Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2018”

Moscow, May 30—June 2, 2018

IMPROVING TOPIC MODELS WITH SEGMENTAL STRUCTURE OF TEXTS

Skachkov N. A. (nikolaj-skachkov@yandex.ru)
Lomonosov Moscow State University

Vorontsov K. V. (voron@forecsys.ru)
Dorodnicyn Computing Centre of RAS,
Moscow Institute of Physics and Technology, Russia

Probabilistic topic modeling is a powerful tool of text analysis, that reveals topics as distributions over words and then softly assigns documents to the topics. Even though the aggregated distributions can be good with basic models, a sequential topic representation of each document is often unsatisfactory. This work introduces a method that allows to increase the quality of topical representation of each single text using its segmental structure. Our approach is based on Additive Regularization of Topic Models (ARTM), which is a technique for imposing additional criteria into the model. The proposed method efficiently avoids a bag-of-words assumption by considering the topical connections of words that co-occur in a local segment. We assume, that sequential sentences are topically and semantically coherent, while the number of topics in each particular text fragment is low. We apply our model to topic segmentation task and achieve a better quality than the current state-of-the-art TopicTiling algorithm. In further experiments we demonstrate that the proposed technique reveals an interpretable sequential structure of documents, while keeping a number of topics low, i.e. the sparsity of the model increases. Apart from topic segmentation, the constructed topical text embeddings can be used in any other applications, where the analysis of the document structure is desirable.

Keywords: Topic modeling, text segmentation, topic segmentation, topical embeddings, sparse embeddings, EM-algorithm

ИСПОЛЬЗОВАНИЕ СЕГМЕНТНОЙ СТРУКТУРЫ ДОКУМЕНТОВ ДЛЯ ПОСТРОЕНИЯ ТЕМАТИЧЕСКОЙ МОДЕЛИ

Скачков Н. А. (nikolaj-skachkov@yandex.ru)
Московский Государственный Университет
им. М. В. Ломоносова

Воронцов К. В. (voron@forecsys.ru)
Вычислительный центр им. А. А. Дородницына РАН,
Московский Физико-Технический Институт, Россия

1. Introduction

Topic modeling is a rapidly developing branch of statistical text analysis. Topic model uncovers a hidden semantic structure of the text collection and finds a highly compressed representation of each document by a set of its topics. From the statistical point of view, each topic is a set of words or phrases that frequently co-occur in many documents. The topical representation of a document captures the most important information about its semantics and therefore it is useful for many applications including information retrieval, classification, categorization, summarization and segmentation of texts [Vorontsov, 2014].

Despite many advantages, topic models are known to fail modeling the structure of the text inside the documents. Usually all the topics that are presented in the document are evenly distributed along the text. That is strongly connected with the bag of words assumption in the modeling of the texts. This assumption significantly simplifies the theoretical inference that allows to receive an iterative solution known as EM algorithm. But in many tasks, such as analysis of large documents, intra-document search, or dialog systems, it is important to model intra-document topic behavior with a good granularity.

One of the most popular topic models is Latent Dirichlet Allocation (LDA) proposed in [Blei, 2003]. LDA is a two-level Bayesian generative model, which assumes that topic distributions over words and document distributions over topics are generated from prior Dirichlet distributions.

Many authors successfully tried changing LDA generative model in such a way that some assumptions about text structure are incorporated.

For example, in [Balikas, 2016] a model called senLDA was built. In this model, all words in a sentence could have only one and the same topic label. In the experiments, this model converged faster than LDA, and the representation of documents provided by this model successfully complemented LDA representation for a task of document classification.

In [Du, 2013] a more complex LDA-based Topic Segmentation Model (TSM) was proposed. It assumes that documents consist of segments, whose topical subjects are also

present in the document subjects. To model the segments subjects, a Pitman-Yor process is used. It represents each segment as a Chinese restaurant, where customers represent words, dishes represent topics and tables represent monothematic subsets of words.

For all these models, any assumptions on the text structure change the generative model, thus making it hard to design and infer new modifications.

In this work, we offer a new method based on Additive Regularization of Topic Models (ARTM) [Vorontsov, 2014]. Our method allows to reconstruct the segmental structure of the text in the topic model. The estimated segment boundaries are being used to reduce the number of topics within a segment. This is made in an assumption that words within a text fragment share the same small set of topics. The result topic model granulates topics in the segments and increases the model sparsity. This topic structure may be used for automatic intra-document analysis.

Finally the proposed enhancements don't complicate the structure of model training and theoretical inference, but increase topic model quality, sparsity and interpretability. All of that reveals new opportunities for applications of topic models.

To evaluate the segmentation quality an artificially generated corpus is used. It is generated from PostScience collection. We are using artificial documents for evaluation as it was done in many works before [Galley, 2003; Du, 2013; Riedl, 2012]. We use them because the comparison on real texts is a complicated challenge as there are no golden standard segment boundaries provided. We create artificial documents by concatenating full-source documents from the PostScience collection. This method was shown to be more justified comparing to the other ways of artificial documents construction [Riedl, 2012].

The paper is organized as follows. The next section gives an overview of ARTM and section 3 introduces our approach to topic segmentation. Section 4 provides details about parameter evaluation. Finally, section 5 presents our final results on an artificial corpus.

2. Additive regularization of topic models

A topic model describes a collection D by a finite set of topics T. In ARTM [Vorontsov, 2014] and in more basic PLSA model [Hoffman, 1999], the distribution of words in documents is modeled as a mixture of topics:

$$p(w|d) = \sum_{t \in T} p\,(w|t)p(t|d), d \in D, w \in W \tag{1}$$

The model is parametrized stochastic matrices Φ and Θ with the elements:

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

Topic modeling can be also interpreted as a task of approximate matrix factorization $F \approx \Phi\Theta$. The solution of matrix factorization task is non-unique, thus we follow ARTM [Vorontsov, 2012] approach and consider additional criteria to learn better Φ and Θ matrices. Particularly, we maximize the weighted sum of log-likelihood and some additive regularizers R_i:

$$\sum_{d \in D} \sum_{w \in W} n_{dw} \ln \sum_{t \in T} \varphi_{wt}\theta_{td} + \sum \tau_i R_i(\Phi, \Theta) \to \max_{\Phi, \Theta} \tag{2}$$

Regularizers R_i impose additional problem-specific criteria on the model parameters. Regularizer coefficients τ_i balance the importance of regularizers and log-likelihood. If no regularizers are added, the described model simplifies to PLSA.

The stationary point of the problem (2) satisfies the system of equations, that yields expectation-maximization algorithm as the fixed point iteration method. E-step of this algorithm calculates probabilities of word assignments to topics in the context of a document $(t \mid d, w) \equiv p_{tdw}$. M-step uses these probabilities to update matrices Φ and Θ.

3. Using the segmental structure of documents to improve EM-algorithm

According to (1), each document is represented as a bag-of-words. Additive regularizers are normally applied on M-step and they also cannot make any assumptions about the word order.

However, during the E-step, we compute p_{tdw} probabilities for each position in the document sequentially. It means, we can impose additional assumptions on topic distributions for the words that occur in the same part of the text. According to these assumptions, p_{tdw} values can be modified and then used at the M-step of the EM algorithm.

In real texts, authors usually convey their thoughts in a sequential way. That is why we can expect to see only a few topics in any small piece of text. This can be formulated as a sparsity assumption of subjects within a sentence $(t|s)$, where we define the subject of the sentence as an average sum of its word distributions over topics:

$$p_{ts} \equiv p(t|s) = \frac{1}{n_s} \sum n_{sw}\, p_{tdw},$$

where n_s is the length of sentence, and n_{sw} is the number of occurrences of word in sentence.

One can show that the provided sparsity assumption influences p_{tdw} in the following way:

$$p_{\widetilde{tdw}} = p_{tdw}\left(1 - \frac{\tau}{n_{dw}} \sum_{s \in S_d} \frac{n_{sw}}{n_s} \left(\frac{1}{p_{ts}} - \sum_{z \in T} \frac{p_{zdw}}{p_{zs}}\right)\right) \tag{3}$$

where S_d is a set of all sentences in a document.

We omit derivation of this formula due to space limitations, but provide some intuition behind. If p_{ts} for some sentence is close to zero, then $1/p_{ts}$ is big and the resulting sign for this sentence term is negative. It means, that the probability of the corresponding topic in the word will be decreased. On the contrary, if p_{ts} is close to 1, this sentence term may be positive and the probability of the corresponding topic will increase. Shortly speaking, each term of the sum in this formula brings the distributions p_{tdw} closer to the main topics of the sentences where the word occurs.

It is worth being mentioned that the $p_{\widetilde{tdw}}$ may not determine a distribution over topics, because the values can be negative, but this doesn't break the EM algorithm.

The τ parameter in the formula determines the strength of the influence of sentence subjects on word distributions over topics.

Now let us elaborate the idea of sparsity of the subjects of text fragments and use it for topic segmentation task. We assume that any text consists of *segments* that can be represented by a small number of topics. Topics are supposed to stay constant along each segment. Two sequential segments are supposed to have low intersection of topics.

We estimate borders of the segments gradually while the model is being learnt. At the first iteration of EM-algorithm we use sentences as an initial approximation of the segments. Then at each E-step we find the subjects of the segments and merge the sequential segments if they share the same topics. Equation (3) in this method is applied to the segments that have been built so far by the current iteration.

4. Topic model quality evaluation

To evaluate topic model quality, we will take into account two factors: topic model segmentation quality and intra-document sparsity. Segmentation quality shows the ability of the model to restore topic borders and semantic changes in the text. To evaluate this, we will use an artificially generated text corpus. It will provide us with the gold standard for segment boundaries in the documents. To check how the golden standard boundaries overlap with the estimated ones, we will use P_k and WindowDiff measures as it was done in the prior work [Riedl, 2012].

The intra-document sparsity shows the ability of the model to describe semantic segments with the smallest possible number of topics. The sparsity of segment subjects implies the sparsity of the whole document, so we will use θ-matrix average sparsity to evaluate this.

4.1. Building segment boundaries using topic models

For all topic models, we will use a special topic segmenter algorithm to find segment boundaries. This method was applied in TopicTiling algorithm [Riedl, 2012] and showed good results compared to other segmentation algorithms. The idea of this method is to calculate a similarity between left and right windows for each sentence ending. The sentence endings with the lowest values of this similarity are considered to be candidates for segment boundaries. Then some smoothing transformation is applied to the similarity function to obtain a so called *depth score*. The candidates with the depth score exceeding a certain threshold are selected as the final segment boundaries. The depth score can be also interpreted as a probability of the boundary in the corresponding sentence ending.

Our version of the segmenter algorithm differs from the original one, as we use sentence subjects to calculate similarities between the windows of sentences. In the original version, the authors used the topic IDs assigned to the words during the inference.

4.2. Segmentation quality metrics P_k and WindowDiff

P_k measure uses a sliding window with a length of k tokens, which is moved over the text to calculate the segmentation penalties. For each pair of words at a distance of k it is checked whether both words belong to the same segment or to different segments. This is done separately for the golden standard boundaries and the estimated segment

boundaries. If the gold standard and the estimated segments do not match, a penalty of 1 is added. Finally, the error rate is computed by normalizing the penalty by the number of pairs. A value close to 0 denotes a perfect segmentation quality of the estimation.

The value of parameter k is assigned to half of the number of tokens in the document divided by the number of segments, given by the gold standard.

A drawback of the P_k measure is its unawareness of the number of segments between the pair of words. WindowDiff is an enhancement of P_k: the number of segments between the pair of words is counted. Then the number of segments is compared between the gold standard and the estimated segments. If the number of segments are not equal, 1 is added to the penalty, which is again normalized by the number of pairs to get an error rate between 0 and 1. [Riedl, 2012]

4.3. Artificially generated corpus description

We use PostScience corpus as a basis for the generated collection. We apply lemmatization, delete stop words and all documents that contain more than 200 sentences or less than 10 sentences. Then we compose artificial documents by concatenating full source documents. As it was mentioned in [Riedl, 2012], using full documents makes the corpus more realistic compared to the case when only the fragments of documents are concatenated.

To avoid topic repetition in sequential segments, we build a simple topic model on PostScience dataset and use only documents with different topics for sequential segments. Moreover, we use only the documents with the probability of one topic exceeding the threshold of 0.8. All of this allows us to assume that the golden standard segmentation boundaries also appear to be the topical boundaries.

The number of segments in a document varies from 2 to 4. The resulting number of documents in the generated corpus is 700.[1]

4.4. Experiments Setup

In all the provided experiments, BigARTM open-source library [Vorontsov, Frey, 2015] was used for topic model constructing. To find the optimal parameter values, we use the 5-fold cross-validation on the training subset. It includes 500 artificial documents. Here WindowDiff metrics is used for the evaluation.

Let us describe all the parameters that are the subject of our exploration:

- I is the number of iterations in EM-algorithm. This parameter is strongly connected with overfitting and model convergence.
- α is the strength of Theta sparsity regularizer. By tuning this parameter, we investigate whether a simple sparsity decreases the segmentation quality of topic models.
- τ_1 is the τ parameter for equation (3), which is used when segment boundaries match the sentence boundaries.

[1] The artificially generated corpus ArtPostScience is available here: https://yadi.sk/d/fSswtwqV3SbsCD

- τ_2 is the τ parameter for equation (3), which is used when the segment boundaries are being calculated iteratively by merging sequential segments.
- w is the size of window, which is used in segmenter algorithm to calculate final boundaries.

Table 1. Topic model parameter evaluation

Parameter	Optimal value	WindowDiff
I	40	0.253
α	0.2	0.248
τ_1	0.1	0.242
τ_2	11	0.232

The results of the parameters tuning are provided in Table 1. Once the the optimal number of iterations was found, we define the structure of the training process. During the first 5 iterations the topic model works without any regularizers. At the 5th iteration the regularizer of Θ sparsity starts to work. For the last 25 iterations, we apply equation (3) on each E-step of the algorithm. This structure of training is essential, because the ability to use segment subjects requires the convergence of the topic model.

As we can see from Table 1, the optimal value of α parameter is very small. That means that strong influence of Θ sparsity regularizer lowers the ability of the topic model to restore the segment boundaries.

One can also note, that iterative merging of segments gets a high τ_2 coefficient in equation (3), while the strategy with fixed sentence boundaries keeps this coefficient low. That means that making sparse small segments like sentences is undesirable. That can be explained by the fact that small sentences depend more on separate words topics and their subjects are more unstable.

The optimal value for the window parameter was 11. It may be explained by the fact that the shortest gold standard segments in our collection have the size of 10 sentences. Thus, in real collection we would recommend to set this parameter equal to the length of the shortest segment.

5. Main results

All the models with the optimal parameter values were evaluated on the test documents of our artificial dataset. The number of test documents is 200. The train documents were used only to build the topic model.

The final results in segmentation are shown in table 2 and compared against TopicTiling model from [Riedl, 2012]. For TopicTiling baseline model, we reproduce the original estimation of similarities between windows based on topic assignments in LDA inference. In our models, we use sentence subjects to calculate the similarities. *PLSA + Θ sparsity* is the model that uses Θ sparsity regularizer. *PLSA + SentenceSparse* is the model that applies equation (3) using sentence subjects. *PLSA + SegmentSparse* is the model with iterative segment merging. We used estimated parameters for all models provided.

Table 2. Final results in segmentation

Model	WindowDiff metrics	P_k metrics
TopicTiling	0.258	0.145
PLSA + Θ sparsity	0.173	0.100
PLSA + SentenceSparse	0.159	0.099
PLSA + SegmentSparse	**0.155**	**0.095**

One can see that the proposed method with regularization (3) and iterative merging of segments gives the best segmentation quality. Also both TopicTiling and PLSA + Θ-sparse models, that are built in a bag-of-words assumption, show the worst results in segmentation.

To better explore the effective number of topics in the models, we compare the sparsity levels of Θ matrix in Table 3. The SegmentSparse model decreases the average number of topics almost in 3 times. Without implementation of equation (3), such result in sparsity could have only been achieved with a loss in segmentation quality.

Table 3. Results of Θ sparsity for different models

Model name	Ration of non-zeros in Θ
PLSA + Θ sparsity reg.	4%
PLSA + SentenceSparse	1.8%
PLSA + SegmentSparse	**1.5%**

Let us look more closely into how topics are spread along a document. In Fig. 1 we represent dominant topics for each word in the sequential text with a color. The PLSA + Θ-sparse model reveals only two semantic segments in the text and in the second segment all the topics are mixed up. Whereas the SegmentSparse model catches all three segments and makes them topically different. Yellow topic appears in both second and third segments so we don't mark its words in the given text fragments. Also note, that the second model used up 32 topics, while for the first model 9 topics were enough. So in this document the SegmentSparse model outperformed PLSA + Θ-sparse in both sparsity and segment boundaries estimation.

Now we provide the results of the SegmentSparse model application to the documents of the original PostScience collection. We still use the model that was trained on artificial documents. On Fig. 3 we represent an original document, where SegmentSparse model found two different segments. The first topical segment contains historical information when the second segment is more about natural features of the region. So this segmentation of the document is justified.

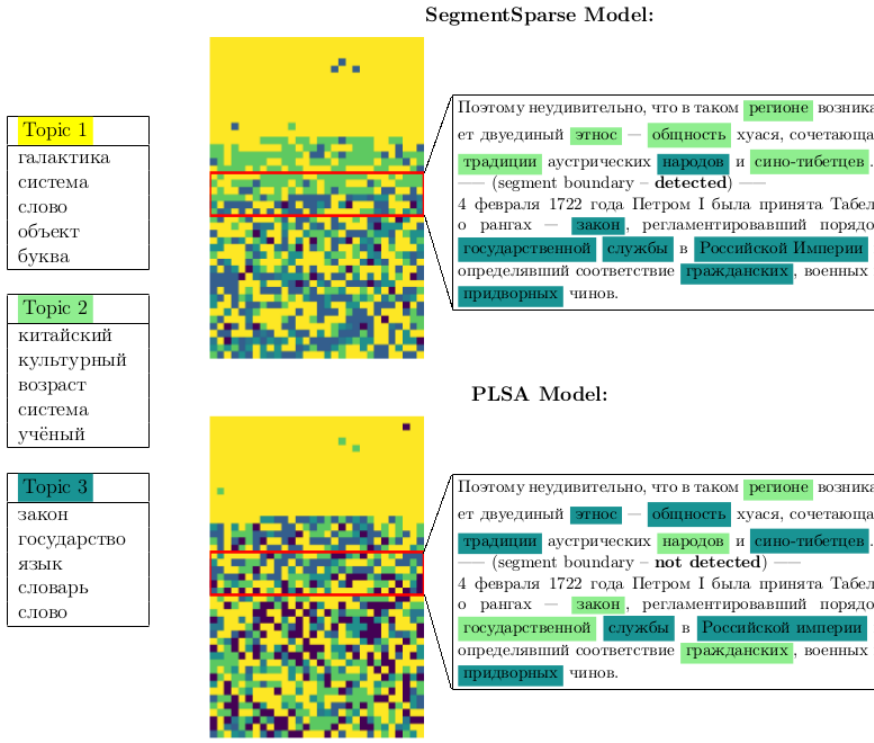

Figure 1. The visualization of PLSA + Θ sparsity reg. (bottom) and SegmentSparse (top) models applied to the same test document. Words on the figure are represented with pixels, which follow each other from left to right and from top to bottom. The text fragment where the PLSA model has failed is marked with red.

... Казанская губерния, наоборот, вошла по просьбе Коржинского. Это интересная историческая ситуация, но постепенно, уже к 20-м годам XX века средняя Россия охватывает территорию от Ярославской и Костромской губерний на севере до Воронежской и Саратовской на юге. Вот эта вся территория находится в пределах европейской части России на левобережье Волги. В среднем, по некоторым подсчётам, природная флора этого региона насчитывает примерно 4,5 тысячи видов, очень немного. Для сравнения флора Турции, которая меньше по площади, включает больше 15 тысяч видов. Связана бедность флоры, с одной стороны, с тем, что это равнинная территория...

Figure 3. The topic representation of an original document from PostScience where SegmentSparse model found 2 segments

6. Conclusion

In this work we have shown, that going beyond the bag-of-words assumption in topic modeling gives a significant improvement in determining text structure. Our enhancements of the EM-algorithm allow to consider words co-occurrences without complicated modifications of the iterative process. Furthermore, the proposed method increases sparsity of documents subjects. This means, the topic models we have built are simpler and provide a better and more interpretable text representation than the models that are built within the bag-of-words assumption. Our iterative segments merging procedure highly increases the segmentation quality of the topic model. As it was shown, the model's confidence in segment boundary identification increased.

For the further research, we are going to implement more assumptions about topic structure in texts, following the same approach. Besides, we would like to focus on the applications of such topic models. Better reflection of the real text structure in a topic model can bring significant improvements to many down-stream tasks.

7. Acknowledgements

The work was supported by Government of the Russian Federation (agreement 05.Y09.21.0018) and the Russian Foundation for Basic Research grant 17-07-01536.

References

1. *Blei D. M., Ng A. Y., Jordan M. I.* (2003), Latent Dirichlet allocation, Journal of Machine Learning Research, Vol. 3, pp. 993–1022.
2. *Hofmann T.* (1999), Probabilistic latent semantic indexing, Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA: ACM, pp. 50–57.
3. *Galley M., McKeown K., Fosler-Lussier E., Jing H.* (2003), Discourse Segmentation of Multi-Party Conversation, ACL '03 Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, Vol. 1, pp 562–569
4. *Vorontsov K., Potapenko A.,* (2014). Additive regularization of topic models. Machine Learning. 101. 1–21. 10.1007/s10994-014-5476-6.
5. *Vorontsov K. V.* (2014), Additive Regularization for Topic Models of Text Collections, Doklady Akademii Nauk, Vol. 455, no. 3.
6. *Vorontsov K., Frey A., Romov P., Yanina A., Suvorova M., Apishev M.* (2015) Bigartm: open-source library for topic modeling of big text collections [Bigartm: biblioteka s otkritim kodom dlya tematicheskogo modelirovaniya bolshich tekstovich kollektsiy] In Analytics and data management in areas with intensive use of data [Analitika I upravleniye v oblastyach s intensivnim ispolzovaniem dannich]. DAMDID/RCDL'2015, Obninsk, pp 28–36.
7. *Martin Riedl, Chris Biemann.* (2012), Text Segmentation with Topic Models. Journal for Language Technology and Computational Linguistics (JLCL) Vol. 27 — pp. 47–69

8. *Lan Du, Wray Buntine, Mark Johnson,* (2013), Topic Segmentation with a Structured Topic Model, Proceedings of NAACL-HLT 2013, pp. 190–200
9. *J. Pitman and M. Yor.* (1997) The two-parameter Poisson-Diriclet distribution derived from a stable subordinator. Annals Probability, Vol.25, pp. 855–900.
10. *G. Balikas, M. Reza Amini, M. Clausel,* (2016). On a Topic Model for Sentences. 10.1145/2911451.2914714.