PhraseCTM: Correlated Topic Modeling on Phrases within Markov Random Fields

Weijing Huang[1], Tengjiao Wang[1], Wei Chen[1], Siyuan Jiang[2], Kam-Fai Wong[1,3]
[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.
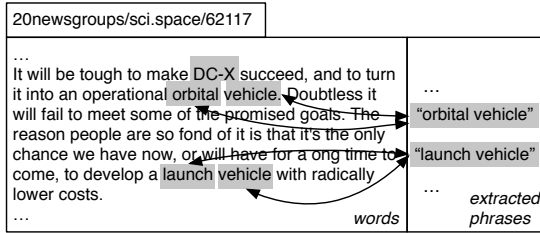
Figure 1: An example of phrases' contextual information. The phrases and words marked gray are about the same topic. The arrows show the topics of the phrases and their component words tend to be same, which tendency are modeled within Markov Random Field.

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.

2 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, so it can capture the correlated structure of topics. And the experiments in the further works (Chen et al., 2013; Roberts et al., 2016; He et al., 2017) showed its usefulness in exploring the text corpus by using

the correlated word-level topics. Note that our work is not a simple combination of these two methods, because the existing topic models on phrases lack the ability to capture the correlation structure while CTM cannot be directly applied on phrases due to the sparseness of phrases in each document. And as we used Markov Random Fields (Kindermann and Snell, 1980), our work is different from previous ones (Daume III, 2009; Sun et al., 2009; Xie et al., 2015) because we don't put all links into Markov Random Fields but only choose the semantic coherent links.

3 The proposed method

By preparing data as described in the subsection 3.1, our method is carried out in two stages, shown in the subsection 3.2, and 3.3 respectively.

3.1 Semantically Coherent Links for MRF

Given the input data in the form of raw text documents, we transform each document into the format as "words, phrases, semantically coherent links between phrases and component words". Extracting words is trivial, and extracting phrases can be conducted by using an existing tool, *e.g.*, AutoPhrase (Shang et al., 2018). In this process, each extracted phrase is counted twice, one as the phrase itself (represented in the phrase vocabulary), the other is divided into component words (represented in the word vocabulary).

In a given document, we denote the i-th phrase as $w_i^{(\mathcal{P})}$, and its component words as $\boldsymbol{w}_{l(i)}$. We use the Equation (1) to determine the semantic coherent score between $w_i^{(\mathcal{P})}$ and $\boldsymbol{w}_{l(i)}$ by utilizing NPMI (Bouma, 2009) . The NPMI metric is defined upon two word types as $\text{NPMI}(x, y) = \log \frac{p(x,y)}{p(x)p(y)} / (-\log p(x, y))$, where $p(x)$ is estimated by the document frequency $\frac{|d(x)|}{D}$, and $p(x,y) = \frac{|d(x) \cap d(y)|}{D}$.

$$s(w_i^{(\mathcal{P})}, \boldsymbol{w}_{l(i)}) = \min_{j,k \in l(i)} \{\text{NPMI}(w_j, w_k)\} \quad (1)$$

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

1 suggests that the phrase and its component words should be aligned to the same topic in whole corpus. By a reasonable threshold τ, we can add the semantically coherent link for $w_i^{(\mathcal{P})}$ and $\boldsymbol{w}_{l(i)}$, if $s(w_i^{(\mathcal{P})}, \boldsymbol{w}_{l(i)}) > \tau$. In practice, we set τ to 0.4.

Assuming the topic of the phrase $w_i^{(\mathcal{P})}$ is $z_i^{(\mathcal{P})}$ and the topics of words $\boldsymbol{w}_{l(i)}$ are $\boldsymbol{z}_{l(i)}$, when they have the above mentioned semantically coherent link, we put $z_i^{(\mathcal{P})}$ and $\boldsymbol{z}_{l(i)}$ in a Markov Random Field. More specifically, for $z_i^{(\mathcal{P})}$ and z_j, $j \in l(i)$, there's the edge potential function $\exp\{\kappa \cdot \mathbb{I}(z_i^{(\mathcal{P})} = z_j)/|l(i)|\}$, where κ is the weight to adjust how much the link is introduced to constraint the topics to be same. In the following experiment, κ is set to be 10^{-3}.

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 $\boldsymbol{\beta}^{(\mathcal{P})}$. 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 η_d, which is a K-dimension vector sampled from a Gaussian distribution $\mathcal{N}(\mu, \Sigma)$. Like CTM, Σ is the covariance matrix, modeling the correlation between topics.

As a part of MRF, the unary potential on the topic node $z_{d,i}^{(\mathcal{P})}$ or $z_{d,j}$ is defined by a logistic-normal distribution like CTM $p(z_{d,j} = k|\eta_d) = \exp \eta_{d,k}/\sum_k \exp \eta_{d,k}$. 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.

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

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

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 $\phi_{d,i}^{(\mathcal{P})}$ for the latent topic $z_{d,i}^{(\mathcal{P})}$ of the phrase $w_{d,i}^{(\mathcal{P})}$. 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), λ_d is the variational parameter for η_d such that $\eta_{d,k} \sim \mathcal{N}(\lambda_{d,k}, \nu_{d,k}^2)$.
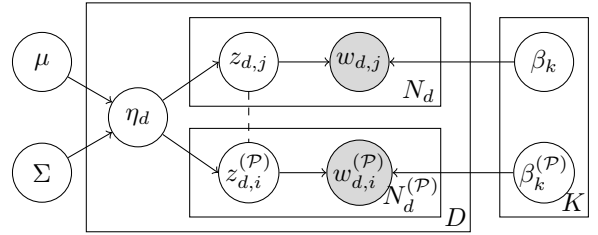
$$\hat{\phi}_{d,i,k}^{(\mathcal{P})} \propto \beta_{k,w_i}^{(\mathcal{P})} \exp(\lambda_{d,k} + \frac{\kappa}{|l(d,i)|} \sum_{j\in l(d,i)} \phi_{d,j,k}) \quad (2)$$

$$\hat{\phi}_{d,i,k}^{(\mathcal{P})} \propto \beta_{k,w_i}^{(\mathcal{P})} \exp(\lambda_{d,k}) \quad (3)$$

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.

$$\hat{\phi}_{d,i,k} \propto \beta_{k,w_i} \exp(\lambda_{d,k} + \frac{\kappa}{|l(d,j)|} \phi_{d,j,k}^{(\mathcal{P})}), \ i \in l(d,j) \quad (4)$$

$$\hat{\phi}_{d,i,k} \propto \beta_{k,w_i} \exp(\lambda_{d,k}) \quad (5)$$

(a) The first stage: training on our proposed model PhraseCTM. When observed words W and phrases $W^{(\mathcal{P})}$, we learn word topics $\boldsymbol{\beta}$, and phrase topics $\boldsymbol{\beta}^{(\mathcal{P})}$.
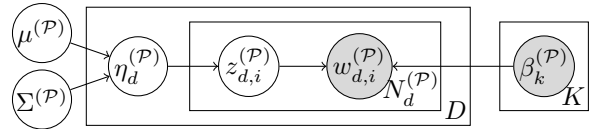
(b) The second stage: inferring the phrase topics' correlation. When given the phrases $W^{(\mathcal{P})}$, and the phrase topics $\boldsymbol{\beta}^{(\mathcal{P})}$ learned from the first stage, we infer the phrase-level topics' covariance $\Sigma^{(\mathcal{P})}$ as the correlation result.

Figure 2: Illustration of two stages of our method

3.3 Generation of Phrase Topics' Correlation

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

	$|V|$	$|V^{(\mathcal{P})}|$	$|W|$	$|W^{(\mathcal{P})}|$	$|D|$	$|W|/|D|$	$|W^{(\mathcal{P})}|/|D|$
20 Newsgroup	22,787	4,245	1,361,843	51,024	18,828	72.3	2.7
Argentina@Wiki	20,847	5,505	1,052,674	98,502	8,617	122.2	11.4
Mathematics@Wiki	43,779	27,371	6,062,815	594,704	27,947	216.9	21.3
Chemistry@Wiki	76,265	67,979	11,346,781	1,546,088	60,375	187.9	25.6
PubMed Abstracts	34,125	24,233	11,274,350	968,928	99,214	113.6	9.8

Table 1: The statistics of the datasets. In average, phrases appear more sparse than words. Phrases are extracted by AutoPhrase (Shang et al., 2018).

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* (Varmus 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 $\beta = 0.005$. 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.
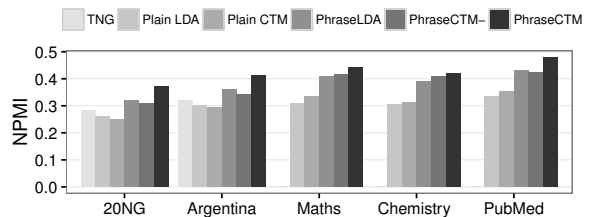

Figure 3: The quality of the learned topics.

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
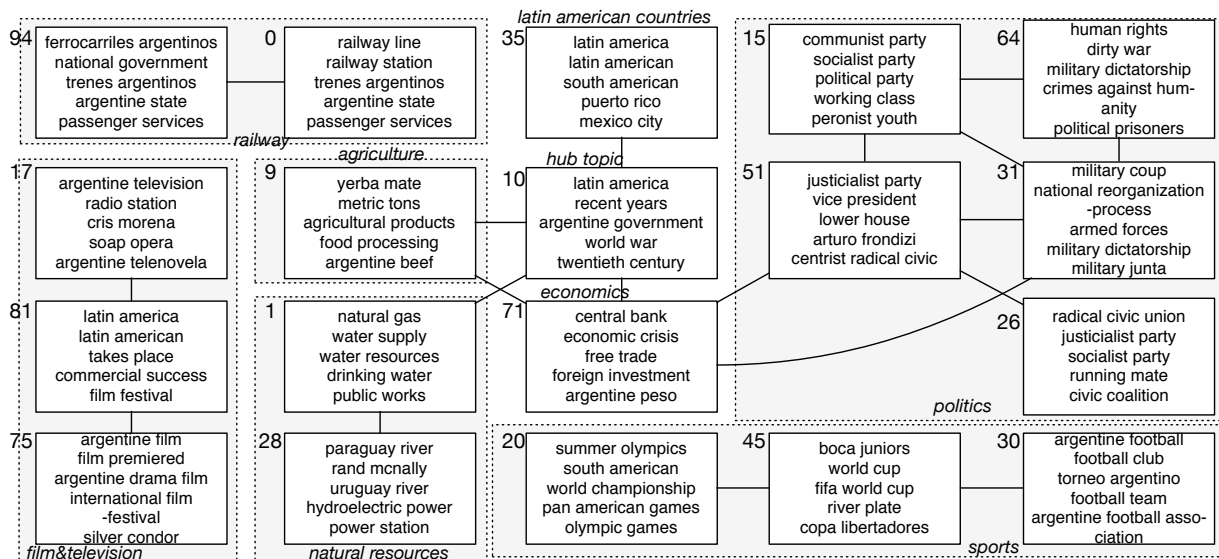
[1] https://github.com/shangjingbo1226/AutoPhrase
[2] https://github.com/blei-lab/ctm-c
[3] http://mallet.cs.umass.edu/

524

Figure 4: A part of the topic graph (K=100) generated by our method on the Argentina-related Wikipedia pages, where each node shows the top-5 phrases in each topic, and edges connect correlated topics. The left-aligned numbers in the graph represent the topic ID.

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 *poli-*

	CTM		PhraseCTM	
	Maths	Argentina	Maths	Argentina
Group A	12.4	-	-	7.5
Group B	-	14.0	6.7	-
In Average		13.2		7.1

Table 2: Human time consumption on topic labeling for correlated topics generated by CTM and PhraseCTM, measured in minutes.

tics. 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

References

John Aitchison. 1982. The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):139–177.

David M. Blei and John D. Lafferty. 2005. Correlated topic models. In *Proceedings of the 18th International Conference on Neural Information Processing Systems*, NIPS'05, pages 147–154. MIT Press.

David M. Blei and John D. Lafferty. 2007. A correlated topic model of science. *The Annals of Applied Statistics*, 1(1):17–35.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the German Society for Computational Linguistics & Language Technology*, pages 31–40.

Jianfei Chen, Jun Zhu, Zi Wang, Xun Zheng, and Bo Zhang. 2013. Scalable inference for logistic-normal topic models. In *Advances in Neural Information Processing Systems 26*, pages 2445–2453.

Hal Daume III. 2009. Markov random topic fields. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 293–296. Association for Computational Linguistics.

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.

Junxian He, Zhiting Hu, Taylor Berg-Kirkpatrick, Ying Huang, and Eric P. Xing. 2017. Efficient correlated topic modeling with topic embedding. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'17, pages 225–233. ACM.

Yulan He. 2016. Extracting topical phrases from clinical documents. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 2957–2963. AAAI Press.

Weijing Huang, Tengjiao Wang, Wei Chen, and Yazhou Wang. 2017. Category-level transfer learning from knowledge base to microblog stream for accurate event detection. In *International Conference on Database Systems for Advanced Applications*, DASFAA'17, pages 50–67. Springer.

Noriaki Kawamae. 2014. Supervised n-gram topic model. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 473–482. ACM.

Ross Kindermann and J Laurie Snell. 1980. *Markov random fields and their applications*, volume 1. American Mathematical Society.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Proceedings of the 12th International Conference on Machine Learning*, pages 331–339.

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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Margaret E. Roberts, Brandon M. Stewart, and Edoardo M. Airoldi. 2016. A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 111(515):988–1003.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 399–408. 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. itopicmodel: Information network-integrated topic modeling. In *Ninth IEEE International Conference on Data Mining*, pages 493–502. IEEE.

Jian Tang, Zhaoshi Meng, Xuanlong Nguyen, Qiaozhu Mei, and Ming Zhang. 2014. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *Proceedings of the 31st International Conference on Machine Learning*, pages 190–198.

Harold Varmus, David Lipman, and Patrick Brown. 1999. Pubmed central: an nih-operated site for electronic distribution of life sciences research reports.

Xuerui Wang, Andrew McCallum, and Xing Wei. 2007. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, ICDM '07, pages 697–702. IEEE.

Pengtao Xie, Diyi Yang, and Eric Xing. 2015. 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*, pages 725–734. Association for Computational Linguistics.