Multilingual topic search and categorization of scientific publications

Eugene V. Tishchenko, Konstantin V. Vorontsov, Polina S. Potapova

This paper considers the problem of cross-lingual search for textual plagiarism. The goal of the paper is to obtain a model that extracts information about the distribution of words in the text regardless of their linguistic origin. In addition, the model must satisfy the size and training time constraints for practical use.

Keywords: *machine learning, topic modeling, multilingual search for duplicates, natural language text processing, multilingual text processing*

1 Introduction

The task of searching and recognizing subject-related documents is important because of the impossibility to check all documents of common subject for plagiarism due to the size of the global network. The task becomes especially difficult as there are many documents that are translations of the original works. According to research by Donald L. McCabe, [1] 36% of American university students have paraphrased or copied information from the World Wide Web without citing the source.

Document vectorization methods for plagiarism detection [2, 3] are mostly limited to a single language. In this case, the problem of creating a uniform system for obtaining vector representations of multilingual document collections arises. To be able to apply the model beyond scientific experiments, technical constraints are set by the server resources, i. e., model size and time complexity of training.

The objects of the research are multilingual topic models, algorithms for searching documents in the text collection by words and documents, methods of vector representation of words and documents of the query and collections used for searching duplicates regardless of language affiliation.

The topic model of a collection of text documents determines which topic each document belongs to and which words or terms form each topic. A probabilistic topic model describes each topic by a discrete probability distribution of words, and each document − by a discrete probability distribution of topics. A topic model converts any text into a vector of topic probabilities.

A multimodal topic model is used to solve this problem. This topic model describes documents that contain metadata in addition to the main text. Metadata refers to the language affiliation of the document, authorship, date of publication, links to other sources, etc. Metadata improves the accuracy of detection of document topics. As modalities are used 100 languages, as well as scientific rubrics of the State Rubricator of Scientific and Technical Information GRNTI [4] and the Universal Decimal Classifier UDC [5]. Vector representation of the text is obtained by using languages as modalities. This representation is independent of the original text language. During text preprocessing, BPE tokenization − iteratively replacing the most occurring character pair with a character that does not occur in the word. This method significantly reduces the size of the original dictionary for practical use of the model.

The goal of the experiment is to develop a topic model for obtaining vector representations of texts. The experiment analyzes the effect of regularization and preprocessing of text data on the quality of the search. Articles from the Wikipedia site, and a sample of scientific articles from the scientific electronic library eLIBRARY.ru are used as training data.

2 Problem Statement for Document Duplicate Search

Let D – a collection of documents, T – a set of document topics. As a candidate for a duplicate document $d \in D$ we denote such an element of the collection $f(d)$ that

$$f(d) = \underset{d' \in D \setminus \{d\}}{\arg\min} \rho(m(d), m(d')).$$

The function ρ is an arbitrary vector space metric. The function $m(d)-$ is a model that constructs a vector representation of the document. The quality of the duplicate search model is measured on a test sample using two document mapping metrics:

1. The average frequency with which a query document ranks in the top 10%. For the document in the test sample it is already known which documents are translations of the original document. We consider the fraction of those documents for which there are translations among the 10% ranked duplicate candidates.

2. The average percentage of documents in the top 10% of translation documents that have the same rubric as the query document. Similar to the previous metric, but considering the percentage of those documents for which documents of the same rubric turn out to be the closest in vector representation.

It is demanded to solve the problem of searching for duplicates, and the quality should exceed 0.9 on the first metric, 0.3 on the second. It is also necessary to limit the size of the model to 100 GB, and model training must be done within no more than 24 hours.

Topic modeling is used as a transformation of text to vector representation. The topic model [6] on a collection of documents constructs a probability distribution $p(w|t)$ of terms w – of words, phrases, and terms in topics $t \in T$. According to the conditional independence hypothesis as well as the total likelihood formula, the distribution of terms w in document d is a probabilistic mixture of distributions of terms in topics $\varphi_{wt} = p(w|t)$ with weights $\theta_{td} = p(t|d)$:

$$p(w|d) = \sum_{t \in T} p(w|t,d)p(t|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \varphi_{wt}\theta_{td}.$$

This expression can be rewritten in matrix form. The matrix of frequencies of terms in the documents

$$F \approx \Phi\Theta.$$

Since the number of topics is usually much larger than the number of documents, we need to find a decomposition whose rank does not exceed $|T|$.

3 Computational experiment

The problem is to determine the optimal parameters of the topic model, the methods of text tokenization, which will meet the requirements on the quality of the model and satisfy the technical constraints. After selecting the optimal parameters, experiments will be conducted to show the contribution of each individual solution to the overall quality of the model.

The test sample is generated in a certain way. The search for a topic related document is performed not on the entire search collection, but on a subsample of documents in which 10% of documents have the same UDC heading as the query document, and the remaining 90% of documents have a different UDC heading.

The analogy metric is used as a metric in vector space. Let A — the language of the original document d, and search among documents d' in language B. Then

$$\rho(d, d') = 1 - cos(d', d - a + b)$$

where a and b — are the average vectors for all documents in the collection belonging to languages A and B, respectively.

3.1 Data description

The articles of the eLibrary [7] scientific digital library and the Wikipedia [8] multilingual online encyclopedia were used to train the topic model. The GRNTI and UDC rubrics were known for the majority of the eLibrary scientific articles. The sample of articles in 100 languages was generated as follows. Twenty-four thousand eLibrary articles in Russian and English were selected. Then the articles were translated into 42 languages. The languages were chosen because of their high representation among scientific texts. The translation was accomplished with the Moses statistical machine translation system [9]. Wikipedia articles were also collected for all languages, including these 42 languages.

The distribution of the training data by the UDC and GRNTI rubrics is presented below. It should be pointed out that the samples are unbalanced by the volume of documents, which leads to problems during training, since the model is supposed to be trained on balanced data. Besides, among the rubrics of the GRNTI there is a rubric «no», which is a set of documents which do not belong to other rubrics. This rubric contains documents related to various topics and is the fourth largest among all the rubrics.

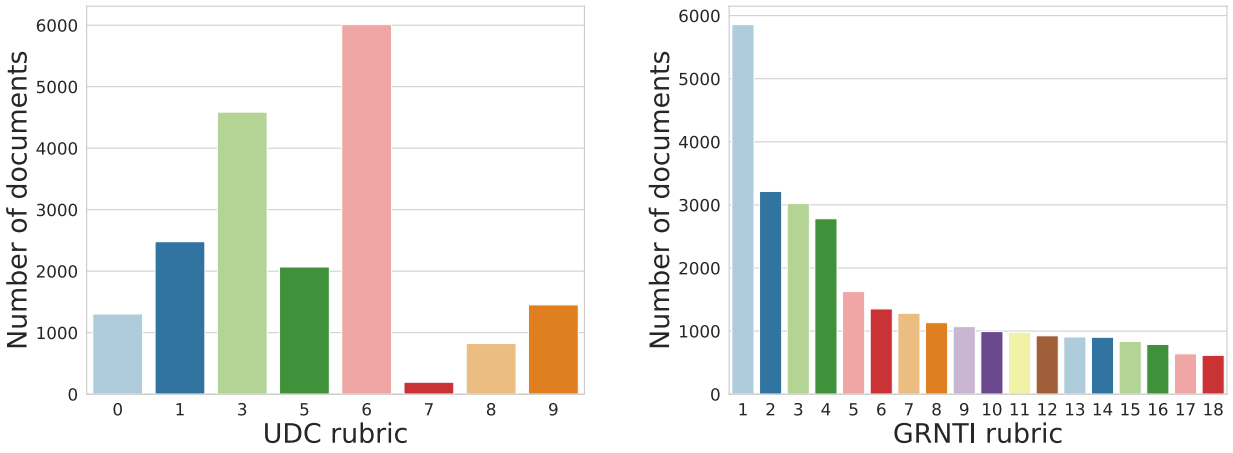

Figure 1 Distribution of data by UDC and GRNTI rubrics

3.2 Base models

A topic model is used as the base model, using 100 languages as modalities, see Table 3. Learning is based on Wikipedia data. 300 themes are used, of which one is a background theme. The background theme contains all common words that are not subject words. These include prepositions, particles, conjunctions, and other non-informative parts of speech. During text preprocessing the BPE tokenization method is used, and the original dictionary volume of 120 thousand tokens is reduced to 2 thousand for each individual language. Training the model takes about 6 hours and the size of the model is 4.6 GB.

The XLM-RoBERTa [10] neural network model is considered as an alternative baseline model. This model is trained on 2.5TB of data containing documents in 100 different languages. Vector text representation is obtained by averaging the embeddings of individual sentences. Obtaining a vector representation of an individual text takes an average of 1.6 seconds, but due to the large size of the document collection, the total time of obtaining embeddings takes 120 hours.

4 Proposed solution

Applying various heuristics to the underlying topic model, as well as searching for optimal parameters, allows us to achieve the required results. Below is a list of the solutions applied to improve the quality of the model:

1. When teaching the model, the optimal number of topics between 10 and 1,000 is selected. The best results were achieved with the selection of 125 individual topics. A significant gain is made by adding the modalities of the UDC and GRNTI rubrics.
2. The training of the topic model is iterative, with a random subsample of documents selected at each iteration. To solve the problem of imbalanced data with respect to rubrics, each iteration of training generates a subsample of documents for which the distribution of GRNTI rubrics turned out to be equal. So, in each subsample of documents, the number of documents with different rubrics is equal.
3. Reducing the vocabulary down to 2 thousand tokens per language allows us to meet the constraints on model size, but this number of tokens is insufficient to describe a single language. Using 11 thousand tokens makes it possible not only to improve the expressiveness of the model, and thus its quality, but also to meet the restrictions on the training time, which is no more than 24 hours.
4. The use of the heading «no» reduces the quality of the model, because it contains documents of different topics. To improve the quality of the model it was decided not to take this heading into account and not to set it as a modality for documents. In addition, it was decided to exclude stop words from the 15 main languages.

The training takes 17 iterations, and Fig. 2 shows how the metrics change depending on the iteration.

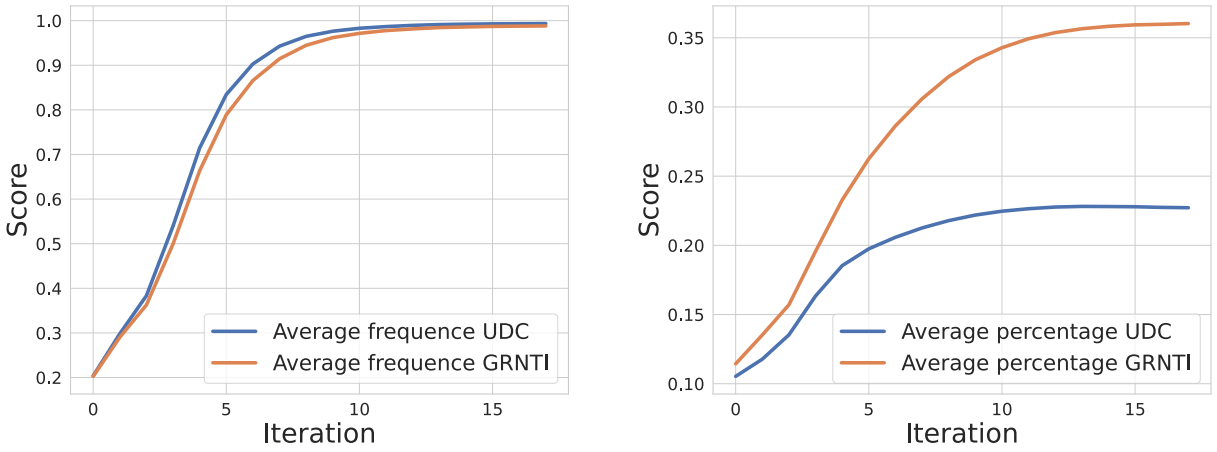

Figure 2 Scores on train

4.1 Experimental results

A comparison of the base solutions with the final solution is shown below.

Table 1 Model comparison

Model name	Average frequency UDC	Average percentage UDC	Average frequency GRNTI	Average percentage GRNTI
Basic topic model	0.558	0.165	0.536	0.220
XLM-RoBERTa	0.835	0.179	0.832	0.288
Final topic model	**0.995**	**0.225**	**0.852**	**0.366**

As can be seen from the table, the quality of the final topic model differs greatly from the base solution, as well as from the XLM-RoBERTa model. By all metrics, the final model showed better quality.

Table 2 shows the quality of the topic model, which was trained without a certain feature. As you can observe, the decision to expand the dictionary to 11 thousand tokens made the greatest contribution to the quality of the model.

Table 2 Feature importance comparison

Model name	Average frequency UDC	Average percentage UDC	Average frequency GRNTI	Average percentage GRNTI
Final topic model	**0.995**	**0.225**	**0.852**	**0.366**
Using 300 themes instead of 125	0.714	0.185	0.663	0.232
Using 11,000 tokens instead of 2,000.	0.539	0.163	0.499	0.195
Generating samples with the same distribution of topics	0.834	0.197	0.789	0.262

5 Conclusion

In this paper we propose a method for obtaining a vector representation of the text using topic modeling, which allows to obtain a vector of probabilities of belonging to individual topics on the document, regardless of their linguistic affiliation.

The underlying topic model as well as the XLM-RoBERTa neural network model were considered. Both models were tested on a collection of documents belonging to 42 different languages.

Various heuristics were proposed, as well as a text preprocessing method to improve the quality of the thematic model, while meeting the limitations on model size and training time for practical use. A comparison of the final thematic model with the baseline model was conducted.

References

[1] Donald L. McCabe Cheating among college and university students: A North American perspective// *International Journal for Educational Integrity*, 2005

[2] Zdenek Ceska, Michal Toman, and Karel Jezek Multilingual Plagiarism Detection// *Artificial Intelligence: Methodology, Systems and Applications*, 2008

[3] Duygu Ataman, Jose G. C. de Souza, Marco Turchi, Matteo Negri Cross-lingual Semantic Similarity Measurement Using Quality Estimation Features and Compositional Bilingual Word Embeddings//

[4] *grnti.ru*

[5] *udcsummary.info*

[6] Vorontsov K. V. Overview of probabilistic thematic models.//

[7] *elibrary.ru*

[8] *wikipedia.org*

[9] *Koehn P. et al.* Moses: Open source toolkit for statistical machine translation //Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions. – 2007. – С. 177-180.

[10] *Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, Veselin Stoyanov* Unsupervised Cross-lingual Representation Learning at Scale.

6 Appendix

The table shows all the languages that were used to train the topic model.

Table 3 Presented languages

1	English	26	Arabic	51	Uzbek	76	Burmese
2	German	27	Persian	52	Interlingua	77	Karakalpakian
3	French	28	Finnish	53	Galician	78	Amharic
4	Russian	29	Serbo-Croatian	54	Malayalam	79	Somali
5	Spanish	30	Serbian	55	Albanian	80	Samoan
6	Italian	31	Slovenian	56	Welsh	81	Gujarati
7	Portuguese	32	Hebrew	57	Bosnian	82	Laotian
8	Chinese	33	Afrikaans	58	Bengali	83	Yiddish
9	Japanese	34	Lithuanian	59	Tajik	84	Kirundi
10	Czech	35	Indonesian	60	Nepali	85	Javanese
11	Polish	36	Greek	61	Swahili	86	Romani
12	Turkish	37	Hindi	62	Turkmen	87	Tatar
13	Danish	38	Thai	63	Moldovan	88	Chechen
14	Kazakh	39	Catalan	64	Esperanto	89	Bashkirian
15	Kyrgyz	40	Malay	65	Gaelic	90	Chuvashian
16	Korean	41	Belorussian	66	Sinhalese	91	Avar
17	Swedish	42	Estonian	67	Mongolian	92	Kabardian-Circassian
18	Hungarian	43	Icelandic	68	Basque Language	93	Ingush
19	Dutch	44	Azeri	69	Malagasy	94	Ossetian
20	Romanian	45	Latvian	70	Tamil	95	Yakutian
21	Norwegian	46	Latin	71	Khmer	96	Karachai-Balkar
22	Bulgarian	47	Georgian	72	Urdu	97	Lezghinian
23	Croatian	48	Macedonian	73	Maori	98	Erzyan
24	Ukrainian	49	Armenian	74	Aleutian	99	Mari