Context-Aware Latent Dirichlet Allocation for Topic Segmentation

Wenbo Li[1(✉)], Tetsu Matsukawa[1], Hiroto Saigo[1], and Einoshin Suzuki[1,2]

[1] Graduate School and Faculty of Information Science and Electrical Engineering, Kyushu University, Fukuoka, Japan
liwenbo_923@hotmail.com, {matsukawa,saigo,suzuki}@inf.kyushu-u.ac.jp
[2] Graduate School of Systems Life Sciences, Kyushu University, Fukuoka, Japan

Abstract. We propose a new generative model for topic segmentation based on Latent Dirichlet Allocation. The task is to divide a document into a sequence of topically coherent segments, while preserving long topic change-points (coherency) and keeping short topic segments from getting merged (saliency). Most of the existing models either fuse topic segments by keywords or focus on modeling word co-occurrence patterns without merging. They can hardly achieve both coherency and saliency since many words have high uncertainties in topic assignments due to their polysemous nature. To solve this problem, we introduce topic-specific co-occurrence of word pairs within contexts in modeling, to generate more coherent segments and alleviate the influence of irrelevant words on topic assignment. We also design an optimization algorithm to eliminate redundant items in the generated topic segments. Experimental results show that our proposal produces significant improvements in both topic coherence and topic segmentation.

1 Introduction

Topic segmentation is the task of dividing a document into a sequence of topically coherent segments [19]. Specifically, besides the topic distribution, the order of topic segments is also an essential part of document semantic information [18]. Even with the same topic distribution, different orders might represent different or even opposite standpoints. For example, a commentary at the end often determines the guidance of the public opinion, such as the coverage of politics, in particular, election campaigns [6,12]. The challenge of this task is to ensure both the coherency and the saliency of the topic segments, where the coherency refers to keeping long topic segments without being split, while the saliency reserving short topic segments without being absorbed with longer ones.

Conventional topic modeling, such as Latent Dirichlet Allocation (LDA) [5], has made significant progress in various specific applications by handling sparse

A part of this work is supported by Grant-in-Aid for Scientific Research JP18H03290 from the Japan Society for the Promotion of Science (JSPS) and the State Scholarship Fund of China Scholarship Council (grant 201706680067).

H. W. Lauw et al. (Eds.): PAKDD 2020, LNAI 12084, pp. 475–486, 2020.
https://doi.org/10.1007/978-3-030-47426-3_37

high dimensional features and finding latent semantic relationships [14,27]. Nevertheless, the "bag of words" based models are unable to capture the order of topics within each document. A simple solution is to consider the physical structure [2] (e.g., sentences and paragraphs) of each document and use a Hidden Markov Model (HMM) structure [4,9,21,23,24] or predefine a common canonical topic ordering to model the order of topics [8]. However, in recent decades, massive document data are continuously generated in various forms (e.g., news and postings) and from multiple modes (e.g., voice and video). The above models cannot handle these documents with no physical structure information.

Another way is to use high-frequency words as keywords of topics [22]. Detecting and utilizing keywords on the topic assignments improve the coherency of topic segments, especially in documents with well-proportioned topic distribution and sufficient keywords. However, relying heavily on extracted keywords limits the saliency of topic segments. For example, for a document with an uneven topic distribution, extracting enough keywords for all the segments is difficult. As a result, less proportionate topic segments are likely to be absorbed by topic segments with higher proportions, due to insufficient keywords.

The fundamental reason for the limited saliency and coherency is that the topic assignment of each word is highly uncertain. Most words can represent multiple topics, due to their polysemy. The distributional hypothesis [20], which states that words in similar contexts have similar meanings, is one of the primary theories used to quantify the meaning of words according to their context (e.g., Word2vec [11]). Inspired by it, we assume that the topic of each word in a document is related to its context, that is, similar contexts correspond to similar topics. Intuitively, even if a word can be assigned to multiple topics, given its context, we can assign a corresponding topic more certainly. For example, the word "Liverpool" can belong to a topic of sports, geography or art, etc. However, if we combine it to the words in its context (e.g., "Liverpool" & "football" or "Liverpool" & "Beatles"), the assignment is much clearer.

In this paper, we propose a new generative model, Context-Aware Latent Dirichlet Allocation (C-LDA), for document segmentation. In the topic assignment, we consider both the topic distributions and the topic-specific occurrence of word pairs in contexts. Our model enjoys two substantial merits over the state-of-the-art methods: (1) a word is generated by both the document-specific topic distribution and the topic distribution associated with each word and its context; (2) it is independent of physical structures.

2 Related Work

Document segmentation has long been studied in various topic models [4,8,9,21,23,24], such as segHMM [4] and Bayesseg [9]. The traditional methods mainly rely on the document physical structure, which refers to the text-spans in each document, such as sentences or paragraphs [2]. They basically assume that words in the same text-span share the same topic or topic distribution. They conduct segmentation by introducing HMM structure in their topic models and modeling dependencies between consecutive text-spans. However, these approaches

are unable to handle data with no structural information, which significantly limits their applicability. Moreover, in most cases, topics might evolve in long paragraphs or sections, and thus a text-span might contain multiple topics.

Recent studies have been focusing on physical structure-independent segmentation [1,7,22,25]. Topic Keyword Model (TKM) [22] is a topic model based on keywords and their contexts. Its main weakness lies in handling short topic segments, which are likely to be absorbed by long topic segments due to their small number of keywords. Biterm Topic Model (BTM) [7] learns topics by modeling the generation of word co-occurrence patterns, which improves the sensitivity of the discovery of phrases in short text data. On the basis of the former, Bursty Biterm Topic Model (BBTM) introduces a new variable to discover bursty topics[1] [25]. These phrase-level topic modeling methods can achieve good results in discovering word co-occurrence patterns in individual short documents and require no physical structure information. However, high-frequency phrases only make up a tiny proportion of the corpus, which limits their ability to generate coherent topics in topic segmentation tasks. The main difference from our model is that they consider all distinct word pairs of each fixed-size window, while we focus on the topic-specific word pairs, which only concern the target word in the corresponding context. Copula LDA with Segmentation (SegLDA) [1] is an LDA-based model which automatically segments documents into topically coherent sequences of words. SegLDA predefines segments for each document before modeling. For each word in a segment, a topic is assigned either from the segment-specific topic distribution or the document-specific topic distribution. These distributions differentiate the main topics of a document from potential segment-specific topics, which improves the saliency of short segments. However, the two distributions are independent. Specifically, in the former distribution, a topic assignment depends only on the words within the segment, which leads to a loss of much context information in the original document.

In addition, context information is also utilized in other topic models to solve various specific problems in document semantic analysis [16,26], such as Contextual Topic Model (CTM) [26] and Contextual Latent Dirichlet Allocation (Contextual-LDA) [16]. CTM considers the dependencies of topics between each sentence in document summarization while Contextual-LDA uses the topic position of each physical structure-based segment for key information detection. Different from them, we focus on solving the problem of topic segmentation by considering topic-specific word pairs in contexts.

3 Context-Aware Topic Modeling

3.1 Context Word Pairs-Topic Distribution

For conventional LDA and its extended models, topic assignment for each word mostly relies on topic distribution and word distribution. Although the constraints

[1] In their study [25], a topic is considered to be bursty in a time slice if it is heavily discussed, but not in most of the other slices.

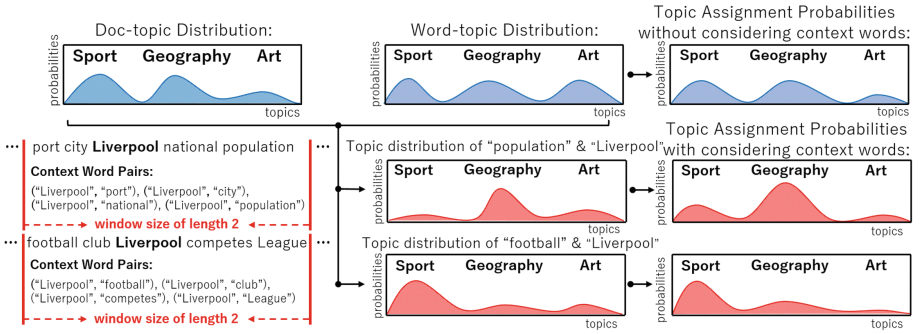


Fig. 1. Schematic illustration of topic assignment for word "Liverpool" with and without considering its context words (respectively labeled by red and blue). We see that if "Liverpool" co-occurs with word "football" in the same context, it is more likely to be assigned to the topic of "sports", while "geography" if co-occurs with "population". (Color figure online)

of topic distribution can alleviate the uncertainty in the topic assignment, it is still insufficient to handle documents containing multiple main topics. For example, a document on the study of modern football and the geographical distribution of England, should at least belong to two topics (geography and sports). We study the topic assignment of the word "Liverpool" in a specific location and consider its 3 related topics: sports, geography, and art. As shown in Fig. 1, for traditional topic models, although the topic distribution reduces the probability of being assigned to the topic of "art", there is still a large uncertainty between "sports" and "geography". However, by considering the frequency of co-occurrence of context words on various topics, this uncertainty can be further reduced, which also coincides with the distributional hypothesis.

Therefore, in our model, we give each word w a context window of length L and define a set of words within the window as context words $\boldsymbol{c}_w$. For the topic assignment of w, we consider the topics of word pairs $\boldsymbol{b}_w$ which consist of w and $\boldsymbol{c}_w$. $\boldsymbol{b}_w$ is defined as:

$$\boldsymbol{b}_w \triangleq \{(w, w') | w' \in \boldsymbol{c}_w\}.$$

Following LDA [5], we also assume that the topic distribution $\boldsymbol{\lambda}_w$ of all the sets of word pairs follows a Dirichlet distribution and name it Context Word Pairs-Topic Distribution (CWTD):

$$\boldsymbol{\lambda}_w \sim Dir(\boldsymbol{\gamma}).$$

$\boldsymbol{\lambda}_w$ depends on the topic distribution of the word pairs of $\boldsymbol{b}_w$ in all other documents. By the definition of Dirichlet distribution [15], the expectation can be calculated as:

$$E_{Dir(\boldsymbol{\gamma})}(\lambda_{w,k}) = \frac{n^{\boldsymbol{b}_w}_{k,-(d,l)} + \gamma_k}{\sum_{s=1}^{K}(n^{\boldsymbol{b}_w}_{s,-(d,l)} + \gamma_s)}, \tag{1}$$

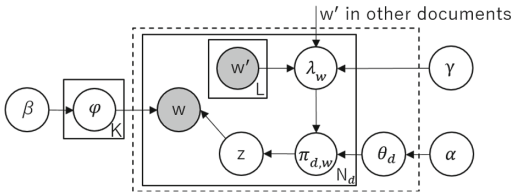


Fig. 2. Graphical model for Context-Aware LDA.

where $n^{\boldsymbol{b}_w}_{s,-(d,l)}$ is the total number of word pairs which are in $\boldsymbol{b}_w$ and belong to topic k in all documents without containing the lth word of document d. In topic assignment, we reorganize the topic distribution $\boldsymbol{\theta}_d$ of a document based on the context of each word and name the reorganized topic distribution as Context-Aware Topic Distribution (CTD), denoted by $\boldsymbol{\pi}_{d,w}$. Therefore, the topic $Z_{d,w}$ for word w in d follows a Categorical distribution which is from the Dirichlet distribution $\boldsymbol{\pi}_{d,w}$ with the prior of both the topic distribution $\boldsymbol{\theta}_d$ and the CWTD $\boldsymbol{\lambda}_w$:

$$\boldsymbol{\pi}_{d,w} \sim Dir(\boldsymbol{\theta}_d + \boldsymbol{\lambda}_w), Z_{d,w} \sim Cat(\boldsymbol{\pi}_{d,w}).$$

3.2 Context-Aware Latent Dirichlet Allocation

As Fig. 2 shows, we introduce four variables $\boldsymbol{\pi}_{d,w}$, $\boldsymbol{\lambda}_w$, w' and $\boldsymbol{\gamma}$ based on traditional LDA, where $\boldsymbol{\pi}_{d,w}$ represents the CTD for word w in document d, $\boldsymbol{\lambda}_w$ is the corresponding CWTD with prior of $\boldsymbol{\gamma}$ and w' refers to a context word of w. Besides, $\boldsymbol{\theta}_d$ represents the topic distribution of document d with prior $\boldsymbol{\alpha}$ and $\boldsymbol{\phi}_k$ is the word distribution of topic k with prior $\boldsymbol{\beta}$. For a dataset of D documents with a vocabulary of size V and latent topics indexed in $\{1, ..., K\}$, C-LDA is associated to the following generative model.

1. Generate the word-topic distribution $\boldsymbol{\phi}_k$ for each topic k: $\boldsymbol{\phi}_k \sim Dir(\boldsymbol{\beta})$.
2. For each document d:
   (a) Generate the topic-word distribution $\boldsymbol{\theta}_d$ of document d: $\boldsymbol{\theta}_d \sim Dir(\boldsymbol{\alpha})$.
   (b) For each word w in d (index by l):
      i. Get context word pairs $\boldsymbol{b}_w$ and generate the CWTD $\boldsymbol{\lambda}_w$ based on Eq. (1): $\boldsymbol{\lambda}_w \sim Dir(\boldsymbol{\gamma})$.
      ii. Generate the CTD $\boldsymbol{\pi}_{d,w}$ of word w according to $\boldsymbol{\theta}_d$ and $\boldsymbol{\lambda}_w$: $\boldsymbol{\pi}_{d,w} \sim Dir(\boldsymbol{\lambda}_w + \boldsymbol{\theta}_d)$.
      iii. Choose a topic $Z_{d,l}$ assignment according to $\boldsymbol{\pi}_{d,w}$: $Z_{d,l} \sim Cat(\boldsymbol{\pi}_{d,w})$.
      iv. Generate $w_{d,l}$ based on the topic $Z_{d,l}$ and $\boldsymbol{\phi}_k$: $w_{d,l} \sim Cat(\boldsymbol{\phi}_{Z_{d,l}})$.

The topic distribution and the context words are combined to further reduce the uncertainty of the topic assignment. As we explain in Sect. 3.4, this reduction ensures a high probability that consecutive words are assigned to the same topic.

Algorithm 1: Gibbs sampling algorithm

Input: A set $\boldsymbol{D}$ of documents with length N_d ($d \in \boldsymbol{D}$); number of iterations N_{iter}; number of topics K

Output: For each document $d \in \boldsymbol{D}$, topic distribution θ_d; for each topic k, word distribution ϕ_k ($1 \leq k \leq K$); word co-occurrence matrix $\boldsymbol{\Lambda}$

```
Initialize topic assignments randomly for all words in D
for iteration = 1 to N_iter do
    for d = 1 to |D| do
        for l = 1 to N_d do
            Generate a topic Z_{d,l} from P_{d,l} according to Eq. (2).
        Update θ_d, φ_k and Λ
return φ_k for each topic k, θ_d for each document d and Λ.
```

3.3 Parameter Estimation

We use Gibbs sampling [10] to estimate parameters. In our sampling procedure, we need to calculate the conditional probability of topic assignment $P_{d,l,k} = P(Z_{d,l} = k | W_{d,l}, \boldsymbol{Z}_{d,-(d,l)}, \boldsymbol{W}'_{d,l}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ for each word, where $W_{d,l}$ represents the lth word in d. $\boldsymbol{Z}_{d,-(d,l)}$ refers to the topic assignments for all words in d except for word $W_{d,l}$. $\boldsymbol{W}'_{d,l}$ are the context words of $W_{d,l}$. The result of $P_{d,l,k}$ is computed as follow (See Appendix A in Supplementary for detailed derivation):

$$P_{d,l,k} \propto \left[(n^{\boldsymbol{b}_w}_{k,-(d,i)} + \gamma_t) + (n_{d,k,-(d,l)} + \alpha_k) \right] \frac{n^t_{k,-(d,l)} + \beta_t}{\sum_{f=1}^{V} (n^f_{k,-(d,l)} + \beta_f)}, \tag{2}$$

where $n_{d,k,-(d,l)}$ is the number of words in d which belongs to topic k without $W_{d,l}$, $n^t_{k,-(d,l)}$ represents the number of word t of topic k without $W_{d,l}$. Compared with the conditional probability of traditional topic models, such as LDA (as Eq. (3)), we see the difference is the probability of topic k for each word, which is affected by the frequency of its context word pairs on topic k in other documents.

$$P'_{d,l,k} \propto (n_{d,k,-(d,l)} + \alpha_k) \frac{n^t_{k,-(d,l)} + \beta_t}{\sum_{f=1}^{V} (n^f_{k,-(d,l)} + \beta_f)}. \tag{3}$$

According to Eq. (2), we obtain the conditional probabilities of topic assignment $P_{d,l,k}$ of each word in document d, so as to compute their corresponding topic distribution $\boldsymbol{P}_{d,l}$. Our sampling algorithm is shown in Algorithm 1. The word co-occurrence matrix $\boldsymbol{\Lambda}$ recording the number of word pairs in each topic is utilized to compute $\boldsymbol{\lambda}$, where the first two dimensions of $\boldsymbol{\Lambda}$ are all the unique words and the third dimension records the accumulated shared topic counts.

3.4 Topic Coherency Ratio

To further study how C-LDA affects the coherency and saliency in modeling, we calculate the joint probability of consecutive words which share the same topic

in two cases: with and without considering context word pairs. For consecutive words $\boldsymbol{W}_{d,i:j}$ from W_i to W_j in document d, we denote the joint probability of sharing topic k by $P(\boldsymbol{W}_{d,i:j}, k)$ in the former case and the one in the latter case by $P'(\boldsymbol{W}_{d,i:j}, k)$. Taking their logarithms and computing their ratios as well as removing constant terms, we obtain the result as shown in Eq. (4). We retain the fraction of the right-hand side and name it Topic Coherency Ratio (TCR) as Eq. (5), denoted by R_t (See Appendix B in Supplementary for detailed derivation).

$$\frac{\log P(\boldsymbol{W}_{d,i:j}, k)}{\log P'(\boldsymbol{W}_{d,i:j}, k)} \propto 1 + \frac{\sum_{w \in \boldsymbol{W}_{d,i:j}} \log n_{k,-\boldsymbol{W}_{d,i:j}}^{\boldsymbol{b}_w}}{\sum_{w \in \boldsymbol{W}_{d,i:j}} \log n_{k,-\boldsymbol{W}_{d,i:j}}^{w}} \tag{4}$$

$$R_t(\boldsymbol{W}_{d,i:j}, k) \triangleq \frac{\sum_{w \in \boldsymbol{W}_{d,i:j}} \log n_{k,-\boldsymbol{W}_{d,i:j}}^{\boldsymbol{b}_w}}{\sum_{w \in \boldsymbol{W}_{d,i:j}} \log n_{k,-\boldsymbol{W}_{d,i:j}}^{w}}. \tag{5}$$

For a set of consecutive words, the TCR is a ratio of occurrence number in the same topic between the context word pairs and words. The ratio ranges from $[0, 1]$ and reflects the intensity of coherency for a set of consecutive words[2]. A higher ratio corresponds to a stronger coherency. By Eq. (4), we see $P(\boldsymbol{W}_{d,i:j}, k)$ is always greater than $P'(\boldsymbol{W}_{d,i:j}, k)$, which proves that C-LDA is more likely to generate coherent topic segments than other conventional topic models, including LDA and most of its extended versions[3]. For short segments consisting of a tiny proportion of words in a document, they can still be assigned to the topic k with a higher probability than others if they contain frequent word pairs in topic k. Thus C-LDA ensures both better coherency and saliency in topic segmentation.

Since the number K of topics is a given empirical value, it is inevitable to generate redundant topic segments in each document. Although we might be able to specify a good K value beforehand, the difference in the number of topics contained in each document also leads to the inevitability of generating redundant segments. Therefore, merging redundant segments with frequent ones is indispensable, where the key is to judge whether the resulting segment has a higher coherency than the original ones. The TCR is a coherency measurement based on the ratio of word pairs and words instead of relying solely on their frequencies. This property ensures the coherency of segments are independent of their lengths; thus, we design a TCR based Redundant Topic Merging (RTM) algorithm to optimize the generated topic segments. The steps of RTM are: for each topic segment, we consider three cases: (1) merging with the previous segment; (2) merging with the next segment; (3) non-merging. For these three cases, the TCRs are calculated separately and the case with the highest ratio is selected. We repeat the above steps until the number of segments stays unchanged.

[2] For the words in $\boldsymbol{W}_{d,i:j}$ belonging to topic k, if and only if they all occur as context word pairs of topic k in all the documents, the TCR gets the maximum value 1, while it gets the minimum value 0 if and only if none of them occurs in a context.

[3] The fraction on the right-hand side is always positive.

4 Experiments

We evaluate our model by a series of experiments. Results were obtained with eight-fold cross-validation on a machine with Intel i9 processor and 128 GB memory. The hyper-parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ were all fixed to 0.05.

We tested our model on three standard datasets[4] (**Wikicities** (Wici), **Cellphones Reviews** (Cell) and **Wikielements** (Wiel)) and three extended datasets based on the former three. **Wikicities** contains Wikipedia articles about the world 100 largest cities by population, **Cellphones Reviews** contains 100 cellphone reviews and **Wikielements** contains 118 English Wikipedia articles about chemical elements. Labeled topic segments of the 3 standard datasets are all of the similar lengths (about 3000 words per document) and uniformly distributed; thus, to simulate the cases of more diverse topic structures, we increase their original total number of documents to 2000 and generated various sizes of topic segments for each document. The detailed generating steps for a document are: (1) select the number of segments based on a uniform distribution from 10 to 50; (2) for each segment, set its length from a uniform distribution of 10 to 100 and randomly assign it to a topic from the topic labels; (3) choose sentences of the corresponding assigned topics from the labeled documents to fill the segments until all segments are loaded.

We compare C-LDA (available on Github[5]) against four topic models: LDA [5], BTM [7], TKM [22] and SegLDA [1]. BTM is a topic model based on word co-occurrence modeling. TKM is a method to generate coherent topics by considering the influence of keywords on their contexts. SegLDA is a LDA-extended model for topic segmentation by introducing an independent topic distribution for each predefined segment.

We use Normalized Point-wise Mutual Information (NPMI)[6] to measure the topic coherence scores [17]. It assumes that a topic is more coherent if the most probable words in the topic co-occur more frequently in the corpus [13]. NPMI scores are in $[-1, 1]$ and a higher value indicates that the topic distributions are semantically more coherent. The performances of topic segmentation is evaluated with two metrics: PK[7] and Window Diff (WD)[8]. They both refer to error rates which are calculated by comparing the inferred segmentation with the gold-standard (ground truth) for each window based on moving a sliding window over the document. Lower scores refer to better segmentation performance.

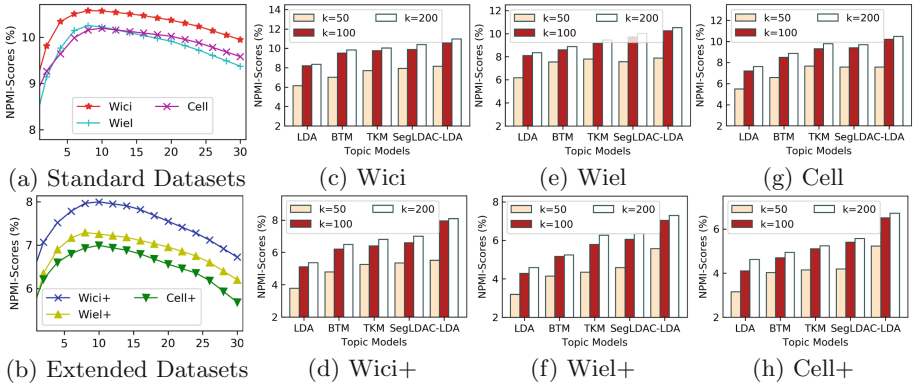
[4] http://groups.csail.mit.edu/rbg/code/mallows/.

[5] https://github.com/liliverpool/C-LDA.git.

[6] $\text{NPMI}(k) = \sum_{1 \le i < j \le T} \frac{1}{-\log P(w_i, w_j)} \log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}$, where $P(w_i, w_j)$ and $P(w_i)$ are the occurrence probabilities of word pair (w_i, w_j) and word w_i, respectively.

[7] $P_k(\text{ref}, \text{hyp}) = P(\text{false}|\text{refer}, \text{hyp}, \text{same}, k)P(\text{same}|\text{refer}, k) + P(\text{miss}|\text{refer}, \text{hyp}, \text{diff}, k)$ $P(\text{diff}|\text{refer}, k)$, where "refer" is the ground truth and "hyp" is the generated segments. k is usually the half of the average gold-standard segment size ($k = 15$ in our experiments). More details are in [3].

[8] $WD(\text{ref}, \text{hyp}) = \frac{1}{N-k} \sum_{i=1}^{N-k} (|b(\text{ref}_i, \text{ref}_{i+k}) - b(\text{hyp}_i, \text{hyp}_{i+k})| > 0)$, where $b(i, j)$ represents the number of boundaries between positions i and j in the text and N is the number of sentences in the document [17].



(a) Standard Datasets (c) Wici (e) Wiel (g) Cell

(b) Extended Datasets (d) Wici+ (f) Wiel+ (h) Cell+

Fig. 3. NPMIs of different L values (a–b) and different topic numbers k (c–h).

4.1 Topic Coherence

Firstly, we calculated NPMIs of C-LDA under different window sizes L (from 1 to 30) with topic number $K = 100$. The results are shown in Fig. 3(a–b). We see that, in both standard and extended datasets, NPMIs increase sharply until around $L = 10$ then begin to decline. Moreover, we see there is a sharp decline in the extended datasets when $L > 15$. This might be because of their more complex topic structures and longer window sizes are more likely to contain irrelevant content. Therefore, we set $L = 10$ in the rest of our experiments.

The results of NPMIs (with $K = 50, 100, 200$) for all baseline models are shown in Fig. 3(c–h). We see that C-LDA shows the best results on all six datasets and more significant improvements in data sets with more complex topic structures (Wici+, Wile+, Cell+), which proves the validity of our model for generating coherent topics. A possible reason is that C-LDA combines the frequency of context word pairs for each topic in modeling, while the other models (such as TKM) either consider only the words frequency in each topic or the frequency of all word pairs in individual documents (such as BTM). Moreover, semantic expressions in a document are usually coherent and segmented, e.g., paragraphs and sections, thus, considering the context in a topic assignment can clarify the semantics of the word, so as to reduce the risk of splitting a coherent semantic segment.

4.2 Topic Segmentation

The results in topics of $K = 50$ and $K = 100$ are shown in Table 1, where the C-LDA-R is the C-LDA with RTM optimization algorithm. We see that C-LDA and C-LDA-R perform the best in all cases of $K = 100$ and dominate in most cases when $K = 50$, which validates their performance for coherency and saliency of different segments in topic segmentation tasks.

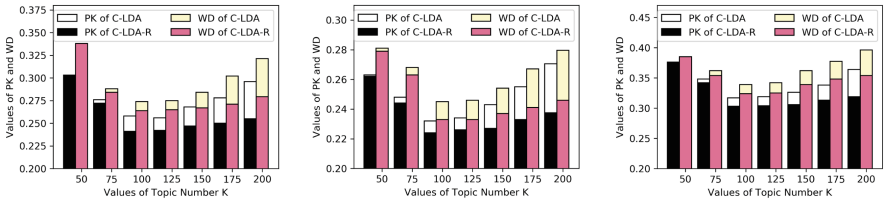
BTM aims to generate all the distinct word pairs within a fixed window given a topic. Therefore, its effect on the topic coherency is achieved by increasing the

Table 1. Topic segmentation results. PK and WD scores are in %. Bold fonts indicate best scores yielded by models except for C-LDA-R and * indicates the best scores among all the models.

K	Models	PK						WindowDiff						Time Cost (hours)					
		Wici.	Wiel.	Cell.	Wici$^+$	Wiel$^+$	Cell$^+$	Wici.	Wiel.	Cell.	Wici$^+$	Wiel$^+$	Cell$^+$	Wici.	Wiel.	Cell.	Wici$^+$	Wiel$^+$	Cell$^+$
50	BTM	35.9	33.6	41.2	42.2	38.7	47.1	38.2	34.5	41.0	45.6	42.1	49.8	9.2	7.2	7.7	4.9	4.2	4.2
	TKM	28.6	23.9	37.8	33.8	28.5	43.2	28.7	33.4	38.7	35.8	31.8	46.4	**1.1***	**0.7***	**0.8***	**0.4***	**0.3***	**0.3***
	SegLDA	26.1	22.7	**35.2**	30.5	27.2	38.9	**27.1***	**25.6**	35.8	**33.4***	28.3	39.3	4.5	3.1	3.3	1.7	1.4	1.5
	C-LDA	**25.3**	**22.2**	35.3	**29.9***	**26.3**	**37.6**	27.7	25.7	**34.1***	33.7	**28.2**	**38.1***	1.9	1.2	1.5	0.9	0.8	0.8
	C-LDA-R	24.8*	20.6*	34.9*	30.3	26.2*	37.5*	27.3	24.8*	33.5	33.8	27.9*	38.5	2.0	1.3	1.6	1.0	0.9	0.9
100	BTM	32.5	30.2	37.5	40.1	36.5	44.3	35.7	33.4	40.2	41.4	39.8	45.7	15.7	12.6	11.5	8.6	8.2	8.5
	TKM	26.7	21.2	30.6	31.3	27.4	37.2	29.9	24.6	36.6	32.8	29.8	41.7	**2.1***	**1.5***	**1.7***	**0.9***	**0.7***	**0.8***
	SegLDA	23.2	20.4	31.3	27.4	24.1	34.8	28.5	23.9	33.5	29.8	24.6	36.3	8.8	6.5	7.6	3.1	2.4	2.6
	C-LDA	**22.1**	**19.7**	**29.8**	**25.8**	**23.2**	**31.7**	**27.5**	**22.6**	**32.2**	**27.4**	**24.5**	**33.9**	4.2	3.2	3.8	2.4	2.3	2.4
	C-LDA-R	21.9*	19.2*	27.6*	24.5*	22.6*	30.4*	25.2*	21.6*	30.7*	26.8*	23.7*	32.4*	4.3	3.3	3.9	2.5	2.4	2.5

joint probability of each word pair and topics. The high frequent word pairs in a corpus are of high joint probabilities. However, in a corpus, the majority are ordinary words but not word pairs, and their topic assignments are still of high uncertainty. Besides, the computation of all distinct word pairs significantly increases its training time. TKM improves the coherency of topic segmentation by considering the influence of keywords on the topic assignment of surrounding words and cost the least time. However, short topic segments with insufficient keywords are likely to be absorbed by long topic segments, which is a possible reason of its low performance. In some cases of insufficient topic number ($K = 50$), SegLDA outperforms other methods. However, as K increased from 50 to 100, its performance growth is inferior to C-LDA. For SegLDA, the topics for words in a segment can be assigned from the segment-specific topic distribution, which improves the saliency of topic segments. However, assigning topics without considering the original document can lead to a loss of context information and degrade the accuracy of topic modeling. That is, a word is possibly assigned to an incorrect topic even if it is not absorbed by others. C-LDA considers both the contextual word pairs and topic distribution. Based on the reorganized topic distribution CTD, it reduces the uncertainty of the topic assignment and increases the joint probability of consecutive words sharing the same topic at the expense of increasing time consumption. Moreover, comparing the results of the original and their extended datasets, we see our method has stronger robustness to more complex topic structures, which also leads to better applicability.

For C-LDA-R, we see that the effect of the RTM algorithm is limited in the case when $K = 50$ since it is insufficient to cover all the occurred topics. When $K = 100$, RTM effectively improves the performance of topic segmentation. To further study the effect of RTM, we calculated the changes of PK and WindowDiff with different numbers K of topics (from 25 to 200). The experiments were conducted on the 3 extended datasets and the results are shown in Fig. 4. We see the measures of both C-LDA and C-LDA-R decrease quickly with the increase in length of K until $K = 100$. For C-LDA, the performance starts to



(a) Wici$^+$ (b) Wiel$^+$ (b) Cell$^+$

Fig. 4. PK and WindowDiff scores with the increase of the number K of topics in (a) Wikicities$^+$, (b) Wikielements$^+$ and (c) CellphoneReviews$^+$. The stacked part above each bar is the improvement from RTM algorithm.

decrease around $K = 150$, while for C-LDA-R, it tends to saturate as K keeps on increasing. The improvement by RTM becomes increasingly remarkable with the increase of K, which also proves the robustness of C-LDA-R for redundant topics. In addition, the time complexity of RTM for each document is $O(L \sum_{S' \in S} |S'|)$, where L is the context window size, S is the list of segments for a document and $|S'|$ represents the length of each segment S' in S. The time consumption of the RTM is acceptable, since L is set less than 30. Besides, the optimization process of each document is independent, which is easy for parallelization.

5 Conclusion

We proposed a new generative model for topic segmentation. By combining topic distribution and context word pairs-topic distribution, our model improves the certainty of the topic assignment and ensures high coherency and saliency in topic segmentation. Besides, we designed an optimization algorithm to merge redundant topic segments for each document. Our experiments show that our proposal outperforms baseline models, in terms of the segmentation scores of PK and WD in topic segmentation. In future work, we will further optimize the parameter estimation steps, such as reducing the size of the word co-occurrence matrix, and use more efficient estimation methods (e.g., Variational Inference).

References

1. Amoualian, H., Lu, W., Gaussier, M.: Topical coherence in LDA-based models through induced segmentation. In: Proceedings of ACL, vol. 1, pp. 1799–1809 (2017)
2. Balikas, G., Amoualian, H., Clausel, M., Gaussier, E., Amini, M.R.: Modeling topic dependencies in semantically coherent text spans with copulas. In: Proceedings of COLING, pp. 1767–1776 (2016)
3. Beeferman, D., Berger, A., Lafferty, J.: Statistical models for text segmentation. Mach. Learn. **34**(1–3), 177–210 (1999)
4. Blei, D.M., Moreno, P.J.: Topic segmentation with an aspect hidden Markov model. In: Proceedings of SIGIR, pp. 343–348 (2001)

5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
6. Cappella, J.N., Jamieson, K.H.: Spiral of Cynicism: The Press and the PublicGood. Oxford University Press, New York (1997)
7. Cheng, X., Yan, X., Lan, Y., Guo, J.: BTM: topic modeling over short texts. IEEE TKDE **26**(12), 2928–2941 (2014)
8. Du, L., Pate, J.K., Johnson, M.: Topic segmentation with an ordering-based topic model. In: Proceedings of AAAI (2015)
9. Eisenstein, J., Barzilay, R.: Bayesian unsupervised topic segmentation. In: Proceedings of EMNLP, pp. 334–343 (2008)
10. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. PAMI **6**(6), 721–741 (2009)
11. Goldberg, Y., Levy, O.: Word2vec explained: deriving Mikolov et al'.s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722 (2014)
12. Jenks, J.W.: The guidance of public opinion. Am. J. Sociol. **1**(2), 158–169 (1895)
13. Lamprier, S., Amghar, T., Levrat, B., Saubion, F.: On evaluation methodologies for text segmentation algorithms. In: Proceedings of ICTAI, vol. 2, pp. 19–26 (2007)
14. Liang, S., Ren, Z., Yilmaz, E., Kanoulas, E.: Collaborative user clustering for short text streams. In: Proceedings of AAAI, pp. 3504–3510 (2017)
15. Ng, K.W., Tian, G.L., Tang, M.L.: Dirichlet and Related Distributions: Theory, Methods and Applications, vol. 888. Wiley, Oxford (2011)
16. Peng, D., Guilan, D., Yong, Z.: Contextual-LDA: a context coherent latent topic model for mining large corpora. In: Proceedings of BigMM, pp. 420–425. IEEE (2016)
17. Pevzner, L., Hearst, M.A.: A critique and improvement of an evaluation metric for text segmentation. Comput. Linguist. **28**(1), 19–36 (2002)
18. Purver, M.: Topic Segmentation. Spoken Language Understanding: Systems for Extracting Semantic Information from Speech, pp. 291–317 (2011)
19. Reynar, J.C.: Topic segmentation: algorithms and applications. Ph.D. thesis, Institute for Research in Cognitive Science Technical, University of Pennsylvania (1998)
20. Sahlgren, M.: The distributional hypothesis. Ital. J. Disabil. Stud. **20**, 33–53 (2008)
21. Sauper, C., Haghighi, A., Barzilay, R.: Content models with attitude. In: Proceedings of ACL, pp. 350–358 (2011)
22. Schneider, J., Vlachos, M.: Topic modeling based on keywords and context. In: Proceedings of ICDM, pp. 369–377 (2018)
23. Wang, H., Zhang, D., Zhai, C.: Structural topic model for latent topical structure analysis. In: Proceedings of ACL, pp. 1526–1535 (2011)
24. Wang, X., McCallum, A., Wei, X.: Topical n-grams: phrase and topic discovery, with an application to information retrieval. In: Proceedings of ICDM, pp. 697–702 (2007)
25. Yan, X., Guo, J., Lan, Y., Xu, J., Cheng, X.: A probabilistic model for bursty topic discovery in microblogs. In: Proceedings of AAAI (2015)
26. Yang, G., Wen, D., Chen, N.S., Sutinen, E., et al.: A novel contextual topic model for multi-document summarization. Expert Syst. Appl. **42**(3), 1340–1352 (2015)
27. Yin, J., Wang, J.: A Dirichlet multinomial mixture model-based approach for short text clustering. In: Proceedings of SIGKDD, pp. 233–242 (2014)