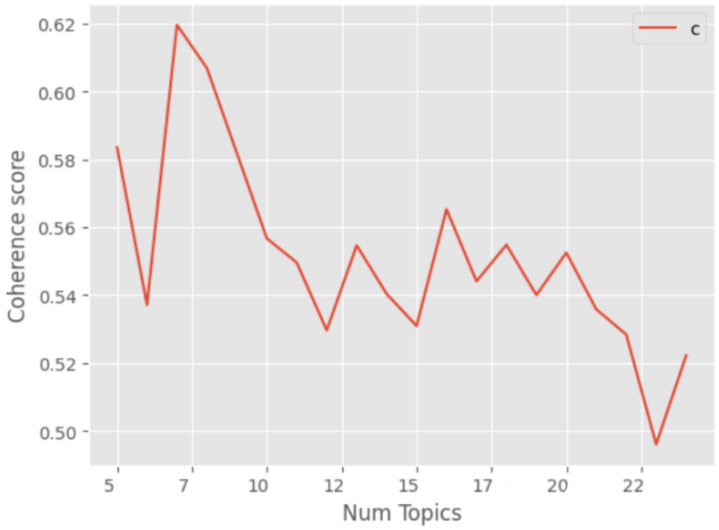
Furthermore, the corpus had to be additionally filtered in the course of the work, since some words with high frequency were assessed as irrelevant for the topic in the model we first obtained. For example, the terms *nauka* 'science' (with an absolute frequency above 10,000 in the corpus), *statya* 'article' (8,000), and *issledovatel'* 'researcher' (above 12,000) all appeared in the initial results. Such words cannot provide a meaningful representation of the topic because they are common to most texts in the corpus. We identified the most frequent words in the model dictionary that are not of interest in the topic model and excluded them from the corpus (e.g., *nauchny* 'scientific', *issledovaniye* 'research, study', *statya* 'article, paper', *universitet* 'university', *experiment* 'experiment', *resultat* 'result').

4 Results of the experiment

To train the LDA model, we used the Gensim machine learning library [23]. The number of topics was selected on the coherence measure (Figure 1.). The coherence index is the highest and reaches 0.62 specifically with 7 topics. For each selected topic a name, or label, was assigned manually. The perplexity level of the model has reached -8.57, which is a solid indicator of successful performance.

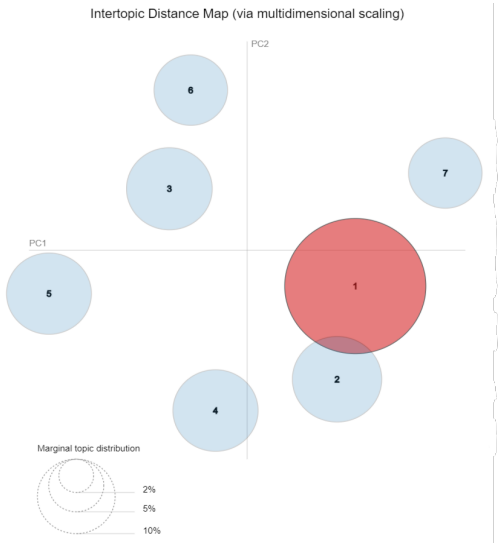
Table 1 shows the results of the LDA topic modeling. For the convenience of the presentation, we have chosen the top 15 topical words out of 30 extracted. In the process of topic labeling, we aimed to reflect the main idea that unites the words that constitute the topic. If a variety of topics were distinguished in a single theme, the name was assigned according to the main idea and in accordance with the content of the documents most typical for it. The presented

Fig. 1: Coherence



lists are vast and hence they can suggest other interpretation. Some labels reflect the topics more accurately, which is dependent on the degree of semantic homogeneity of the words within the topic. Figure 2 shows a visualization of the distances between the topics that are calculated in the system using multidimensional scaling.

Fig. 2: Coherence



We should note that the obtained topics are rather broad topics-directions, and more specific topics can be identified within them. For example, topic_1 which was determined as genetic engineering includes topic words from articles

Table 1: Result of topic modeling with LDA.

#	Label	Topic words (Russian)	Translation
Topic_1	Genetic Engineering	*kletka*, *gen*, *belok*, *dnk* *bolezn'*, *bakteriya*, *immunnyy*, *mysh*, *virus*, *molekula*, *tkan'*, *sistema*, *mutaciya*, *kletochnyy*	'cell', 'gene', 'protein', 'dna', 'disease', 'bacterium', 'immune', 'mouse', 'virus', 'molecule', 'tissue', 'system' 'mutation', 'cellular'
Topic_2	Language Psychology	*socialnyy*, *rebyonok*, *spat*, *vyigrysh*, *igra*, *yazyk*, *uchastnik*, *povedeniye*, *vliyat*, *rech*, *svyaz*, *kontekst*, *kotik*	'social', 'child', 'to sleep', 'win', 'game', 'language', 'participant', 'behaviour', 'to influence', 'speech', 'connection', 'context', 'kitty'
Topic_3	Astrophysics	*zvezda*, *sistema*, *chyornyy*, *planeta*, *temperatura*, *kosmicheskiy*, *atmosfera*, *gaz*, *obyekt*, *massa*, *dyra*, *solnechnyy*	'star', 'system', 'black', 'planet', 'temperature', 'cosmic', 'atmosphere', 'gas', 'object', 'mass', 'hole', 'solar'
Topic_4	Evolution	*derevo*, *rastenie*, *zhivotnoye*, *samec*, *samka*, *ryba*, *ptica*, *zver'*, *dinozavr*, *rasti*, *bolshiy*, *gruppa*, *telo*, *pochva*, *morskoy*	'tree', 'plant', 'animal', 'male', 'female', 'fish', 'bird', 'beast', 'dinosaur', 'to grow', 'larger', 'group', 'body', 'soil', 'marine'
Topic_5	Archaeology	*zemletryasenie*, *drevnij*, *region*, *kost'*, *kolco*, *kamen*, *nahodka*, *tysyacha*, *arheolog*, *naslat*, *godichnyj*, *peshchera*, *sovremennyj*	'earthquake', 'ancient', 'region', 'bone', 'ring', 'stone', 'find', 'thousand', 'archaeologist', 'to lay', 'year', 'cave', 'modern'
Topic_6	Molecular Biology	*ispolzovat*, *molekula*, *himicheskiy*, *veshchestvo*, *material*, *kvantovyj*, *svojstvo*, *pomoshch*, *metod*, *struktura*	'use', 'molecule', 'chemical', 'substance', 'material', 'quantum', 'property', 'help', 'method', 'structure'
Topic_7	Neurobiology	*mozg*, *neyron*, *signal*, *mysh*, *aktivnost'*, *son*, *sistema*, *dofamin*, *nervnyy*, *ritm*, *nejronnyy*, *pamyat*, *glimfaticheskiy*	'brain', 'neuron', 'signal', 'mouse', 'activity', 'sleep', 'system', 'dopamine', 'nerve', 'rhythm', 'neural', 'memory', 'glymphatic'

tagged as “medicine”, “molecular biology”, “methodology of science”, and “cognitive psychology”. This indicates that although the texts can be grouped into one large domain, they are not qualitatively homogeneous. For example, within one topic, at least 10 subtopics can be identified, which suggests a great thematic diversity of news materials.

The same applies, for example, to topic_5, the topic words of which were selected from texts in the fields of anthropology, human evolution, and archaeology. A particular interest is topic_2, which initially raised difficulties in assigning a marker. In this case topic words that represent the names of subfields within one broad topic were assigned to one topic. This topic includes texts about language acquisition by children, learning a foreign language, the brain and language, and also how pets react to speech signals. From these latter texts, the terms “kitty” and “dog” (from the top 30) were among the extracted words. Therefore, the model captures the most general topics that could serve as sections on the website and would also contain some subsections.

The distance visualization also reveals that the model has clustered the texts into seven distinct groups, which, with the exception of the topics 1 and 2 which are similar in some respects, do not overlap and are at a remote distance from each other.

It is also necessary to dwell on the comparison of the author’s tags and the results obtained. As an example, we present a few article titles and their tags from the collected dataset (Table 2).

We can observe that while hashtags carry valuable information about the thematic content of the article, in some cases they are assigned rather arbitrarily (both broadly and narrowly defined topics are observed). For this reason, it is difficult to make an automatic comparison between them and the topics identified in the course of our work. The tags that could most fully and generally relate the text to a particular field of scientific knowledge are not always attributed by the authors (Table 2). The text could have been attributed to the field of biology, but the authors do not assign a more generalizing tag.

Also, in spite of our careful attention to the preprocessing stage, some errors in the output could not be avoided. For example, the top 30 in topic_4 contains the term *nibyt*, which is not really a word, but a negation and a verb (‘notbe’). Similarly, the comparative degree of an adjective, which should have been converted to the initial form bolshiy (‘bigger’), appeared in topic_4. These observations show that further error checking of the corpus is necessary to obtain better results in the future.

5 Conclusion

In the paper, we considered an experiment of clustering a text collection of popular science news articles by rubrics. A probabilistic topic-based LDA model was used, which is a tool that discovers the distribution of topics across documents and the distribution of topic words across topics. The paper evaluated the ef-

Table 2: Result of topic modeling with LDA.

#	Title	Tags
1	*Samki lyubyat umnyh samcov* 'Females like smart males'	*#povedenie zhivotnyh*, *#brachnoye povedeniye*, *#intellekt zhivotnyh*, *#evolyuciya* 'animal behaviour', 'mating behaviour', 'animal intelligence', 'evolution'
2	*Muzhchiny i zhenshchiny pomnyat bol' po-raznomu* 'Men and women remember pain differently'	*#nejrobiologiya*, *#gormony*, #sensornye sistemy 'neurobiology', 'hormones', 'sensory systems'
3	*Ryby ne lyubyat staryh znakomyh* 'Pisces don't like old acquaintances'	*#povedenie zhivotnyh*, *#ryby* 'animal behaviour' 'fish'
4	*Nejrony mozga rassmotreli po sloyam* 'The neurons of the brain have been examined layer by layer'	*#nejrobiologiya*, *#nejrony*, *#mozg*, *#kora mozga*, *#mikroskopiya* 'neurobiology', 'neurons', 'brain' 'brain cortex', 'microscopy'

fectiveness of the model and the obtained results were compared with the tags manually attributed by the authors of the texts.

The experiment demonstrated the high efficiency of the LDA statistical model applied to the corpus of popular science articles in Russian. In future, the preprocessing stage will involve additional corpus filtering at in order to eliminate possible lemmatization errors. The corpus can be supplied with older materials to study the dynamics of the journal's thematic diversity.

Acknowledgements. The work by Maria Khokhlova in the presented research was supported by the Russian Science Foundation, project No. 24-28-00937, `https://rscf.ru/en/project/24-28-00937/`.

References

1. Kirina, M.A.: Comparison of thematic models based on LDA, STM and NMF for qualitative analysis of Russian fiction prose of small form. NSU Vestnik, Linguistics and Intercultural Communication 2(20), 93–109 (2022)

2. McFarland, D.A., Ramage, D., Chuang, J., Heer, J., Manning, C.D., Jurafsky, D.: Differentiating language usage through topic models. Poetics 41, 607–625 (2013)
3. Liu, L., Tang, L., Dong, W., Yao, S., Zhou, W.: An Overview of Topic Modeling and Its Current Applications in Bioinformatics. SpringerPlus, 5, 1608 (2016)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. The Journal of Machine Learning Research, vol. 3, pp. 993–1022 (2003)
5. Lin, C., He, Y.: Joint sentiment/topic model for sentiment analysis. In: Proceedings of the 18th ACM conference on Information and knowledge management, pp. 375–384 (2009)
6. Gong, Y., Liu, X.: Generic text summarization using relevance measure and latent semantic analysis. In: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 19–25 (2001)
7. Deerwester, S., Dumais, S.T., Furnas G.W., Landauer. T. K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science, vol. 41 (6), pp. 391–407 (1990)
8. Fei-Fei, L., Perona, P.: A Bayesian hierarchical model for learning natural scene categories. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 2, pp. 524–531. San Diego, CA, USA (2005)
9. Blei, D.M.: Probabilistic topic models. Communications of the ACM, 55(4), 77–84 (2012)
10. Hoffman, M.D., Blei, D.M., Wang, C., Paisley, J.: Stochastic variational inference. Journal of Machine Learning Research, vol. 14(40), pp. 1303–1347 (2013)
11. Papadimitriou Ch., Tamaki, H., Raghavan, P., Vempala, S.: Latent semantic indexing: a probabilistic analysis. In: Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems (PODS '98), pp. 159–168. Association for Computing Machinery, New York, NY, USA (1998)
12. Hofmann, T.: Probabilistic Latent Semantic Analysis. In: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, pp. 289–296 (1999)
13. Lee, D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature (401), 788–791 (1999)
14. Blei, D.M., Lafferty, J. D.: Topic models. Text mining, pp. 101–124. Chapman and Hall/CR (2009)
15. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics (2019)
16. Odden, T., Marin, A., Rudolph, J. L.: How has Science Education Changed over the last 100 years? An analysis using natural language processing. Science Education (105), 653–680 (2021)
17. Jacobi, C., Van Atteveldt, W., Welbers, K.: Quantitative analysis of large amounts of journalistic texts using topic modelling. In: Rethinking Research Methods in an Age of Digital Journalism, pp. 89–106. Routledge (2018)
18. Rhody, L.M.: Topic Modelling and Figurative Language. Journal of Digital Humanities, 19–35 (2012)
19. Schöch, C.: Topic modeling genre: an exploration of French classical and enlightenment drama. Digital Humanities Quarterly, vol. 11(2) (2017). `https://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html` Last accessed 1 Nov 2024.
20. BeautifulSoup library, `https://www.crummy.com/software/BeautifulSoup`. Last accessed 1 Nov 2024.

21. NLTK library, https://www.nltk.org. Last accessed 1 Nov 2024.
22. PyMorphy3 library, https://pypi.org/project/pymorphy3. Last accessed 1 Nov 2024.
23. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45–50 (2010). https://is.muni.cz/publication/884893/en.