Estimating Aspects in Online Reviews Using Topic Model with 2-Level Learning

Takuya Konishi, Taro Tezuka, Fuminori Kimura, and Akira Maeda

***Abstract*—In this paper, we propose a method for estimating latent aspects in online review documents. Review aspects represent features of items or services evaluated by users. We can expect to acquire useful features for users by discovering their review aspects. We apply topic models to this problem. Existing work proposed methods for estimating the topics of the whole document or sets of sentences with various window sizes. In this paper, we propose two-level learning approach that connects adjacent sentences when their topics are similar, and re-estimates topics once again using the determined processing units. In the experiments of precision using perplexity, we confirm our proposed method improves on the existing method.**

***Index Terms*—online review, topic model, sentiment analysis, text mining.**

I. Introduction

The quantity of various online documents published on the Web continues to increase. Therefore, we need elaborate and effective methods to acquire useful information from these documents on the Web. However, it is sometimes difficult to analyze these documents by general frameworks, because user generated information is not well organized; for example, sparseness of information, noises, or biases. For instance, user generated documents are not necessarily long enough (e.g. blogs, twitter, or BBS). These shortness of descriptions in documents is a major problem when these documents are analyzed by general data mining techniques, natural language processing, or sophisticated statistical models. In these cases, properties of the target document sets must be account for.

We focus on online review documents. Review documents describe the properties of items or services and are provided by reviewers through blogs or consumer generated media. They provide readers useful information for decision-making when buying the reviewed item or service. Since we can obtain these review documents in bulk nowadays, many researchers have been studying them by means of statistical approaches over the last ten years. One of the main topics in this domain is sentiment analysis. Pang et al. thought of this problem as supervised document classification [1]. They applied some classification models (Naive Bayes, Maximum entropy model, and SVM) to review documents given sentiment aspect (positive or negative), and evaluated effectiveness of these precision.

T. Konishi is with the Graduate School of Science and Engineering, Ritsumeikan University, 1-1-1 Noji-Higashi, Kusatsu, Shiga, Japan, e-mail: (cm005069@ed.ritsumei.ac.jp).

F. Kimura, and A. Maeda are with the Colleage of Information Science and Engineering, Ritsumeikan University, 1-1-1 Noji-Higashi, Kusatsu, Shiga, Japan, e-mail: ({amaeda@media, fkimura@is}.ritsumei.ac.jp).

T. Tezuka is with the Graduate School of Library, Information and Media Studies, University of Tsukuba, 1-2 Kasuga, Tsukuba City, Ibaraki, 305-8550, Japan, email: (tezuka@slis.tsukuba.ac.jp).

In this paper, we consider existence of review aspects. Each subject comes into some categories; for instance, consumer electronics has digital cameras, televisions, desktops, or laptops. We assume that a review document set in which each evaluated subject belongs to the same category has shared aspects. The discovery of these aspects is beneficial for users to understand subjects and review documents.

In this task, we apply topic models used for document modeling [2]. This method assumes each document has multiple topics. Topics are represented by a probabilistic generative model and are estimated from the document set. We assume that these topics correspond to the above-mentioned review aspects. Topic models hereby represent each review aspect as word distribution. While ordinary topic models can treat some general documents, they cannot extract meaningful topics from review documents because of the similarity across each document.

In previous work, Titov and McDonald developed a more enhanced topic model for this problem called Multi-grain LDA [3]. It is possible to extract review aspects by incorporating the topic model with a nonparametric framework. It is a sophisticated model using a variety of latent variables.

In this paper, we propose a two-level learning topic model algorithm for review documents. It consists of two steps. First, it estimates topics for each sentence in the first learning, and connects sentences using the result of first learning. Second, it learns all grouped sentences data in the second learning. While this method is somewhat simple, we show that it can reduce complexity in a quantitative experiment.

The remainder of this paper is structured as follows. In Section 2, we explain the problem when we treat review documents in topic models. In Section 3, we show Multi-grain LDA (MG-LDA), which catches the property of review document, and also show simplified Multi-grain LDA (sMG-LDA), which focuses on only local topics and provides our baseline model. Additionally, we mention the problem of using window elements in terms of complexity. In Section 4, we describe our proposed algorithm. Section 5 provides an empirical experiment of between the baseline model and our method. Finally, we summarize our paper and discuss future work.

II. The Characteristics of Review Documents

In this section, we describe the characteristics of review documents that are the problems in extracting topics. Topic models basically require document sets to be represented as “bag-of-words”. Bag-of-words is a simple assumption that ignores the order of words when processing words in each group (e.g. a document) and only accounts for the frequency of each word. This assumption is convenient to employ to the probabilistic models for extracting global properties of

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

the target document set like topics. General topic models apply this assumption to document level; that is to say, the data has only information about how many times each word appears in each document.

However, this assumption causes the problem for topic extraction in review documents. All reviews that describe the same category are very similar to each other. For instance, consider reviews of digital cameras. We can suppose "image quality" of a digital camera is a review aspect. This topic is one of the important aspects characterizing digital camera reviews. Consequently, it will appear in almost all the documents. In other words, words representing image quality will appear in any document. This problem makes extracting these topics difficult because words related to this topic that appear are very similar to each other. In fact, through the experiment of applying latent Dirichlet allocation, which is the most general topic model [2], to review documents, we found that we cannot extract the desired meaningful topics we want. Therefore, we have to account for more local distinguishing features in each document.

The simple idea is to assume sentence level bag-of-words. Sentences represent more local features than documents. In fact this assumption works well to some extent. One problem in this idea is the lack of features in each sentence. If we remove the non-important feature words (e.g. stop words) in all sentences, they hardly have any feature words left. This is a problem when the model composes sentence-specific topic distribution because each distribution is somewhat sparse.

In conclusion, we hope that topic models capture local topics representing features of smaller units than a whole document. In addition, the distribution composed by the models is more informative. The key point for solving the problem is how to determine the processing units. We need to decide the appropriate size of units, which is larger than a sentence but smaller than a document.

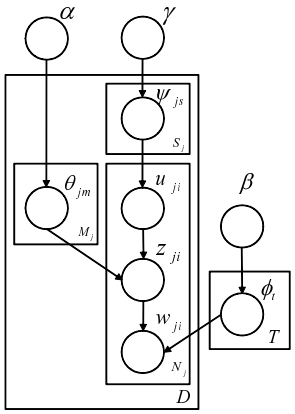

Fig. 1. A graphical representation in simplified Multi-grain LDA.

Algorithm 1 Generative process of sMG-LDA

for all topic t (=1...T) **do**
  Draw $\boldsymbol{\phi}_t \sim Dir(\boldsymbol{\beta})$
end for
for all document j (=1...D) **do**
  for all sentence s of document j (s=1... S_j) **do**
    Draw $\boldsymbol{\psi}_{js} \sim Dir(\boldsymbol{\gamma})$
  end for
  for all sliding window m of document j (m=1...M_j) **do**
    Draw $\boldsymbol{\theta}_{jm} \sim Dir(\boldsymbol{\alpha})$
  end for
  for all token i in sentence s of doucment j(i=1...N_j) **do**
    Draw $u_{ji} \sim Multi(\boldsymbol{\psi}_{js})$
    Draw $z_{ji} \sim Multi(\boldsymbol{\theta}_{ju_{ji}})$
    Draw $w_{ji} \sim Multi(\boldsymbol{\phi}_{z_{ji}})$
  end for
end for

III. Related Work

In this section we briefly introduce the specific topic model for review documents called Multi-grain LDA [3]. The model is represented as a probabilistic generative model. We overview the model and discuss the role of windows in this model.

A. MG-LDA

Titov and McDonald developed the Multi-grain LDA, which captures the global and local topics [3]. They assumed that review documents had global topics that represent global properties appearing in document level (e.g. item specific topics) and local topics that represent local properties appearing in some sentence levels (i.e. ratable aspect). We tested the effectiveness of this model. However it did not necessarily extract meaningful global topics. In contrast it works well for extracting local topics. In addition, since our method only focuses on local topics, we ignore the existence of global topics in the model. We call the simplified Multi-grain LDA model "sMG-LDA" in this paper.

sMG-LDA is a probabilistic generative model. In general, probabilistic generative models define the generative process of data, and all data are generated in accordance with this process. sMG-LDA follows the generative process of Algorithm 1.

Before we start to explain this process, we denote some notation. First T is the number of topics, D is the number of documents and S_j is the number of sentences in the document j. M_j is the number of windows in document j. This is decided by window size K and S_j, i.e. $M_j = K + S_j - 1$. Also $Dir()$ denotes the Dirichlet distribution, and $Multi()$ denotes the multinomial distribution in the above generative process. Additionally, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ are hyperparameter vectors for each Dirichlet distribution.

First, this model samples word probability vectors ϕ_t from the Dirichlet prior $Dir(\boldsymbol{\beta})$ for each topic t. For each document, this model samples window probability vectors ψ_{js} from Dirichlet prior $Dir(\boldsymbol{\gamma})$ for each sentence, and samples topic probability vectors θ_{jm} from Dirichlet prior $Dir(\boldsymbol{\alpha})$ for each window. Finally, it samples three variables for each word in all documents. u_{ji} denotes a window assignment of token i in document j, z_{ji} denotes a topic assignment of token i in document j, and w_{ji} denotes a word of token i in document j. These three kinds of variables are sampled from corresponding multinomial distribution.

Under these generative processes, we need to estimate topic assignment z and window assignment u for each token.

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

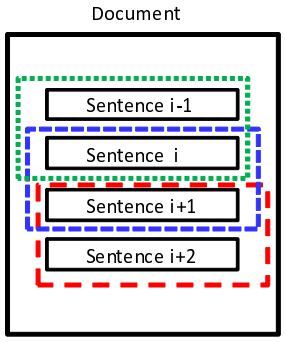

Fig. 2. An illustration of window elements in Multi-grain topic model. This instance denotes two window sizes.

We employ the Gibbs sampling to infer this model. The conditional probabilities for Gibbs sampling are given by :

$$p(z_{ji} = t|w_{ji} = v, u_{ji} = m, \mathbf{w}^{-ji}, \mathbf{z}^{-ji}, \mathbf{u}^{-ji}) = \frac{n^{v}_{t,-ji} + \beta}{n^{\cdot}_{t,-ji} + V\beta} \frac{n^{j,m}_{t,-ji} + \alpha_t}{n^{j,m}_{\cdot,-ji} + \alpha_0} \quad (1)$$

$$p(u_{ji} = m|z_{ji} = t, \mathbf{w}, \mathbf{z}^{-ji}, \mathbf{u}^{-ji}) = \frac{n^{j,m}_{t,-ji} + \alpha_t}{n^{j,m}_{\cdot,-ji} + \alpha_0} \frac{n^{j,s}_{m,-ji} + \gamma}{n^{j,s}_{\cdot,-ji} + K\gamma} \quad (2)$$

where $n^{v}_{t,-ji}$ is the number of times word v appears in topic t except token ji (i.e token i in document j), $n^{j,m}_{t,-ji}$ is the number of times topic t appears in window m of document j except token ji, and $n^{j,s}_{m,-ji}$ is the number of times window m appears in sentence s of document j except token ji. Each dot with count notations in denominators represents marginal counts. In addition, all values of hyperparameter vector $\boldsymbol{\alpha}$ affect estimation result sensitively, so we estimate them using Minka's fixed point iteration [4]. Remaining hyperparameter vectors set the same value for the elements of each vector (specifically, $\boldsymbol{\beta} = 0.1$ and $\boldsymbol{\gamma} = 1.0$). We show the graphical model for this model in Figure 1.

B. Complexity of Previous model

We consider the window element of Multi-grain topic models in this subsection. The window size affects how many sentences the model accounts for (see Figure 2). Since the model can take account of not only one sentence but also neighboring sentences using a window, we may need to consider the role of the window. sMG-LDA (also MG-LDA) requires a decision on the window size in model selection as well as the number of topics.

sMG-LDA put a mixture model of window variables into practice, that is to say, this model assumes that each sentence has multiple windows. This is a little complex for the data, and we consider only one window is enough for each sentence. Moreover, this model needs to fix the window size across the target corpus. This results in a lack of flexibility. For an extreme example, a document composed of one sentence is assigned to two or three windows. This is obviously redundant.

In the next section, we propose the model that accounts for neighboring sentences without multiple windows. We will solve the problem using two-level learning.

IV. Proposed Algorithm

We propose a new learning algorithm toward the above, mentioned problem. This mainly consists of the first learning, connecting neighboring sentences, and the second learning. Firstly, it conducts the first learning for every one sentence. The result gives interim topic assignments for each token. These assignments give topic distribution for each sentence indirectly. Specifically, the topic distribution of sentence s in document j is given as follows:

$$p(\boldsymbol{z}|\boldsymbol{\theta}_{js}) = \prod_{t=1}^{T} (\theta_{jst})^{z_t} \quad (3)$$

where $p(\boldsymbol{z}|\boldsymbol{\theta}_{js})$ denotes the topic distribution of sentence s in document j, and is represented as multinomial distribution (also called categorical distribution). $\boldsymbol{z}$ is the 1-of-K representation (and z_t is the tth element of the $\boldsymbol{z}$), and θ_{jst} is the tth element of topic probability vector $\boldsymbol{\theta}_{js}$ (i.e. tth topic probability value $p(\boldsymbol{z} = t|\boldsymbol{\theta}_{js}) = \theta_{jst}$).

Then our method connects each sentence using the estimated topic distribution. We consider boundaries of sentences in each document. The number of boundaries in document j, B_j is $S_j - 1$. The method evaluates the connection of neighboring sentences for each boundary. While we need the criterion to decide whether the sentences are connecting or not, we apply the Jensen-Shannon divergence (JS-divergence) to evaluate it. JS-divergence is applied when measuring the similarity of probability distribution. Unlike the Kullback-Leibler divergence, which is the common measure for similarity of probability distribution, JS-divergence satisfies the symmetric relation. We employ this measure to obtain the similarity of neighboring sentences. Therefore, we can calculate the similarity between the sentence immediately before the boundary s^{-}_{jb} and the sentence immediately behind the boundary s^{+}_{jb} in boundary b of document j using JS-divergence as follows:

$$sim(s^{-}_{jb}, s^{+}_{jb}) = JS(p^{-}||p^{+}) = \frac{1}{2}\left(\sum_{t=1}^{T} p^{-}_{t} \log \frac{2p^{-}_{t}}{p^{-}_{t} + p^{+}_{t}} + \sum_{t=1}^{T} p^{+}_{t} \log \frac{2p^{+}_{t}}{p^{-}_{t} + p^{+}_{t}}\right) \quad (4)$$

where

$$p^{-} = p(\boldsymbol{z}|\boldsymbol{\theta}_{js^{-}_{jb}}), p^{+} = p(\boldsymbol{z}|\boldsymbol{\theta}_{js^{+}_{jb}})$$
$$p^{-}_{t} = p(z_t = t|\boldsymbol{\theta}_{j^{-}_{jb}}), p^{+}_{t} = p(z_t = t|\boldsymbol{\theta}_{j^{+}_{jb}})$$

According to the property of JS-divergence, the similarity has non-negativity and becomes zero if and only if each distribution is equal to each other. If the value is lower, the method interprets the similarity to be higher. Our method evaluates neighboring sentences using this measure, and if the value of similarity is below the threshold, these sentences are connected. The decision of the threshold will be described in the next section. As the result of the above process, we can obtain the new dataset in which semantically similar sentences are connected.

Finally, the method conducts the second learning using the renewed dataset. Our proposed algorithm is shown in Algorithm 2 and illustrated in Figure 3. We denote that

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

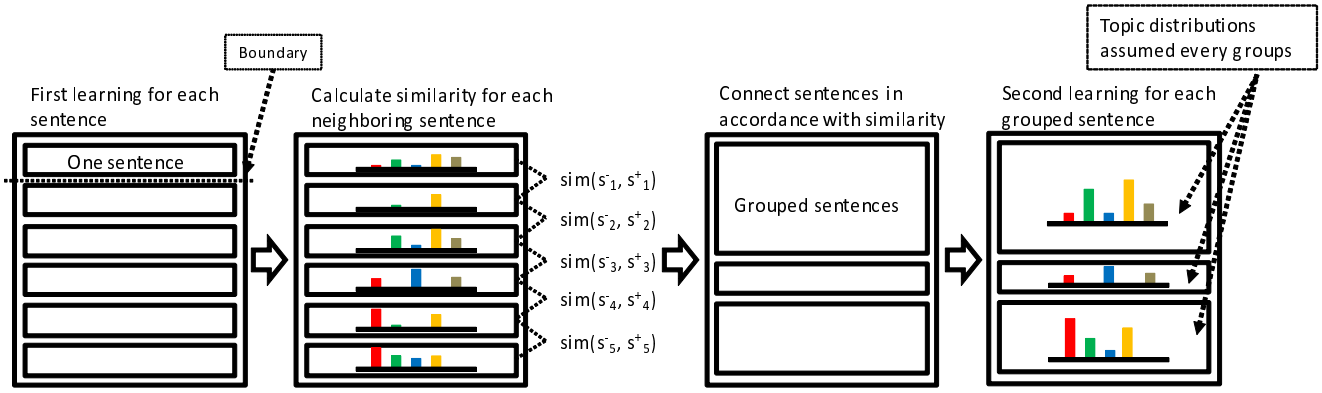

Fig. 3. An illustration of our proposed algorithm. The key idea is to connect neighboring sentences according to the first learning. We can obtain the topic distribution for each group of sentences in the second learning.

Algorithm 2 Our proposed algorithm

1: Execute first learning for processing unit that is in each sentence
2: **for all** document j (=1...D) **do**
3: **for all** sentence boundary b (=1...B_j) **do**
4: Calculate the similarity sim(s^-_{jb}, s^+_{jb}) of neighboring sentences s^-_{jb} and s^+_{jb} according to (4)
5: **end for**
6: **for all** sentence boundary b (=1...B_j) **do**
7: **if** sim(s^-_{jb}, s^+_{jb}) $<$ threshold **then**
8: Connect the neighboring sentences
9: **end if**
10: **end for**
11: **end for**
12: Execute second learning for renewed processing unit that groups semantically similar sentences by the above processes

second learning assumes topic distribution in every grouped sentence in Figure 3. The distribution has more information than the distribution for each sentence.

The idea of the proposed algorithm is considerably simple and does not need additional variables to control the window unlike the sMG-LDA. While the algorithm assumes it is possible that first learning estimates topics for each sentence to some extent, it can determine the adaptive processing units for each document. It does not overfit toward a document that has only a few sentences.

V. Experimental Results

We describe the empirical results of our proposed algorithm in this section. First, we explain the setup for this experiment.

We prepare two Japanese review document sets and one English review document set. For Japanese reviews, we use two real online review document sets about digital cameras and laptops. We obtained these documents from the Japanese online price comparison site, kakaku.com [5]. Kakaku.com provides the information of various items or services, and publishes review documents contributed by users. We collected 13,638 reviews about digital cameras and 12,211 reviews about laptops from this site. We used nouns as feature words in these data sets.

For English reviews, we use one review set about digital camera in Amazon.com [6]. We collected 11,279 digital camera reviews. We removed some stop words and used stemming for this data. We utilized these three data sets as experimental data.

We used perplexity to evaluate models. Perplexity indicates the performance of prediction for new words in each model. We used 90% words as training data, and 10 % words as test data. Perplexity is defined as follows:

$$\exp(-\frac{1}{N^{test}}\log p(\mathbf{w}_{test}|\mathbf{w}_{training})) \quad (5)$$

where N^{test} is the number of words in the test data, $\mathbf{w}_{test}$ is the set of test words, and $\mathbf{w}_{training}$ is the set of training words. We compare the proposed algorithm with the previous model. The model to compare is sMG-LDA introduced in section 3.

Our proposed method needs to determine a threshold of the similarity in each boundary. In this experiment, we determined in accordance with rankings of similarities. A ranking of similarities can be obtained as a sequence of similarities in descending order (i.e. ascending order in the value). In this way, the threshold changes in every model selection (e.g. number of topics). We employ higher-ranking value in the ranking, and demonstrated the preliminary experiment of perplexity in shifting the proportion of the connected sentences on the basis of similarities. For Japanese reviews, the top 20% of higher similarities are the best result in digital camera reviews, and the top 25% are the same in laptop reviews (i.e. thresholds are equal to the value at each percentage). For English reviews, the top 30% are the same in digital cameras. We make use of values gained in this way to compare our results to those in previous work.

First we show the experimental results of Japanese reviews in Figure 4 and 5. Figure 4 shows the result of digital camera reviews, and Figure 5 shows the results of laptop reviews. In both figures, our proposed algorithm surpasses window size 2 and 3 models, and slightly does so for the window size 1. Generally, while all models hardly differ from each other in small numbers of topics, large differences arise in larger numbers of topics. While the improvement is not large when compared to the window size 1 model, our method succeeded in the improving in the situation accounting for surrounding sentences differently from the other two models.

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

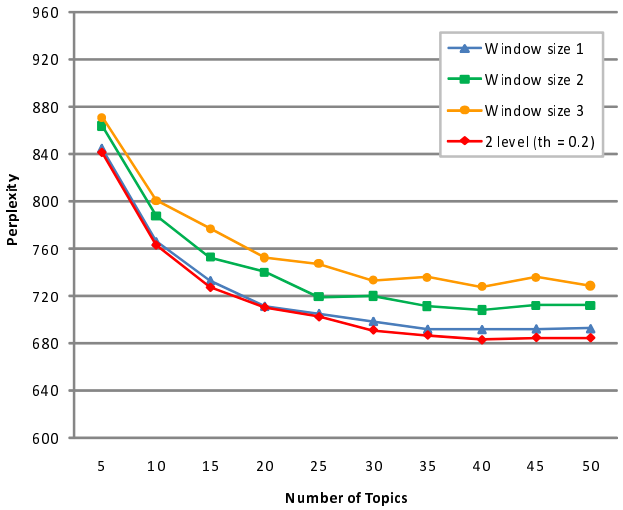

Fig. 4. A comparison between our proposed algorithm and sMG-LDA models in Japanese digital camera reviews.

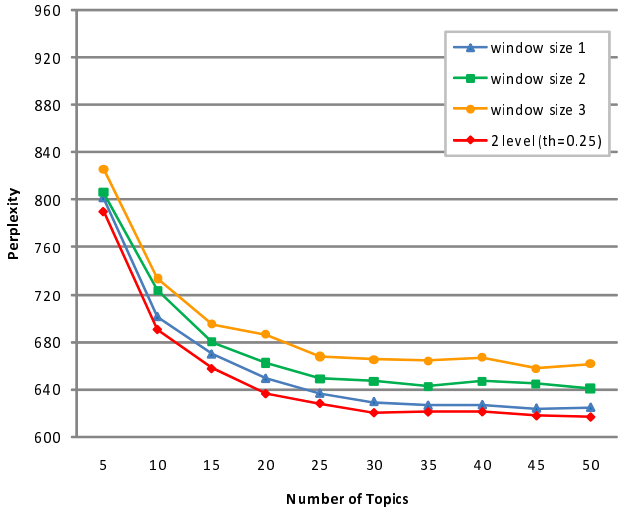

Fig. 5. A comparison between our proposed algorithm and sMG-LDA models in Japanese laptop reviews.

Figure 6 shows the results of English reviews about digital cameras. While the trend is mainly the same as the results of Japanese reviews, it shows the greater improvement than Japanese ones for our method compared to sMG-LDA models. We think these differences are caused by the difference of document structure between Japanese and English reviews. Japanese reviews are comparatively succinct, and are relatively short on average. English reviews are more descriptive than Japanese ones, and are relatively long on average. For those reasons, our algorithm works well in English than Japanese.

In these experiments, we selected fixed thresholds to connect semantically similar sentences as described above. We decided them on basis of empirical results. However, we will consider ways that are more suitable to decide the threshold. If the proposed method decides connections on the basis of more substantiated criterion, we can expect the method to be improved.

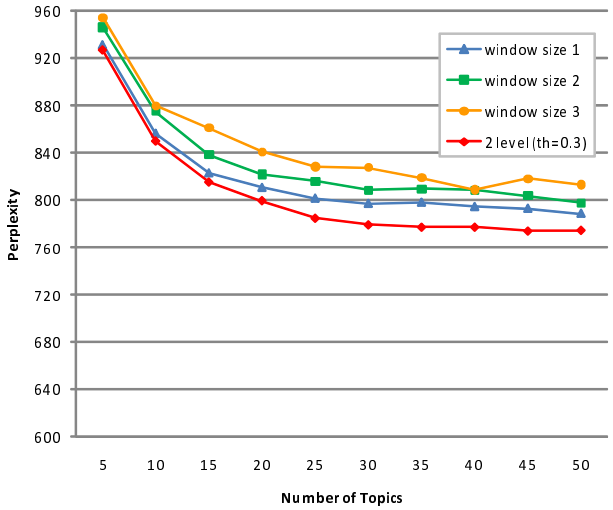

Fig. 6. A comparison between our proposed algorithm and sMG-LDA models in English digital camera reviews.

VI. Conclusion

We proposed the new algorithm for estimating the topics in online reviews. We focused on the complexity of the previous work called MG-LDA, and we considered a more plain framework that uses natural assumption. The results of this model were better than those of the previous work in terms of perplexity, which is used the general index in evaluating topic models.

Here we consider other related works and the future work. We think our proposed idea is closely related to tasks about text segmentation or paragraph detection [7][8]. They need the appropriate text segment and characterizing documents. We expect that our method is applicable to these tasks. Also we executed the our algorithm by using two-step learning. It is not necessarily elegant to use simply separated learning from the point of view of a probabilistic model. Thus, we are going to consider the adaptive model that can execute learning once.

Acknowledgment

This work was supported in part by MEXT Grant-in-Aid for Strategic Formation of Research Infrastructure for Private University “Sharing of Research Resources by Digitization and Utilization of Art and Cultural Materials” (Grant Number: S0991041) and MEXT Grant-in-Aid for Young Scientists (B) “Research on Information Access across Languages, Periods, and Cultures” (Leader: Akira Maeda, Grant Number: 21700271).

References

[1] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? Sentiment Classification using Machine Learning Techniques,” in *Proc of Conference on Empirical Methods in Natural Language Processing*, pp.79-86, 2002.
[2] D. M. Blei, A. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” in *Journal of Machine Learning Research*, Vol.3, No.5, pp.993-1022, 2003.
[3] I. Titov and R. McDonald, “Modeling Online Reviews with Multi-grain Topic Models,” in *Proc of the 17th International World Wide Web Conference*, pp.112-120, 2008.
[4] T. P. Minka, “Estimating a Dirichlet distribution,” Technical report, http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/,2003
[5] kakaku.com, http://kakaku.com/

[6] Amazon.com, http://www.amazon.com/
[7] J. Eisenstein and R. Barzilay, "Bayesian unsupervised topic segmentation," in *Proc of Conference on Empirical Methods in Natural Language Processing*, pp.334-343, 2008.
[8] M. Jeong and I. Titov, "Multi-document topic segmentation," in *Proc of the 19th ACM international conference on Information and knowledge management*, pp.1119-1128, 2010.

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)