World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:10, No:1, 2016

Improving Topic Quality of Scripts by Using Scene Similarity Based Word Co-Occurrence

Yunseok Noh, Chang-Uk Kwak, Sun-Joong Kim, Seong-Bae Park

Abstract—Scripts are one of the basic text resources to understand broadcasting contents. Topic modeling is the method to get the summary of the broadcasting contents from its scripts. Generally, scripts represent contents descriptively with directions and speeches, and provide scene segments that can be seen as semantic units. Therefore, a script can be topic modeled by treating a scene segment as a document. Because scene segments consist of speeches mainly, however, relatively small co-occurrences among words in the scene segments are observed. This causes inevitably the bad quality of topics by statistical learning method. To tackle this problem, we propose a method to improve topic quality with additional word co-occurrence information obtained using scene similarities. The main idea of improving topic quality is that the information that two or more texts are topically related can be useful to learn high quality of topics. In addition, more accurate topical representations lead to get information more accurate whether two texts are related or not. In this paper, we regard two scene segments are related if their topical similarity is high enough. We also consider that words are co-occurred if they are in topically related scene segments together. By iteratively inferring topics and determining semantically neighborhood scene segments, we draw a topic space represents broadcasting contents well. In the experiments, we showed the proposed method generates a higher quality of topics from Korean drama scripts than the baselines.

Keywords—Broadcasting contents, generalized Pólya urn model, scripts, text similarity, topic model.

I. Introduction

LOTS of tools for mining various kinds of text documents have been developed in decades. Topic models are one of the popular tools for modeling such text data [1] because of its flexibility for model extension and language independent characteristic. Broadcasting scripts, however, have not been received much attention from the topic modeling community though the scripts are useful to understand broadcasting contents. As text features, scripts contain rich semantic information of broadcasting contents and its scene segment structure provides proper environment to discover topics of a script by dealing with each scene segment as a document. Learning a script by a conventional topic model, however, may suffer from lack of statistics such as word co-occurrences. This is because many scene segments are short and words in speeches tend not to appear repeatedly over scene segments. These phenomena make a topic model difficult to learn topics of high quality. Therefore, a solution to overcome the sparse statistical information should be required.

In order to handle a script with topic modeling, the scene segment structure can be used. Because scene segments can be regarded as basic semantic units of a script, learning scene segments means learning storylines in the script. In general, a broadcasting script is written for filming, hence scenes are segmented based on shooting locations. This causes occasionally very short scene segments consecutively. For example, phone call scenes of two characters are often divided into lines by each character though all those scenes belong to same storyline. To obtain topics of good quality under this circumstance, we turn the characteristic of scene segments into the clue for the better modeling. A scene segment is incomplete itself but have some neighborhood segments semantically strongly dependent. Going back to the previous example, all segmented scenes about the phone call can be treated as one single semantic scene. With the semantic dependency over scene segments, one can be further assumed that topically dependent scene segments are highly similar in terms of the topic. We can then reach the idea that learning topics and finding topical dependencies over scene segments can help each other. If the assumption about dependencies among scene segments are accepted, it is obvious that topics of higher quality lead to finding more accurate topical dependencies over scene segments. Given topical dependencies can then be used to force topics of scene segments topically dependent more similar [13], [7].

In this paper, we study topic modeling on a broadcasting script with scene segment similarities and word co-occurrence expansion based on the similarities. Our goal is to uncover salient topics in terms of storylines despite of the lack of statistics in a script. The intuition is to jointly improve the confidence of dependencies among scene segments and topic quality iteratively. Specifically, topical dependency among scene segments are determined by topical similarity among them. By using the determined dependency, topics of scene segments are inferred again. The more this procedure is conducted repeatedly, the higher confidence of the dependency and topic will be achieved.

The generalized Pólya urn framework [10] is adopted for obtaining more semantically coherent topics. In the GPU model, co-occurred words are explicitly forced to be put into the same topic to improve topical coherence in unsupervised fashion. In the case of the scene segments, because of its data sparsity, the word co-occurrence is needed to be expanded to topically dependent segments. That is, the dependency affects topics indirectly through the GPU framework. Learning the GPU model with the expanded word co-occurrence encourages uncovered topics to fit into the semantic scenes. Hence, once

Y. Noh and S.-J. Kim are with Smart Media Platform Section, ETRI, Daejeon, Korea (e-mail: noh60085@AT etri.re.kr, kimsj@AT etri.re.kr).

C.-U. Kwak and S.-B. Park are with the School of Computer Science and Engineering, Kyungpook National University, Daegu, 702-701, Korea (e-mail: cukwak@AT sejong.knu.ac.kr, sbpark@AT sejong.knu.ac.kr).

the trustworthy topical dependencies over scene segments are grasped, semantically coherent topics in terms of storylines can be obtained.

The proposed method was evaluated with a Korean drama script, the first episode of *'Heard it through the grapevine'*. Experimental results show that the proposed method outperforms baselines in terms of perplexity and topic coherence.

II. Related Work

The main concerns of this study are twofold: (1) Utilizing topical dependencies of scene segments to learn topics of higher quality; (2) improving topical coherence.

A. Utilizing Topical Dependencies

Modeling the relationship among documents have been studed in many topic models. Purver et al. [13] model the utterances of meeting transcripts for topic segmentation. They explicitly modeled the dependencies between the consecutive utterances. In their model, the topic distribution of current utterance may be affected by the previous utterance. Current utterance, of course, can influence the next utterance. Dirac delta function is used to compel topic distribution of an utterance to be same with that of the previous utterance when they are dependent. In [6], hyper-linked texts on the web are modeled with some additional random variables. Instead of assuming direct dependency between topic distributions of web pages, they introduced the latent link variable generates a link and the latent link variable is dependent on the topic distibution of other web pages. These studies models text data with dependencies under well-designed generative process. However, their methods are for explicit dependencies which are observable while the scene segments do not have.

Du et al. [5] proposed a method of topic modeling that exploits document relative similarities as a regularizer. They maximized the model log-likelihood with the constraint where the distance among similar documents should be much smaller with certain margin than the distance among dissimilar documents. This work supports us in terms of utilizing intra corpus document similarities to enhance the model quality. Because the volume of a script is small, however, it is necessarily to find more elaborate and direct relationship among texts rather than the macroscopic relative similarities in the entire dataset.

B. Improving Topic Coherence

Recently, improving topic coherence has been paid more attention in the research area. Xie et al. [14] used external knowledge to learn coherent topics. They incorporated word correlation into the topic model where the correlation is obtained by using global word semantics from external knowledge such as word2vec [9]. But their approach may not work well in the context of the broadcasting script, because some words of broadcasting scripts do not appear in general knowledge base e.g. character name, and words in a script are not related together in terms of general sense, but are associated according to its storyline. In contrast, Mimno et al. [10] enhanced LDA in terms of topic coherence by integrating the metric for topic coherence into the progress of topic assignment. Chen et al. [4] also utilized the idea of [10], but their goal was to adopt knowledge from the topics of external domain data for learning a new domain. These two works introduced a generalized Pólya urn framework to topic models successfully, and they showed the potential to improve topic model quality in unsupervised fashion. By standing on the shoulders of these works, we study topic models to fit more to the broadcasting scripts.

There are rarely studies of analyzing the broadcasting scripts using topic model framework, but [11] is one of them. Misra et al. explored the topic segmentation of TV news using text features - closed caption of TV news and some image features. LDA was used for obtaining topic distributions of closed captions, and then they suggested an algorithm for TV news story segmentation based on dynamic programming. Our work not just find scene segment dependencies - this may correspond to story segmentation, but also seek superior topics in terms of coherence. We resolve two goals in a single method.

III. Learning Topics of a Script

Let S be the script to learn topics, then we can define $S = \{s_1, s_2, \ldots, s_{|S|}\}$, where s_i is a scene segment and $|S|$ is the number of scene segments in the script. Then our objective is to infer Θ and Φ that maximize the following likelihood function:

$$\log(p(S|\Theta, \Phi)p(\Theta|\alpha)p(\Phi|\beta)). \tag{1}$$

In our problem, however, each s_i has few information which makes a topic model learn topics difficult. Instead it is assumed that there are latent dependencies among scene segments if they belong to the same storyline. Here, we introduce a dependency matrix $\mathbf{B}$, a $|S| \times |S|$ binary valued matrix, represents those dependencies over all scene segments. In this paper, we infer the dependencies using topic distributions of scene segments simply. That is, two scene segments are regarded as being dependent if the similarity of two topic distributions of two scene segments are high enough. By adding a term of decay $f(\cdot)$ for the distance between the indexes of scene segments, b_{ij}, the element of $\mathbf{B}$, can be defined as follows:

$$b_{ij} = \begin{cases} 1 & \text{if } sim(\theta_i, \theta_j) f(|i-j|) > threshold \\ 0 & \text{others,} \end{cases} \tag{2}$$

where θ_i is obviously the topic distribution of s_i. Scene segments generally tend to follow time stream, that is why we add the decay term by the scene segment index.

A. A Naïve Dependency Model

It is needed that the method ties words in dependent scene segments together and puts those words into the same topic. One can be a naïve method that uses topic distributions of dependent scene segments where $b_{ij} = 1$ as base measures for inferring θ_i of s_i [7]. Then, the conditional distribution of the topic assignment is defined as

$$P(z_k = t|z^{-k}, S, \alpha, \beta, \mathbf{B}) \propto \\ (n_{s_i,t}^{-k} + \alpha\bar{\theta}_{dep(s_i)}) \times \frac{\sum_v n_{t,v}^{-k} + \beta}{\sum_{v'}(\sum_v n_{t,v}^{-k} + \beta)} \quad (3)$$

where n^{-k} is the count except the current assignment of z_k, that is z^{-k}, w_k is the current word to be sampled with a topic z_k. $n_{s_i,t}$ refers to the number of times that topic t was assigned to words in scene segment s_i and $n_{t,v}$ denotes the number of times that word v appears in topic t. α and β are symmetric Dirichlet parameters. $\bar{\theta}_{dep(s_i)}$ is the expectation of topic distributions come from dependent scene segments of s_i. $\bar{\theta}_{dep(s_i)}$ works as the prior of s_i where it presses topics of high probability in $\bar{\theta}_{dep(s_i)}$ to arise more in the procedure of topic sampling of s_i. This naïve model, however, just make topic distributions similar, but do not get involved in topic quality. Hence, we explore further to find the method to deal with the topical dependencies.

B. Generalized Pólya Urn Model

Generalized Pólya urn (GPU) model [10] is a modified version of latent Dirichlet allocation (LDA) [2]. The GPU model was devised to improve topic quality in terms of coherence. Co-document frequency of words are widely used in many studies to evaluate topic coherence, and the GPU model incorporates those kinds of metrics into topic model framework directly.

The essential difference between the GPU model and LDA is the update scheme of topic-word component i.e. Φ. The GPU model employs a generalized Pólya urn framework [8] for topic-word component, while it follows the update scheme of LDA for document-topic component. In LDA which follows the simple Pólya urn model in contrast with the GPU model, a ball of certain color may be drawn from an urn. Then the ball is put back to the urn along with another ball of the same color. Under the topic model setting, a ball of certain color can be seen as a word of certain type and an urn as a topic.

In the GPU model, having drawn a word w of particular type, A_{vw} additional words of each type $v \in \{1, \ldots, V\}$ are put back to the topic. $\mathbf{A}$, a $V \times V$ real-valued matrix where V is the size of vocabulary, has the correlations among words as its elements. Therefore, when w is assigned to a topic t, each word v is also assigned to the topic t with the amount of A_{vw}. For learning the GPU model, $\mathbf{A}$ is constructed by using word co-occurrence information. Since the GPU model is nonexchangeable, sequential Monte Carlo methods [3] or the method of approximating the true Gibbs sampling distribution can be used for posterior inference. See [10] for more details.

C. Topical Dependencies Based Word Co-Occurrence

The GPU model uses the statistics of word co-occurrence information in its own dataset. This unsupervised fashion have an advantage of needing no additional knowledge to improve the model. However, as mentioned before, scene segments have little information of word co-occurrence. As the result, $\mathbf{A}$ tends to be limited to have only non-zero values for words within the same scene segment. This may cause the learned topics are fitted to each scene segment not to meaningful storylines. To tackle this problem, the dependency matrix $\mathbf{B}$ is used to release the judgement of being co-occurred from a scene segment s_i to all dependent scene segments s_j where $b_{ij} = 1$. $\mathbf{A}'$, the modified $\mathbf{A}$ using $\mathbf{B}$, can then be defined as

$$\mathbf{A}'_{vw} \propto \lambda_v \sum_i^{|S|} \sum_j^{|S|} b_{ij} s^i(w) s^j(v) \quad (4)$$

where $s^i(w)$ returns 1 if the scene segment s_i contains w, 0 if not. We follow [10], λ_v is set to $\log(|S|/D(v))$, where $D(v)$ is the number of scene segments contain word v. Each column of $\mathbf{A}'$ is also normalized to sum to 1.

When a topic is sampled and assigned to a word with $\mathbf{A}'$, only co-scene segment words have not the chance for follwing to the topic, but words in dependent scene segments can also be guided to the same topic in this modified GPU model. Topical dependency information $\mathbf{B}$ has an effect on topics of scene segments indirectly in the modified GPU model compared to Equation (3) of the naïve dependency model. This indirect reflection of topical dependency also lets the proposed model be apart from the influence of the wrong inferred dependency.

D. Posterior Inference

The conditional distribution for the topic assignments is as

$$P(z_k = t|z^{-k}, S, \alpha, \beta, \mathbf{A}') \propto \\ (n_{s_i,t}^{-k} + \alpha) \times \frac{\sum_v n_{t,v}^{-k} \times A'_{v,w_k,t} + \beta}{\sum_{v'}(\sum_v n_{t,v}^{-k} \times A'_{t,v,v'} + \beta)} \quad (5)$$

where $\mathbf{A}'$ is the schema defined in (4). The approach to compute the conditional distribution is to approximate the true Gibbs sampling distribution by dealing with each word as if it were the last, where the approach was presented by [10]. It is needed to be devised that an algorithm updates those parameters jointly and iteratively because the dependency matrix $\mathbf{B}$ and the topic model parameters Θ, Φ have to affect each other.

Algorithm 1 summarizes the process of learning topics with $\mathbf{A}'$ and $\mathbf{B}$. In the first, the algorithm starts learning with $\mathbf{A}'$ and $\mathbf{B}$ initialized to be $\mathbf{I}$. This initialization reduces the proposed GPU model to the standard LDA. After certain iteration, $\mathbf{A}'$ and $\mathbf{B}$ are updated based on Θ somewhat confident. The algorithm then again learns topics using $\mathbf{A}'$ that is based on $\mathbf{B}$ somewhat confident. By performing this process iteratively, Θ and $\mathbf{B}$ promote each other to become more confident, so that the objective can be achieved.

IV. Experiments

A. Data and Setting

To evaluate our method, we used a Korean TV drama script, the first episode of *'Heard it through the grapevine'*. This script consists of 55 scene segments, which were segmented by '#' mark that means the beginning of a scene. Table I

Algorithm 1: Jointly iterative topic learning with $\mathbf{A}'$ and $\mathbf{B}$

```
Input  : A script with scene segments $s_1, s_2, \ldots, s_{|S|}$, # iterations $nIter$, period to update $\mathbf{A}', \mathbf{B}$ $nUpdate$
Output: A topic model with $\Theta, \Phi$
initialize topic assignments randomly for all tokens
initialize $\mathbf{A}'$ and $\mathbf{B}$ as identity matrix
$iter \leftarrow 1$
repeat
    for $i = 1, \ldots, |S|$ do
        for $w_n \in s_i$ do
            $n_{s_i,z_k} \leftarrow n_{s_i,z_k} - 1$
            for all $v$ do
                $n_{z_k,v} \leftarrow n_{z_k,v} - A'_{v,w_k}$
            end
            draw $z_k \propto (n_{s_i,t+\alpha})\frac{n_{w_k,t}+\beta}{\sum_{z'}(n_{z',w_k}+\beta)}$
            $n_{z_k,v} \leftarrow n_{s_i,z_k} + 1$
            for all $v$ do
                $n_{z_k,v} \leftarrow n_{z_k,v} + A'_{v,w_k}$
            end
        end
    end
    compute the posterior estimates of $\Theta$ and $\Phi$
    if $iter\%nUpdate = 0$ then
        update $\mathbf{B}$ as in Equation (2)
        update $\mathbf{A}'$ as in Equation (4)
    end
until $iter < nIter$;
```

shows detailed information of the script. As shown in Table I, the number of scene segments and the number of words in each scene segments are small. In addition, the average document frequency per word is just 1.48. To think that the entire scene segments are involved in the identical content, we can be aware that rare occurrence of words is too severe to obtain good topics compared to general dataset such as news data. We remained the last 10 scene segments as held-out data to measure the perplexity of presented models.

In experiments, four methods were evaluated including the proposed method.

- LDA. The standard LDA model.
- NDM. The naïve dependency model in (3).
- GPU. The generalized Pólya urn model [10].
- MGPU. The modified GPU model, that is our proposed model.

For the proposed model, we performed 200 runs of Gibbs sampling to get somewhat confident topics and used the expectation of topics of those 200 runs for computing $\mathbf{B}$ and $\mathbf{A}'$. Exponential decay function was used and the threshold was set to 0.5 for (2). We set Dirichlet parameters α to 1, β to 0.001 and the number of topics to 20. These hyper parameters were shared with all topic models in experiments.

We measured all models in two quantitative evaluation tasks. One is held-out perplexity and the other is topic coherence. Perplexity on held-out data is one of the basic metrics for topic model evaluation [2]. This metric presents how the trained model predicts unseen data well, and the perplexity for our held-out dataset is defined as follows.

$$perplexity(S^{test}|\Theta, \Phi) = \exp\left\{ - \frac{\sum_{d=1}^{|S^{test}|} \log p(\mathbf{w}_d|\Theta, \Phi)}{\sum_{d=1}^{|S^{test}|} |s_d|} \right\}. \quad (6)$$

Lately, there have been trends for refining topics in human sense rather than leaving topics as data driven raw stuff. Topic coherence is the metric follows the trends and presented in many studies of topic model evaluation [10], [12]. In this paper, the point-wise mutual information (PMI) [12] is used to measure topic coherence, which is defined as:

$$\mathrm{PMI}(\mathbf{w}) = \frac{2}{N(N-1)} \sum_{1 \le i < j \le N} \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}, \quad (7)$$

where $\mathbf{w}$ are the top-N words of a topic, $p(w_i, w_j)$ is the probability that words w_i and w_j co-occur in the same scene segment and $p(w_i)$ is the probability of word w_i.

B. Results and Discussion

We give the results of perplexity and topic coherence in Table II. As shown in the table, the proposed method outperformed all baselines in terms of both evaluation measures. In a broadcasting script, a continuous story may be divided into several scene segments, and words in those segments may not co-occur. This may do the role of penalty when topic models predict unseen data because words have not co-occured in training data can co-occur in unseen segments belong to the same story. The GPU model which showed the worst score in perplexity aggravates this circumstance by forcing words explicitly co-occurred to have high probability in the same topic. On the other hand, the proposed method showed the ability of predicting unseen data by using topical dependencies. We also found that the proposed model produced more coherent topics. Even the naïve dependency model showed better performance than the GPU model in topic coherence measure. This result supports our intuition that topical dependencies and topics of scene segments can co-promotes together. The proposed model reflects both aspects of the GPU model and the naïve dependency model, then eventually discovered topics are more coherent and more predictable.

V. Conclusion

We proposed an algorithm of learning topic model for a broadcasting script. The conventional topic models have the limits for modeling scene segments of a script. Because the scene segments are relatively small corpus and have few word co-occurrence statistics, a method that makes the best use of additional information. The proposed method utilizes the similarities among topic distributions of scene segments, then finally leads topical dependencies over scene segments and coherent topics of given script. By iteratively learning topics and inferring topical dependencies, both encourage to be better

TABLE I
SIMPLE STATISTICS OF THE 1ST EPISODE OF *'Heard it Through the Grapevine'*

Data	# of scene segments	average # of tokens per scene segements	average document frequency of words
Train	45	23.18	1.48
Test	10	24.6	1.13

TABLE II
EXPERIMENTAL RESULTS

Metric	LDA	GPU	NDM	MGPU
perplexity	3.4855	4.3392	3.7597	**3.1232**
coherence	-0.153	-0.0717	-0.0271	**-0.0104**

together. This idea is adopted to the GPU model, then the proposed method can obtain more coherent topics as well.

We proved that the proposed method is more appropriate to uncover topics in a script than the baselines through a series of experiments. By achieving good performance in both perplexity and topic coherence, we empirically proved our intuition makes sense, and our method models the broadcasting script effectively.

Although we showed the superiority of our proposed model, there still remain rooms for improvement. In this paper, we inferred the latent dependencies among scene segments with very simple method, so more elaborate and reasonable method is demanded. We remain this for our future work.

ACKNOWLEDGMENT

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No. B0126-15-1002, Development of smart broadcast service platform based on semantic cluster to build an open-media ecosystem).

REFERENCES

[1] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[3] K. R. Canini, L. Shi, and T. L. Griffiths, "Online inference of topics with latent dirichlet allocation," in *International conference on artificial intelligence and statistics*, 2009, pp. 65–72.

[4] Z. Chen and B. Liu, "Topic modeling using topics from many domains, lifelong learning and big data," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 703–711.

[5] J. Du, J. Jiang, D. Song, and L. Liao, "Topic modeling with document relative similarities," in *Proceedings of the 24th International Conference on Artificial Intelligence*. AAAI Press, 2015, pp. 3469–3475.

[6] A. Gruber, M. Rosen-Zvi, and Y. Weiss, "Latent topic models for hypertext," in *Proceedings of the Twenty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-08)*. AUAI Press, 2008, pp. 230–239.

[7] Y.-J. Han, S.-Y. Park, and S.-B. Park, "A single-directional influence topic model using call and proximity logs simultaneously," *Soft Computing*, pp. 1–17, 2015.

[8] H. Mahmoud, *Pólya urn models*. CRC press, 2008.

[9] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[10] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 262–272.

[11] H. Misra, F. Hopfgartner, A. Goyal, P. Punitha, and J. M. Jose, "Tv news story segmentation based on semantic coherence and content similarity," in *Advances in Multimedia Modeling*. Springer, 2010, pp. 347–357.

[12] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 100–108.

[13] M. Purver, T. L. Griffiths, K. P. Körding, and J. B. Tenenbaum, "Unsupervised topic modelling for multi-party spoken discourse," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2006, pp. 17–24.

[14] P. Xie, D. Yang, and E. Xing, "Incorporating word correlation knowledge into topic modeling," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2015, pp. 725–734.