International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:10, No:1, 2016

Online Topic Model for Broadcasting Contents Using Semantic Correlation Information

Chang-Uk Kwak, Sun-Joong Kim, Seong-Bae Park, Sang-Jo Lee

***Abstract*—**This paper proposes a method of learning topics for broadcasting contents. There are two kinds of texts related to broadcasting contents. One is a broadcasting script, which is a series of texts including directions and dialogues. The other is blogposts, which possesses relatively abstracted contents, stories, and diverse information of broadcasting contents. Although two texts range over similar broadcasting contents, words in blogposts and broadcasting script are different. When unseen words appear, it needs a method to reflect to existing topic. In this paper, we introduce a semantic vocabulary expansion method to reflect unseen words. We expand topics of the broadcasting script by incorporating the words in blogposts. Each word in blogposts is added to the most semantically correlated topics. We use word2vec to get the semantic correlation between words in blogposts and topics of scripts. The vocabularies of topics are updated and then posterior inference is performed to rearrange the topics. In experiments, we verified that the proposed method can discover more salient topics for broadcasting contents.

***Keywords*—**Broadcasting script analysis, topic expansion, semantic correlation analysis, word2vec.

I. Introduction

BROADCASTING contents are broadcasted through TV or internet. It can provide interesting information to people such as recommendation of TV program or merchandise by analyzing broadcasting contents. Topic model is one of the methods, which discovers topics in documents. If similar subjects such as cast actor, product information, cluster topics discovered from broadcasting contents and so on, these topics help to find useful information of broadcasting contents.

There are two types of text, which are used to analyze broadcasting contents – blogposts and a broadcasting script. Blogposts are written by viewer and contain various information such as cast actors, merchandise, and so on. A broadcasting script is composed of detailed texts like directions and dialogues. Since these two texts include different aspects of broadcasting contents, it is required to using both a broadcasting script and blogposts for topic learning.

When topics are discovered by using a broadcasting script and blogposts, two points should be considered. First, blogposts are made consistently. It is inefficient to learn topics whenever new blogposts are generated. Therefore, it is required to a method for process continuously. Second, word-distribution between blogposts and a broadcasting script are different each other. Thus, unseen words appeared when each blogpost is added. Because of inconsistency in word distribution, it is difficult to obtain sufficient statistics of inter correlation of unseen words. These characteristics cause that topic model is hard to discover topic when blogposts and a broadcasting script are learned at the same time. To solve these problems, this paper proposes a method of topic learning for two types of texts. A broadcasting script is representing detailed storylines. Therefore, topic model using a broadcasting script do the role for a guide to learn blogposts. Thus, blogposts are learnt after building a broadcasting script topic. In order to reflect blogposts continuously, online learning is adopted. Blogposts are basically added one by one. That is, we learn latent topic by incorporating each blogposts into existing topic. Second, in learning process, unseen words, which are not contained in existing topics, appeared from blogposts. When unseen words are added for topic learning, it is hard to obtain relation with existing topic. In order to reflect unseen words semantically, we use Word2vec [1]. Using Word2vec it is possible to consider unseen words in learning process. Word2vec, one of popular semantic representation of words, is exploited to compute correlation between a word from blogposts and topics of a broadcasting script. By averaging similarities between a word to be added and top-k words in each topic of scripts, we can obtain the topical correlation scores. When the biased vocabulary expansion is done, the posterior distribution can be inferred by stochastic gradient optimization [2].

The proposed method is evaluated with LDA model in various conditions of broadcasting contents document. The experimental results show that the proposed method outperforms all the compared models, where the performance of topic coherence is measured by PMI and Word2Vec. These results prove that the proposed method is more plausible in representing broadcasting contents topic by using a broadcasting script and blogposts.

II. Related Works

Document modeling finds the method of representing of document corpus. If document modeling is achieved, it is easy to grasp of contents of documents and be used for various applications. Topic model is widely used in document modeling. Latent Dirichlet Allocation (LDA) [3] and Hierarchical Dirichlet process (HDP) [4] is well-known topic model. Topic models have been used to discover latent topic in document. Generally, the number of topic is less than the number of words in documents; it can be efficient tools for

Chang-Uk Kwak and Sang-Jo Lee are with School of Computer Science and Engineering, Kyungpook National University, Daegu 41566, Korea (e-mail: cukwak@sejong.knu.ac.kr, sjlee@sejong.knu.ac.kr).

Seong-Bae Park is with School of Computer Science and Engineering, Kyungpook National University, Daegu 41566, Korea (corresponding author phone: +82 53 950 7574; e-mail: sbpark@sejong.knu.ac.kr).

Sun-Joong Kim is with Smart Media Platform Section, ETRI, Daejeon 34129, Korea (e-mail: kimsj@etri.re.kr).

mapping high dimensional document corpus to low dimensional vector space. Topic model can obtain topic, which is consisted of similar words from low dimension vector space.

Recently, topic model was applied to online learning [5], [6]. The motivation of online topic model is to analyze large data set. In addition, online topic models also can be used to analyze streams of data. However, existing online topic model must have predefined vocabulary for discovering topic. To solve this problem, [2] proposed a method of online LDA with infinite vocabulary. They suggested that when new data are added to existing topic, topic model updates own vocabulary by reflecting unseen words.

There were some studies to analyze broadcasting contents by using topic model. Misra et al. [7] proposed a method of scene segmentation on TV news using LDA. In this study, transcript is used for scene segmentation. Thus, topic model can be obtained story boundaries and topic distribution each scene. Engels et al. [8] proposed a method of automatic annotation of location in video. In this study, topic model is used to find location information from topic distribution, and transcript is also used. These two studies use texts, which contain broadcasting contents. However, because of these studies focus on internal information of broadcasting content, their methods cannot obtain topics which are related broadcasting contents likes product information.

III. Topic Model for Broadcasting Contents Using Semantic Correlation Information

A. Topic Learning Using Broadcasting Script

In order to discover topic from a broadcasting script, we use a Latent Dirichlet Allocation (LDA). As LDA is a generative model for documents, it is widely used for document analysis. Given corpus of documents in *D*, LDA discovers *k*-topics after training. The generative process is as follow:

1. **for** each document d **in** D **do**:
2. Choose a $\theta_d \sim Dirichlet(\alpha)$
3. **for** each word indexed n = 1, … N **in** d **do**:
4. Choose a topic $z_n \sim Multinomial(\theta_d)$
5. Choose a word $p(w_n|z_n, \beta)$

If this process was used to our topic learning from broadcasting script *D*, we discover *k*-topics about broadcasting contents. Due to topic model is considered co-occurrence of words in documents, we discover topic, which based on story of broadcasting contents.

B. Topic Expansion Using Blogposts

When we use blogposts for topic expansion, all of blogposts are sequentially processed. That is, blogposts are added one by one. In this condition, unseen words occur necessarily when blogposts added in learning process. Because existing topic is hard to obtain sufficient statistics of unseen words, a method of reflecting unseen word was needed.

For this, we allocate unseen words to similar topic. Word2vec learns semantic vector representations of words from the large corpus; we can obtain semantic correlation between two words. In order to obtain the semantic correlation value, we calculate cosine similarity between two vectors by

$$cor(\rho_1, \rho_2) = \frac{\overrightarrow{wv_{\rho_1}} \cdot \overrightarrow{wv_{\rho_2}}}{\|\overrightarrow{wv_{\rho_1}}\| \|\overrightarrow{wv_{\rho_2}}\|} \quad (1)$$

where $\overrightarrow{wv_{\rho_1}}$ and $\overrightarrow{wv_{\rho_2}}$ are the vector of ρ_1 and ρ_2. In order to allocate unseen word ρ, topical similarity between a word ρ and topic k is calculated in (2). Topical correlation score is calculated by

$$s_k(\rho) = \frac{1}{n} \sum_{i=1}^{n} cor(\rho, t_{ki}) \quad (2)$$

where t_{ki} denotes the top-*i*th word in topic *k*. $s_k(\rho)$ represent a topical correlation score between topic k and word ρ. score $s_k(\rho)$ obtains by averaging $cor(\rho, t_{ki})$ in regard to top-n words. We allocate unseen word to the highest value topic of topical correlation score.

After unseen words are added to topic, we need to update topic model, which is reflected blogposts. To do this, we perform variational inference for conditional distribution of topic given words [2]. Therefore, the conditional distribution of z_{dn} is as:

$$q(z_{dn} = k | \boldsymbol{z}_{-dn}, w_{dn}) \propto \left(\sum_{\substack{m=1 \\ m \neq 1}}^{N_d} \prod_{z_{dm}=k} + \alpha_k \right) \exp\{\mathbb{E}_{q(v)}[\log \beta_{kt}]\} \quad (3)$$

where $\boldsymbol{z}_{-dn}$ denotes words except w_{dn}. q is a distribution. For document *d* contains N_d words, there are two cases to calculate (2). First, if a word w_{dn} is existed in topic *k*, the conditional distribution of a word is computed by:

Algorithm 1: Topic Expansion using Blogposts

Input : A broadcasting script set D_s,
Blogposts sets $D_b^1, D_b^2, \ldots, D_b^N$
Output: Broadcasting contents topic T_k

Learning initial topic T_k from D_s.

for *n* = 1, …, *N* **do**
  Select words sets ρ from D_b^n.
  for *t* = 1, …, *T* **do**
    if word ρ_t is unseen word
      for *k* = 1, …, *K* **do**
        Calculate topical correlation $S_k(\rho_t)$ using (2).
      end
      Allocate ρ_t to highest value topic of $S_k(\rho_t)$.
    end if
  end for
  Update topic using (3).
end for
Return topic T_k

$$q(z_{dn} = k) \propto \left(\sum_{\substack{m=1 \\ m \neq 1}}^{N_d} \phi_{dmk} + \alpha_k \right) \cdot \exp\{\Psi(\nu_{kw}^1) + \sum_{s=1}^{s \leq w} \Psi(\nu_{ks}^2) - \sum_{s=1}^{s \leq w} \Psi(\nu_{ks}^1 + \nu_{ks}^2)\} \quad (4)$$

Second, for reflecting of unseen words into topic, we calculate distribution of unseen words in topic. If a word w_{dn} is

not exist in topic k, the conditional distribution of an unseen word is computed by:

$$q(z_{dn}=k) \propto \left(\sum_{\substack{m=1 \\ m\neq 1}}^{N_d} \phi_{dmk} + \alpha_k\right) \cdot \exp\left\{\sum_{s=1}^{s\leq w}\left(\Psi\left(\nu_{ks}^2\right) - \Psi\left(\nu_{ks}^1 + \nu_{ks}^2\right)\right)\right\} \quad (5)$$

The different between (4) and (5), reflecting unseen words or not. This second process has repeated until all blogposts were processed.

Algorithm 1 explains the proposed method in detail. This algorithm takes a broadcasting script D_s and blogposts sets $\{D_b^1, D_b^2, \ldots, D_b^N\}$ as input, and returns expanded topic T_k. At first, initial topic learning is performed using a broadcasting script D_s. After that, at each n-th blogposts, word selection is performed for every blogposts D_b^n. For each word ρ_t in blogposts D_b^n, if a word ρ_t is a unseen word, vocabulary expansion is needed for allocating a word ρ_t. For each topic k, topical correlation similarity between ρ_t and T_k is calculated by (2). In this calculation, top-i words in existing topic T_k are used for calculating with ρ_t. Unseen word ρ_t is allocated to the highest value topic of $S_k(\rho_t)$. After finishing expansion of vocabularies, topic is updated using (3).

IV. Experiments

A. Experiment Setting

For the experiments, two types of data set are used. First, Korean drama script which is 'I heard it by the grapevine, 6th inning' is used for initial topic learning. Second, blogposts are collected from web. It was crawled by using given keywords, which is name of drama and cast actor during 1 week after broadcasting. For topic learning, a broadcasting script is considered only dialogues. All of documents are extracted noun, and remove stopwords.

Table I shows a statistics of data sets. In case of blogposts, the total documents count is 72. The number of unique terms in a broadcasting and blogposts is 303 and 1,206. The number of duplicate words between two texts is 124. In broadcasting script, percent of duplicate terms is 24%, which is contrasted with blogposts. Here, we confirm that the word distribution between two texts is different considerably.

The proposed model is compared with LDA^{script}, LDA^{blog} and LDA^{all}. LDA^{script} and LDA^{blog} used only a broadcasting script and blogposts, respectively. LDA^{all} uses both a broadcasting script and blogposts at the same time.

Derek et al. [9] measured the performance of topic model by using coherence. We used TC-PMI, TC-W2V for measuring coherence given topic K. In each topic, top-N words are consider to measure coherence. TC-PMI measures topic coherence by using Pointwise Mutual Information (PMI). PMI is a measure of how much co-occurrence between two words [10]. The intuition of using PMI is that many co-occurrences between two words is likely to relate coherent topic. TC-PMI is measured by

$$\text{TC-PMI} = \frac{1}{N \cdot K}\sum_{k=1}^{K}\sum_{j=2}^{N}\sum_{i=1}^{j-1}\log\frac{p(w_{ki}, w_{kj})}{p(w_{ki}) \cdot p(w_{kj})}$$

In TC-PMI, the probability $p(w_{ki})$ and $p(w_{ki}, w_{kj})$ are given by

$$p(w_{ki}) = \frac{\#docs(w_{ki})}{\#docs(\cdot)}$$

$$p(w_{ki}, w_{kj}) = \frac{\#docs(w_{ki}, w_{kj})}{\#docs(\cdot)}$$

where $\#docs(\cdot)$ is the total number of documents and $\#docs(w_{ki})$ is the number of document that contains w_{ki}. $\#docs(w_{ki}, w_{kj})$ is the number of document that contains both w_{ki} and w_{kj}.

TABLE I
Statistics of Data Sets

	# of documents	# of unique terms	# of terms	Duplicate terms
A Broadcasting Script	1	303	498	124
Blogposts	79	1517	5296	

TC-W2V measures topic coherence by using Word2Vec. The intuition of using Word2Vec is that whether discovered topic is built semantically or not. TC-W2V is computed by

$$\text{TC-W2V} = \frac{1}{N \cdot K}\sum_{k=1}^{K}\sum_{j=2}^{N}\sum_{i=1}^{j-1} cor(w_{kj}, w_{ki})$$

For TC-W2V, word2vec score between w_{kj} and w_{ki} in topic k is calculated by cosine similarity in (2).

PMI is calculated from all of data, which are using in this experiment. Word2Vec is built with 100 dimensions from Korean Wikipedia dump. The number of topics K is 30 and N is 30.

B. Experimental Results and Discussion

Table II shows the result of topic coherence. As shown in this table, topics of the proposed method are more coherent compared to those of baselines. When learning topic only using each broadcasting contents data, the performance of topic coherence has lower than the proposed method. This means that the proposed method represents high quality of topic.

In addition, LDA^{all} have the lowest coherence value in TC-PMI. This result means that topic model is hard to discover topic when learning two data at the same time. LDA^{all}, LDA^{blog} and LDA^{scene} are processed under not online condition. As shown in Table II, all the compare model has lower value than proposed method. This means that online learning is better to represent broadcasting contents when using a broadcasting script and blogposts than LDA.

Another important thing is that unseen words allocate to similar topics when blogposts added to existing topic for topic learning. In TC-W2V results, all the LDA model has less coherent compared with proposed method. This result proves

that the proposed method reflects unseen words semantically.

TABLE II
THE RESULT OF TOPIC COHERENCE

	TC-PMI	TC-W2V
LDA^{all}	3.2666	0.0436
LDA^{blog}	3.2953	0.0434
LDA^{script}	3.6371	0.0378
Proposed method	**3.8954**	**0.0533**

V. CONCLUSION

This paper has proposed a method of topic model for broadcasting contents using semantic correlation information. In our work, for making salient topics with regard to broadcasting contents, we represent topic, which are obtained from both a broadcasting script and blogposts. The main advantage of our method is to reflecting semantically when unseen words are added in online learning process. The evaluation of the proposed method was measured by topical coherence. The performance of this model is outstanding compared with other LDA model where condition of various data. It proves that the proposed method can be to discover salient topics of broadcasting contents.

ACKNOWLEDGMENT

This study was supported by the BK21 Plus project (SW Human Resource Development Program for Supporting Smart Life) funded by the Ministry of Education, School of Computer Science and Engineering, Kyungpook National University, Korea (21A20131600005) and 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] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 2013.
[2] K. Zhai, and J. Boyd-Graber. "Online Latent Dirichlet Allocation with Infinite Vocabulary." In *Proceedings of The 30th International Conference on Machine Learning*, pp. 561-569, 2013.
[3] 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.
[4] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical dirichlet processes." *The American statistical association*, 2006.
[5] M. Hoffman, F. R. Bach, and D. M. Blei, "Online learning for latent dirichlet allocation," *Advances in neural information processing systems*, pp. 856-864, 2010.
[6] C. Wang, J. W. Paisley, and D. M. Blei, "Online variational inference for the hierarchical Dirichlet process," In *Proceedings of International Conference on Artificial Intelligence and Statistics*, pp. 752-760, 2011.
[7] H. Misra, F. Hopfgartner, A. Goyal, P.Punitha, and J. M. Mose, "TV news story segmentation based on semantic coherence and content similarity." *Advances in Multimedia Modeling*, pp. 347-357. 2010.
[8] C. Engels, K. Deschacht, J. H. Becker, T. Tuytleaars, M-F. Moens, and L. V. Gool, "Automatic annotation of unique locations from video and text," *BMVC*, pp 1-11, 2010.
[9] D. O'Callaghan, D. Greene, J. Carthy, and P. Cunningham, "An analysis of the coherence of descriptors in topic modeling". *Expert Systems with Applications*, Vol. 42(13), pp. 5645-5657. 2015.
[10] G. Bouma. "Normalized (pointwise) mutual information in collocation extraction." In *Proceedings of GSCL*, pp. 31-40, 2009.