2021 18th International Joint Conference on Computer Science and Software Engineering (JCSSE) | 978-1-6654-3831-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/JCSSE53117.2021.9493816

Topic Modeling Enhancement using Word Embeddings

Siriwat Limwattana
Department of Computer Engineering
King Mongkut's University of Technology Thonburi
Bangkok, Thailand
Email: siriwat.grad@mail.kmutt.ac.th

Santitham Prom-on
Department of Computer Engineering
King Mongkut's University of Technology Thonburi
Bangkok, Thailand
Email: santitham.pro@mail.kmutt.ac.th

***Abstract*—Latent Dirichlet Allocation(LDA) is one of the powerful techniques in extracting topics from a document. The original LDA takes the Bag-of-Word representation as the input and produces topic distributions in documents as output. The drawback of Bag-of-Word is that it represents each word with a plain one-hot encoding which does not encode the word level information. Later research in Natural Language Processing(NLP) demonstrate that word embeddings technique such as Skip-gram model can provide a good representation in capturing the relationship and semantic information between words. In recent studies, many NLP tasks could gain better performance by applying the word embedding as the representation of words. In this paper, we propose Deep Word-Topic Latent Dirichlet Allocation(DWT-LDA), a new process for training LDA with word embedding. A neural network with word embedding is applied to the Collapsed Gibbs Sampling process as another choice for word topic assignment. To quantitatively evaluate our model, the topic coherence framework and topic diversity are the metrics used to compare between our approach and the original LDA. The experimental result shows that our method generates more coherent and diverse topics.**

Keywords*—Topic Modeling, Latent Dirichlet Allocation, Word Embedding

I. Introduction

With the rise of the internet, textual information is growing rapidly. People are expressing their thought mostly on text, which becomes a priceless data source. Therefore, the study on Natural Language Processing(NLP) is crucial for extracting insights from the unstructured textual data. In this regard, topic modeling becomes an active area of research that focuses on understanding and grouping massive documents automatically without manual annotation.

Latent Dirichlet Allocation is a widely used topic algorithm, which was proposed by Blei et al. in 2003 [1]. LDA is a generative model that learns the representation of documents from frequencies of each word used in that document using the Bag-of-Word representation. It treats one word apart from each other although some of them may closely related or be a synonym. After an efficient word embedding technique was proposed by Mokolov et al. in 2013 [2], skip-gram with negative sampling becomes a powerful word representation that uses less training time and able to embed the semantic and syntactic information of words into embedding space [3]. This technique then has been successfully applied to various NLP applications. In 2015, Nguyen et al. [4] proposed a latent feature topic model(LF-LDA) that applies word embedding to topic-word inference process. The process is improved by creating latent features for each topic to be used with word embedding with a Bernoulli distribution that controls the sampling process. Recently, Dieng et al. [5] proposed a novel Embedded Topic Model(ETM) that constructs the topic embedding from the word embedding space, and generates the document topic mixture using the variational autoencoder model. This model allows the word embedding to be learned during the topic modeling training process, and also allows to use a pre-trained embedding. The experiment shows that this model was more robust to stop words when using a pre-trained word embedding compared to training the word embedding during the topic modeling inference process. According to the aforementioned studies, applying word embedding to the topic modeling increases performance over the existing LDA.

This paper proposes a new topic modeling approach that focuses on topic modeling enhancement with a pre-trained word embedding. The core concept of Deep Word-Topic Latent Dirichlet Allocation(DWT-LDA) is to apply the neural network that uses a pre-trained word embedding to learn the topic assignment from the embedding space. The network is then used as an alternative word topic assignment during the inference process. To evaluate the model, the original LDA is compared with our method on both Thai and English datasets.

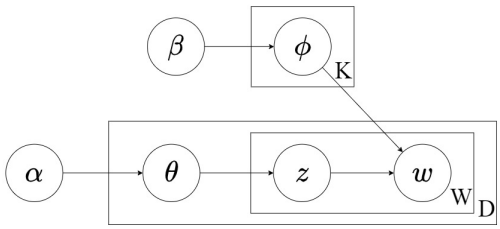

Figure 1. The graphical model of Latent Dirichlet Allocation

The experimental result indicates that our method success in generating a qualitative improvement to the original LDA.

II. Background

A. Latent Dirichlet Allocation

Latent Dirichlet Allocation(LDA) is a generative model [1] that assumes each document contains a distribution over k topic θ with a Dirichlet prior α. Each word w in a document is randomly drawn from each topic $\phi_{1..k}$ with a Dirichlet prior β. For the graphical model of LDA, see figure 1. The generative process for a document is as follows:

For each word w_{ij} in document D_j

1. Draw topic $z_{ij} \sim$ Multinomial(θ_j)
2. Draw word $w_{ij} \sim$ Multinomial($\phi_{z_{ij}}$)

One of the widely used approximation technique is Collapsed Gibbs Sampling [6] that instead of inferencing all posterior distributions, it directly samples the topic z to the observed word w to estimate θ and ϕ as shown in Alogorithm 1. The approximation of ϕ of word w and topic k is shown on Eq. 1, θ of topic k of j^{th} document is shown on Eq. 2 and the probability of topic to be assigned is shown on Eq. 3.

$$\phi_{wk} = \frac{N_{wk} + \beta}{N_k + W\beta} \tag{1}$$

$$\theta_{kj} = \frac{N_{kj} + \alpha}{N_j + K\alpha} \tag{2}$$

$$P(z = k|w, \alpha, \beta) = \frac{\phi_{wk} * \theta_{kj}}{\sum_{k=1}^{K} \phi_{wk} * \theta_{kj}} \tag{3}$$

B. Skip-gram

Skip-gram is a representation of words on lower dimensional space trained by a shallow network which takes a word as input to predict the context word [2]. This technique performed well on capturing both semantic and syntactic between words. By adopting negative sampling method on skip-gram, the training process of skip-gram with negative sampling is to make a shallow network to distinguish between the context word and random sampling. It shows an improvement on speed of model training that performs a logistic regression instead of a softmax layer. It also improves the quality of vector especially for rare words [3]. Therefore, it has become a standard practice for word representation in a current research in NLP.

Algorithm 1 Collapsed Gibbs Sampling For Latent Dirichlet Allocation

Require: a corpus of document D
Require: i^{th} word in the document W_{di}
Require: number of word in each document N_d
Require: frequency of each word in each document N_{di}
Require: frequency of topic assigned to each document N_{kj}
Require: frequency of word assigned to each topic N_{wk}
Require: topic assignment of each word z_{dij}

```
for d in 1... D do
    for i in 1... N_d do
        w = W_di
        for j in 1... N_di do
            k̂ = z_dij
            N_wk̂ = N_wk̂ − 1
            N_dk̂ = N_dk̂ − 1
            k̂ ~ Multinomial(P(z|z^−ij, w, α, β))
            z_dij = k̂
            N_wk̂ = N_wk̂ + 1
            N_dk̂ = N_dk̂ + 1
        end for
    end for
end for
```

III. Dataset

A. Pantip

Pantip.com is the most popular online discussion platform in Thailand. On Pantip, people are discussing or sharing a wide range of topics (e.g., science, finance, traveling, loves, religions, cooking, sports and dramas). We collected the data from the first post of each thread across all forums in total of 58304 posts, and the final vocabulary size of 25847 words. The average length of the documents was 392 words and median of 213 words. Each document was pre-processed as follows:

1) Eliminate HTML tags.
2) Number masking.
3) Normalize the character [7]. e.g, replacing double เ (เเ) with แ, and ◌ำ and า with ◌ำา.
4) Token segmentation using Attacut [8].
5) Remove stop words.
6) Remove common words by filtering out word with document frequency ratio greater than 0.9.
7) Remove rare words by filtering out word with document frequency less than 10 documents.

B. Amazon

For the dataset of customer reviews on Amazon.com [9], we sampled 40000 reviews from the product in various categories including beauty, clothing shoes and jewelry, grocery and gourmet food, sports and outdoor, and video games. After they were pre-processed, the final vocabulary size was 15824 words. The average length of the documents was 260 words and median of 178 words. Each review was pre-processed as follows:

1) Token segmentation using NLTK [10].
2) Lowering all characters.
3) Remove stop words.
4) Remove common words by filtering out word with document frequency ratio greater than 0.9.
5) Remove rare words by filtering out word with document frequency less than 10 documents.

IV. Methodology

This section will describe the Deep Word-Topic Latent Dirichlet Allocation(DWT-LDA) model architecture, training procedures, and the evaluation method used to compare our model with the existing LDA.

A. Model

DWT-LDA is designed to allow LDA to gain a contextual knowledge of words by applying information from word embedding, therefore the model is divided into 2 steps. The first step is a standard LDA using Collapsed Gibbs Sampling method aiming to initialize the topics, as shown in Algorithm 1. The second step is designed to enhance the learned topic by applying the knowledge of words from a pre-trained word embedding. A neural network, as shown in Figure 2, trained on the topic assignment of each word is placed on the sampling process of topic P($z|w$). It is designed to learn the association between the topics from LDA and the contextual information, which the embedding layer allows the model to learn the topic assignment based on the embedding vector of the word. Since similar words are likely to be close to each other on the embedding space, it increases the chance of being predicted in the same topic that the input of the network will be closely similar to each other. To maintain a dirichlet prior, β is added to the predicted probability of the network. A Bernoulli distribution λ is added as a switching between topic assignment P($z|w$) from using ϕ and neural network to avoid repeatedly using all of its own predictions as the training data for later iteration. Therefore, the word-topic agreement ϕ_{wk} on Eq. 3 is changed to $\hat{\phi}_{wk}$ on Eq. 5.

$$\lambda_{ij} \sim Bernoulli(\lambda) \tag{4}$$

```
Model: "topic_assignment"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
word (Embedding)             (None, None, 300)         7376100
_________________________________________________________________
dense1 (Dense)               (None, None, 1000)        301000
_________________________________________________________________
dropout1 (Dropout)           (None, None, 1000)        0
_________________________________________________________________
dense2 (Dense)               (None, None, 1000)        1001000
_________________________________________________________________
dropout2 (Dropout)           (None, None, 1000)        0
_________________________________________________________________
topic (Dense)                (None, None, 30)          30030
=================================================================
Total params: 8,708,130
Trainable params: 1,332,030
Non-trainable params: 7,376,100
_________________________________________________________________
```

Figure 2. The architecture of neural network based word-topic assignment

$$\hat{\phi}_{wk} = (1 - \lambda_{ij}) * \frac{N_{wk} + \beta}{N_k + W\beta} + \lambda_{ij} * \frac{NN(w)_k + \beta}{K\beta + \sum NN(w)} \tag{5}$$

The training process starts with training the original LDA with Collapsed Gibbs Sampling. After the model is converged, the topic assignments of each word are being used as the training set for the neural network. For initialization purpose, the network is trained for some iterations. During the training process of the second step, the network is being fit for several steps with the topic assignment from the previous iteration.

B. Evaluation

To compare the result of our model, all variables from the first stage were duplicated into 2 copies. That enabled us to train the original LDA along with our model with the same conditions as illustrated in figure 3. Two metrics were used to qualitatively compare the model including C_v score from the topic coherence framework [11], and topic diversity which was proposed in [5].

1) Topic Coherence: Topic coherence(C_v) is a framework for a qualitative evaluation of a topic model. C_v is the most correlates with human rating [11]. It is computed by considering the agreement of top words on each topic and the sliding-window.

2) Topic Diversity: Topic diversity is a qualitative measure of redundancies between topics. It measures the ratio of unique words from the top 25 words to all top words [5]. The high value indicates that the same word did not appear across topics which reduce the ambiguity on interpreting the topic.

V. Experimental Result

To conduct the experiment on both dataset, we run the algorithm with k=5 to k=180 where β, α and λ were empirically set to $\frac{1}{k}$, $\frac{1}{k}$ and 0.6. The pre-trained embedding for

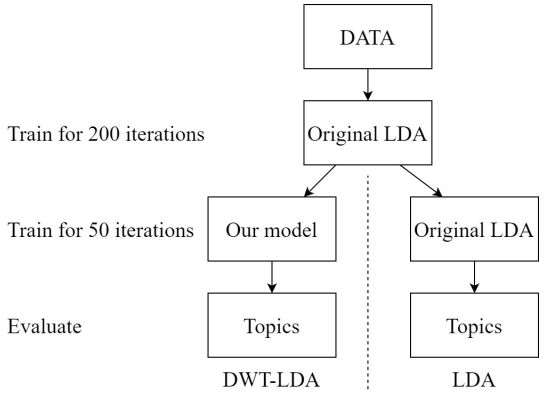

Figure 3. The framework for evaluating the model

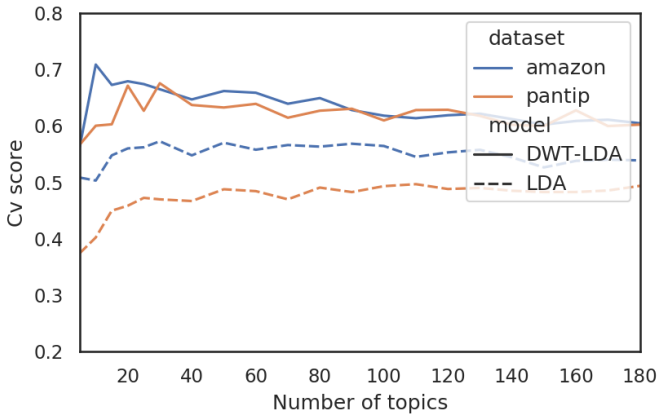

Figure 4. The topic coherence score on Amazon and Pantip dataset

Thai dataset is our self-trained skip-gram model with a very large corpus across multiple domains (e.g., Wikipedia, news, constitutional and etc.), and a pre-trained word embedding on Google News corpus published by Google Open Source for the English dataset. Throughout the experiment, we used top 20 words from each topic to compute the topic coherence, and top 25 words to compute the topic diversity.

The qualitative evaluation shows that our method outperforms the existing LDA. Figure 4 shows that our model archives the topic coherence score of 0.62 on Pantip dataset comparing to the original LDA of 0.47, and 0.64 on Amazon dataset compared to 0.55. Figure 5 shows that there is a gradual decline in the score of topic quality on the increasing number of topics. The average difference between DWT-LDA and LDA is about 0.28 on Pantip, and 0.24 on Amazon dataset. The outcome shows that our method gives more specific keywords to the topics that some topics from LDA are ambiguous while DWT-LDA gives more concrete keywords, e.g., romantic relationship of Table I, and makeup on Table II. Both tables show that DWT-LDA also generates more diverse keywords on topics that frequent word such as "ค น" and "good" are omitted from the top keyword.

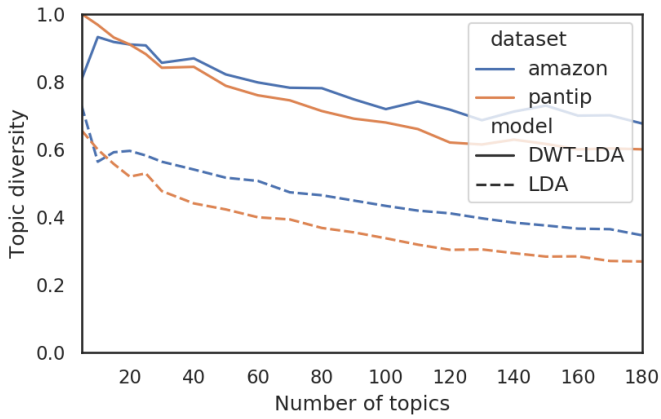

Figure 5. The topic diversity score on Amazon and Pantip dataset

VI. Discussion and Conclusion

In this paper, we propose Deep Word-Topic Latent Dirichlet Allocation(DWT-LDA), an upgraded LDA model by using word information. The word embedding allows the model to learn the topic assignment with the help of information learned by the context from the larger corpus. The word-topic assignment is done by incorporating with embedding dimensional information to assign the topic, therefore the model learns the topics from the latent information of words instead of considering the close or related word apart. However, our method highly depends on the original LDA on the early stage. Therefore, there might be a limitation that this method may perform well if the original LDA could capture good topics.

Our approach shows an improvement to the traditional LDA which yields better topic coherence score and generates more diverse topics on both Thai and English datasets. Although the variety of topics generated on both algorithm were closely related, DWT-LDA is able to generate more specific keywords of each topic that makes a clearer topic annotation. This is another evidence that the LDA could be improved by applying word information to generate more meaningful topics.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, p. 993–1022, Mar. 2003.

[2] T. Mikolov, K. Chen *et al.*, "Efficient estimation of word representations in vector space," in *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2013. [Online]. Available: http://arxiv.org/abs/1301.3781

[3] T. Mikolov, I. Sutskever *et al.*, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*,

Table I. EXAMPLE OF DISCOVERED TOPICS FROM LDA AND DWT-LDA ON PANTIP DATASET

Travel		**Politics**		**Restaurant**		**Buddhism**		**Treatment**		**Finance**		**Relationship**	
LDA	DWT-LDA	LDA	DWT-LDA	LDA	DWT-LDA	LDA	DWT-LDA	LDA	DWT-LDA	LDA	DWT-LDA	LDA	DWT-LDA
เดิน	เที่ยว	คน	พรรค	ร้าน	ร้าน	ท่าน	จิต	หมอ	หมอ	เงิน	เงิน	ผม	ผม
คน	เดินทาง	ทำ	สส	อาหาร	อร่อย	ทำ	ธรรม	ยา	ยา	เดือน	บัตร	คน	แฟน
รถ	รถไฟ	พรรค	รัฐบาล	กิน	ทาน	คน	ท่าน	อาการ	อาการ	ค่า	เดือน	ทำ	คุย
เที่ยว	สถานี	ไทย	เลือกตั้ง	น้ำ	ใส่	โลก	พระ	ทำ	โรค	ทำ	จ่าย	ตอน	คับ
เดินทาง	เมือง	ปี	ประชาชน	ทำ	อาหาร	รู้	ศาสนา	กิน	เลือด	บาท	บาท	รู้	คบ
พัก	บิน	ประเทศ	นายก	ทาน	เนื้อ	ตัว	พระพุทธเจ้า	ตอน	ตรวจ	ซื้อ	แจ้ง	ตัว	เลิก
เวลา	ตั๋ว	เมือง	รัฐมนตรี	ใส่	รสชาติ	ชีวิต	กรรม	ตัว	รักษา	จ่าย	ธนาคาร	เรื่อง	หัก
นั่ง	ทริป	เรื่อง	เมือง	อร่อย	หมู	พระ	วัด	เดือน	ปวด	ขาย	โทร	ดี	ทะเลาะ
ถ่าย	จอง	รัฐบาล	รัฐธรรมนูญ	ดี	หวาน	ใต	ทุกข์	โรค	พยาบาล	แจ้ง	หุ้น	ถาม	ก้
รูป	รถ	ข่าว	คะแนน	ข้าว	น้ำ	สร้าง	บุญ	รักษา	ป่วย	ปี	ประกัน	เหมือน	จีบ

Table II. EXAMPLE OF DISCOVERED TOPICS FROM LDA AND DWT-LDA ON AMAZON DATASET

Orders		**Skin Care**		**Game Console**		**Bike**		**Clothing**		**Makeup**	
LDA	DWT-LDA	LDA	DWT-LDA	LDA	DWT-LDA	LDA	DWT-LDA	LDA	DWT-LDA	LDA	DWT-LDA
amazon	amazon	skin	skin	game	controller	bike	bike	size	shoes	color	brush
price	received	product	face	controller	wii	rack	rack	wear	wear	watch	color
buy	seller	face	product	xbox	xbox	easy	tire	fit	size	brush	polish
product	item	oil	cream	play	ps3	good	seat	shoes	fit	great	nail
review	shipping	dry	oil	ps3	game	work	road	comfortable	pair	love	nails
time	service	cream	lotion	wii	console	ride	lock	pair	comfortable	good	mascara
bought	customer	feel	moisturizer	console	nintendo	seat	ride	great	feet	nail	lashes
money	ordered	smell	dry	great	psp	wheel	chain	feet	socks	time	colors
good	arrived	good	acne	version	sony	road	pump	shoe	bra	polish	coat
quality	shipped	lotion	wash	screen	remote	lock	rear	good	foot	black	makeup

C. J. C. Burges, L. Bottou *et al.*, Eds., vol. 26. Curran Associates, Inc., 2013.

[4] D. Q. Nguyen, R. Billingsley *et al.*, "Improving topic models with latent feature word representations," *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 299–313, 2015. [Online]. Available: https://www.aclweb.org/anthology/Q15-1022

[5] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, "Topic Modeling in Embedding Spaces," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 439–453, 07 2020. [Online]. Available: https://doi.org/10.1162/tacl_a_00325

[6] I. Porteous, D. Newman *et al.*, "Fast collapsed gibbs sampling for latent dirichlet allocation," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 569–577. [Online]. Available: https://doi.org/10.1145/1401890.1401960

[7] W. Phatthiyaphaibun, A. Suriyawongkul *et al.*, "Pythainlp/pythainlp: Pythainlp 2.2.5," Nov. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.4275722

[8] P. Chormai, P. Prasertsom, and A. Rutherford, "Attacut: A fast and accurate neural thai word segmenter," *CoRR*, vol. abs/1911.07056, 2019. [Online]. Available: http://arxiv.org/abs/1911.07056

[9] R. He and J. McAuley, "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering," in *Proceedings of the 25th International Conference on World Wide Web*, ser. WWW '16. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2016, p. 507–517. [Online]. Available: https://doi.org/10.1145/2872427.2883037

[10] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, 1st ed. O'Reilly Media, Inc., 2009.

[11] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, ser. WSDM '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 399–408. [Online]. Available: https://doi.org/10.1145/2684822.2685324