PhraseCTM: Correlated Topic Modeling on Phrases within Markov Random Fields

Weijing Huang¹, Tengjiao Wang¹, Wei Chen¹, Siyuan Jiang², Kam-Fai Wong¹,³
¹ Key Lab of High Confidence Software Technologies (MOE), School of EECS, Peking University, Beijing, China
² Department of Computer Science and Engineering, University of Notre Dame, USA
³ Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong
{huangweijing, tjwang, pekingchenwei}@pku.edu.cn, sjiang1@nd.edu, kfwong@se.cuhk.edu.hk

Abstract

Recent emerged phrase-level topic models are able to provide topics of phrases, which are easy to read for humans. But these models are lack of the ability to capture the correlation structure among the discovered numerous topics. We propose a novel topic model PhraseCTM and a two-stage method to find out the correlated topics at phrase level. In the first stage, we train PhraseCTM, which models the generation of words and phrases simultaneously by linking the phrases and component words within Markov Random Fields when they are semantically coherent. In the second stage, we generate the correlation of topics from PhraseCTM. We evaluate our method by a quantitative experiment and a human study, showing the correlated topic modeling on phrases is a good and practical way to interpret the underlying themes of a corpus.

1 Introduction

In recent years, topic modeling on phrases has been developed for providing more interpretable topics (El-Kishky et al., 2014; Kawamae, 2014; He, 2016). They represent each topic as a list of phrases, which are easy to read for humans. For example, the topic represented in “grounding conductor, grounding wire, aluminum wiring, neutral ground, ...” is easier to read than the topic with words “ground, wire, use, power, cable, wires, ...”, although they are both about the topic of household electricity.

But when the number of topics grows, it’s hard to review all the topics, even they are represented in phrases. The correlation structure is introduced by CTM (Blei and Lafferty, 2005) to figure out the correlated relationship between topics and group the similar topics together. And the correlated topics mined from the scientific papers (Blei and Lafferty, 2007), news corpus (He et al., 2017), and social science data (Roberts et al., 2016), showed their practical utility on grasping the semantic meaning of text documents.

However, it’s nontrivial to apply CTM directly on phrases. The reasons are mainly due to two facts: (1) phrases are much less than words in each document; (2) similar to LDA (Tang et al., 2014), CTM doesn’t perform well on short documents. Therefore, CTM needs more contextual information to build a good enough model, rather than only using the extracted phrases.

To find out the correlated topics at phrase level, we take full advantage of contextual information about the phrases. Firstly, the topic of a phrase in a document is highly related to the topics of other words and phrases in the same document. Secondly, some phrases’ meaning can be implied from their component words. Taking a document in Figure 1 as an example, the phrase “orbital vehicle” shares the same topic as the word “DC-X” (a reusable spaceship), as well as its component words “orbital”, and “vehicle”, which are all about the topic of space exploration. The assumption that the words within the same phrase tend to have the same latent topic is directly used by PhraseLDA (El-Kishky et al., 2014). Note that not all the phrases always have the same topic as their component words (e.g., the newspaper Boston Globe) (Mikolov et al., 2013). It’s difficult to distinguish the “orbital vehicle” type phrases from “Boston Globe” type phrases, but we can use the data-driven method to find out the semantically coherent ones by the NPMI metric (Bouma, 2009), and put them in Markov Random Fields (Kindermann and Snell, 1980) to align the topics of phrases and their component words.
Based on these two kinds of contextual information, we propose a novel topic model PhraseCTM and a two-stage method. In the first stage, we train PhraseCTM, which (1) double counts the phrases as two parts, one as the phrase itself, the other as the component words; (2) models the generation of words and phrases simultaneously by linking the phrases and component words within Markov Random Fields when they are semantically coherent; (3) uses the logistic normal distribution to represent the correlation among the topics, like a previous method CTM. In the second stage, we generate the correlation of topics from PhraseCTM.

We evaluate our method on five datasets by a quantitative experiment and a human study, showing that the correlated topic modeling on phrases is a good way to interpret the underlying themes of a corpus.

3 Related Works

There are two orthogonal lines of research studies related to our work. (1) With the development of phrase extraction techniques (El-Kishky et al., 2014; Liu et al., 2015; Shang et al., 2018), several topic models based on extracted phrases are proposed to provide high-quality phrase-level topics, such as PhraseLDA (El-Kishky et al., 2014), and TPM (He, 2016). Because of the quality of extracted phrases, PhraseLDA performs better than previous n-gram method TNG (Wang et al., 2007), which combines phrase extraction and topic modeling together. (2) CTM (Blei and Lafferty, 2005) uses the logistic normal distribution (Aitchison, 1982) to replace the Dirichlet prior (Aitchison, 1982) and to turn it into an operational orbital vehicle. Despite it will fail to meet some of the extraction limits, the reason people are so fond of it is that it is the only chance we have now, or will have for a long time, to come, to develop a launch vehicle with radically lower costs.

Figure 1: An example of phrases' contextual information. The phrases and words marked gray are about the same topic. The arrows show the topics of the phrases and their component words tend to be same, which tendency are modeled within Markov Random Field.

Inherited from NPMI, the semantic coherent score is defined upon two word types as NPMI \( w_{ij} \) (Bouma, 2009). The NPMI metric is defined upon two word types as NPMI \( (x, y) \) = \( \log \frac{p(x, y)}{p(x)p(y)} / (\log p(x, y)) \), where \( p(x) \) is estimated by the document frequency \( \frac{|d(x)|}{D} \), and \( p(x, y) \) = \( \frac{|d(x) \cap d(y)|}{D} \).

\[
s(w_{ij}^{(P)}, w_{ij}) = \min_{j,k \in \{1,\ldots,i\}} \text{NPMI}(w_j, w_k) \quad (1)
\]

Bouma (2009) pointed out that NPMI has the advantage that it ranges within the fixed interval. Inherited from NPMI, the semantic coherent score also ranges from -1 to 1. A negative semantic coherent score means the phrase does not share the same topic with its component words in the corpus level (e.g., long run, the newspaper Boston Globe). A positive score means the opposite, and the score
I suggests that the phrase and its component words should be aligned to the same topic in whole corpus. By a reasonable threshold $\tau$, we can add the semantically coherent link for $w_i^{(P)}$ and $w_{l(ii)}$, if $s(w_i^{(P)}, w_{l(ii)}) > \tau$. In practice, we set $\tau$ to 0.4.

Assuming the topic of the phrase $w_i^{(P)}$ is $z_i^{(P)}$ and the topics of words $w_{l(ii)}$ are $z_{l(ii)}$, when they have the above mentioned semantically coherent link, we put $z_i^{(P)}$ and $z_{l(ii)}$ in a Markov Random Field. More specifically, for $z_i^{(P)}$ and $z_j$, $j \in l(i)$, there’s the edge potential function $\exp\{\kappa \cdot \mathbb{P}(z_i^{(P)} = z_j)/|l(i)|\}$, where $\kappa$ is the weight to adjust how much the link is introduced to constrain the topics to be same. In the following experiment, $\kappa$ is set to be $10^{-5}$.

### 3.2 PhraseCTM

In the first stage, when given the prepared data as described in subsection 3.1, we are going to train better correlated phrase-level topics $\beta^{(P)}$. The contextual information of phrases include (i) words in the same document, and (ii) their component words within semantically coherent links. Part (ii) has been modeled in the previous subsection. For part (i), we let the phrases and words in a same document $d$ share the topic parameter $\eta_d$, which is a $K$-dimension vector sampled from a Gaussian distribution $\mathcal{N}(\mu, \Sigma)$. Like CTM, $\Sigma$ is the covariance matrix, modeling the correlation between topics.

As a part of MRF, the unary potential on the topic node $z_{d,i}^{(P)}$ or $z_{d,j}$ is defined by a logistic-normal distribution like CTM $p(z_{d,i} = k|\eta_d) = \exp\eta_d,k/\sum_k \exp\eta_d,k$. Therefore, the joint distribution of topics over the phrases and the words in document $d$ are defined as the following equation, where $A_d(\eta_d)$ is used for normalization, and $N_{L_d}$ is the number of semantically coherent links in document $d$.

$$p(z_{d,i}^{(P)}|\eta_d) = \frac{1}{A_d(\eta_d)} \prod_{m=1}^{N_d} p(z_{d,m}^{(P)}|\eta_d)$$

$$\cdot \prod_{i=1}^{N_d^{(P)}} p(z_{d,i}^{(P)}|\eta_d) \cdot \exp\{\sum_{i=1}^{N_{d,i}} \sum_{j \in l(i)} \mathbb{P}(z_{d,i}^{(P)} = z_{d,j})\}$$

The whole generation process is illustrated in Figure 2(a). We train PhraseCTM by variational inference like CTM. The different part lies in the phrases and component words which are in semantically coherent links. For these phrases, we use Eq. (2) to update the variational parameters $\phi_{d,i}^{(P)}$ for the latent topic $z_{d,i}^{(P)}$ of the phrase $w_{d,i}^{(P)}$. For the phrases that are not in semantically coherent links, we use the Eq. (3), which is same as the original CTM’s variational inference. In Eqs. (2) and (3), $\lambda_d$ is the variational parameter for $\eta_d$ such that $\eta_d,k \sim \mathcal{N}(\lambda_d,k, \nu^2_d,k)$.

$$\phi_{d,i,k}^{(P)} \propto \beta_{k,w_i}^{(P)} \exp(\lambda_d,k + \kappa/(l(i,j))) \sum_j \phi_{d,j,k}^{(P)}$$

(2)

$$\phi_{d,i,k}^{(P)} \propto \beta_{w_i}^{(P)} \exp(\lambda_d,k)$$

(3)

Similarly, the variational parameters for the component words in semantically coherent links are updated by Eq. (4), while other words are updated by Eq. (5). In this way, PhraseCTM introduces the impact from words and phrases on each other by Markov Random Fields.

$$\phi_{d,i,k} \propto \beta_{k,w_i} \exp(\lambda_d,k + \kappa/(l(i,j))) \phi_{d,j,k}^{(P)}$$

(4)

$$\phi_{d,i,k} \propto \beta_{w_i} \exp(\lambda_d,k)$$

(5)

![Figure 2](image-url)

(a) The first stage: training on our proposed model PhraseCTM. When observed words $W$ and phrases $W^{(P)}$, we learn word topics $\beta$, and phrase topics $\beta^{(P)}$.

![Figure 2](image-url)

(b) The second stage: inferring the phrase topics’ correlation. When given the phrases $W^{(P)}$, and the phrase topics $\beta^{(P)}$ learned from the first stage, we infer the phrase-level topics’ covariance $\Sigma^{(P)}$ as the correlation result.

### 3.3 Generation of Phrase Topics’ Correlation

In the second stage, we aim to get the correlation for $\beta^{(P)}$. It cannot be directly derived from $\Sigma$, which has also been learned in the first stage, because $\Sigma$ consists the impact from word topics. Thus, given $W^{(P)}$ and $\beta^{(P)}$, we use the variational inference again to learn $\Sigma^{(P)}$ as the illustration of Figure 2(b). Finally, the correlation matrix can be computed by $\text{corr}^{(P)}(i,j) = \frac{\Sigma^{(P)}_{i,j}}{\sqrt{\Sigma^{(P)}_{i,i} \Sigma^{(P)}_{j,j}}}$. 

3
4 Experiments

PhraseCTM is supposed to get benefits from two aspects: (1) generating high-quality phrase-level topics; (2) providing the correlation among phrase topics to help users to understand the underlying themes of a corpus. To check the first claim, we compare PhraseCTM with existing topic models on phrases. To evaluate the second claim, we design a user study to compare PhraseCTM with standard CTM that runs only on words.

Datasets. We choose several public text corpora, including 20Newsgroups (Lang, 1995), subsets of English Wikipedia, a subset of PubMed Abstracts (Varma et al., 1999). Due to efficiency problem, we do not test on the whole Wikipedia corpus. We construct the Mathematics, Chemistry, and Argentina subsets of English Wikipedia as (Huang et al., 2017). For each corpus, we extract the phrases by the implementation\(^1\) of AutoPhrase (Shang et al., 2018), and build the semantically coherent links as subsection 3.1. More specifically, in phrase extraction process, we set the minimum support as 5, and leave other parameters in AutoPhrase as its suggestion. In average, each document of the resulted 20Newsgroups only contains 2.7 phrases while 72.3 words, showing that phrases are much less than words. More statistics about the datasets are shown in Table 1.

4.1 Quantitative Result

Baselines. We compare with PhraseLDA (El-Kishky et al., 2014), the state-of-the-art model on phrases. Besides that, we run plain LDA and plain CTM\(^2\) directly on the extracted phrases (without considering the impact of words). To check the effectiveness of MRF, we run a variant version PhraseCTM(−) by removing all the semantically coherent links from PhraseCTM. We also run a n-gram based topic model TNG (Wang et al., 2007), which has already been implemented in Mallet\(^3\). All the topic numbers are set to be 100. For plain LDA and PhraseLDA, we set β = 0.005. For plain CTM and TNG, we use the default settings in their existing implementations. Since TNG combines phrase extraction and topic modeling together, we run it on the raw datasets.

We use the NPMI metric (Bouma, 2009) to evaluate the semantic coherence of top-10 phrases in each topic (K=100), by taking the entire English Wikipedia as the reference corpus. Roder (2015) has shown that the NPMI metric is highly correlated to human topic coherence ratings, so it’s natural to use it to show how PhraseCTM improves the quality of topics. Although we have already used NPMI in the semantic coherent score for finding semantically coherent links, it does not influence the rationality of the metric used for topic evaluation, because the semantic coherent score is defined upon two word types while the NPMI score on topics is defined upon two phrase types.

![Figure 3: The quality of the learned topics.](https://mallet.cs.umass.edu/)

The result is shown in Figure 3. Due to the computational costs, TNG cannot scale up to large datasets. Plain LDA and plain CTM performs not well on the datasets because of the sparsity of phrases in each document, while TNG performs better than them as it can utilize more words as its contextual information. PhraseCTM(−) is comparable to PhraseLDA in the experiment. PhraseLDA also utilizes all the contextual information with the assumption that the words in a phrase have the same topic. But this assumption

---

\(^1\)https://github.com/shangjingbo1226/AutoPhrase

\(^2\)https://github.com/blei-lab/ctm-c

\(^3\)http://mallet.cs.umass.edu/
is too strong, which can be adjusted by our introduced semantically coherent links in MRF. This experiment demonstrates that our method has generated high-quality phrase-level topics.

### 4.2 Human Study

To compare the correlated topics at different level, we directly run CTM on words. Trained on *Argentina@Wiki* and *Maths@Wiki* by CTM and PhraseCTM respectively, we outputted the topics with top-10 words/phrases in each topic and the correlation of topics for 10 human annotators, and asked them to label the topics. The duration of consuming time for topic labeling is a quite useful metric to check whether the topics are easy to understand for human annotators. It’s based on our following observation in the human study: the confused topic may consume more time to give it an appropriate label, while the good one is easy to understand for human and consumes less time.

There are 2 groups of annotators. The annotators in Group A got the CTM result on *Maths@Wiki* and PhraseCTM’s on *Argentina@Wiki*. The annotators in Group B got the results in the opposite setting. The labeling process was logged to calculate the accumulated time. The labeling time to reach 50 accurate topics’ labels on PhraseCTM is much less than the labeling time on CTM. In average, the annotators spent 7.1 minutes on PhraseCTM while 13.2 minutes on the others, which is listed in Table 2. In Figure 3, it’s easy to label the topics in top right corner as politics. And the edges in the figure illustrate the correlation between topics. As an example, the edge between the topic 71 and the topic 31 represents that the economics and the politics in Argentina is related, helping users to understand the corpus.

### 5 Conclusion

We provide a new topic model PhraseCTM to make the Correlated Topic Modeling available for phrase-level topics. PhraseCTM utilizes more contextual information of phrases, and put them within Markov Random Fields, so it can provide high-quality correlated topics at phrase level. The experiments show that the correlated topic modeling on phrases is a practical tool to interpret the underlying themes of a corpus. In future, we will optimize the efficiency of PhraseCTM to scale it up to large datasets.

### Acknowledgments

This research is supported by the Natural Science Foundation of China (Grant No.61572043) and National Key Research and Development Program (Project Number: 2016YFB1000704).
References


Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R. Voss, and Jiawei Han. 2014. Scalable topical phrase mining from text corpora. In Proceedings of the VLDB Endowment, volume 8, pages 305–316. VLDB Endowment.


Jialu Liu, Jingbo Shang, Chi Wang, Xiang Ren, and Jiawei Han. 2015. Mining quality phrases from massive text corpora. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, SIGMOD ’15, pages 1729–1744. ACM.


Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. 2018. Automated phrase mining from massive text corpora. IEEE Transactions on Knowledge and Data Engineering.

Yizhou Sun, Jiawei Han, Jing Gao, and Yintao Yu. 2009. topic2model: Information network-integrated topic modeling. In Ninth IEEE International Conference on Data Mining, pages 493–502. IEEE.


