Short Text Understanding by Leveraging Knowledge into Topic Model

Shansong Yang, Weiming Lu*, Dezhi Yang, Liang Yao, Baogang Wei
Zhejiang University
Hangzhou, Zhejiang 310000, China
{yangshansong,luwm,deathyyoung,yaoliang,wbg}@zju.edu.cn

Abstract

In this paper, we investigate the challenging task of understanding short text (*STU* task) by jointly considering topic modeling and knowledge incorporation. Knowledge incorporation can solve the content sparsity problem effectively for topic modeling. Specifically, the phrase topic model is proposed to leverage the auto-mined knowledge, i.e., the phrases, to guide the generative process of short text. Experimental results illustrate the effectiveness of the mechanism that utilizes knowledge to improve topic modeling's performance.

1 Introduction

The explosion of online text content, such as twitter messages, text advertisements, QA community messages and product reviews has given rise to the necessity of understanding these prevalent short texts.

Conventional topic modeling, like PLSA (Hofmann, 1999) and LDA (Blei et al., 2003) are widely used for uncovering the hidden topics from text corpus. However, the sparsity of content in short texts brings new challenges to topic modeling.

In fact, short texts usually do not contain sufficient statistical signals to support many state-of-the-art approaches for text processing such as topic modeling (Hua et al., 2015). Knowledge is indispensable to *STU* task, where knowledge-based topic model (Andrzejewski et al., 2009; Hu et al., 2011; Jagarlamudi et al., 2012; Mukherjee and Liu, 2012; Chen et al., 2013; Yan et al., 2013) has attracted more attention recently.

*Corresponding author

We consider, in the *STU* task, the available knowledge can be divided into two classes: self-contained knowledge and external knowledge. Self-contained knowledge, which is focused in this paper, is extracted from the short text itself, such as key-phrase. External knowledge is constructed without special purpose, such as WordNet (Miller, 1995), KnowItAll (Etzioni et al., 2005), Wikipedia (Gabrilovich and Markovitch, 2007), Yago (Suchanek et al., 2007), NELL (Carlson et al., 2010) and Probase (Wu et al., 2012).

PLSA and LDA are the typical unsupervised topic models, that is non-knowledgeable model. In contrast, Biterm topic model (BTM) (Yan et al., 2013) leverages self-contained knowledge into semantic analysis. BTM learns topics over short texts by modeling the generation of biterms in the whole corpus. A biterm is an unordered word-pair co-occurring in short contexts. BTM posits that the two words in a biterm share the same topic drawn from a mixture of topics over the whole corpus. The major advantage of BTM is that BTM explicitly model the word co-occurrences in the local context, which well captures the short-range dependencies between words.

External knowledge-based models incorporate expert domain knowledge to help guide the models. DF-LDA (Andrzejewski et al., 2009) model incorporates domain knowledge in the form of *must-link* and *cannot-link*. *Must-link* states that two words should belong to the same topic, while *cannot-link* states that two words should not be in the same topic. GK-LDA (Chen et al., 2013) leverages lexical semantic relations of words such as synonyms, antonyms and adjective attributes in topic models. A

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 1232–1237,
Denver, Colorado, May 31 – June 5, 2015.

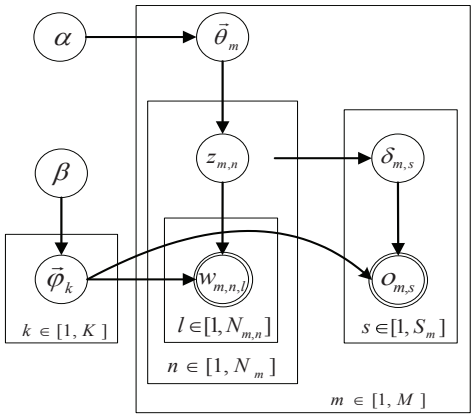

Figure 1: The phrase topic model proposed in this paper.

vast amount of lexical knowledge about words and their relationships, denoted as *LR-sets*, available in online dictionaries or other resources can be exploited by this model to generate more coherent topics.

However, for external knowledge-based models, the incorporated knowledge is too general to be consistent with the short text in the semantic space. On the other hand, BTM, as a typical self-contained knowledge-based model, makes rough assumption on the generated biterms. The generated biterms are inundated with noise, for not any two terms in short text share same topic. Based on the above analysis, we first identify key-phrases from short text, which can be deemed as self-contained knowledge, then propose phrase topic model (PTM), which constrains same topic for terms in key-phrase and sample topics for non-phrase terms from mixture of key-phrase's topic.

2 Phrase Topic Model

2.1 Model

A phrase is defined as a consecutive sequence of terms, or unigrams. In this paper, we focus on self-contained knowledge in short text, i.e., the key-phrases. Key-phrase extraction is a fundamental component in our work. We use CRF++[1] to identify key-phrases in a short text. The training data is built manually, and the features contain the word itself, the part of speech tagged by Stanford Log-linear Part-Of-Speech Tagger (Toutanova et al., 2003). Sample identified key-phrases are shown in Table 2.

In this paper, our phrase topic model is proposed based on three assumptions:

- Key-phrases are the key points of interest in the short text, which should be the focus.
- Terms consisting of the same key-phrase will share common topic.
- Non-phrase term's topic assignment should depend on that of key-phrases in the same text.

Our assumptions is indeed similar to other models (Gruber et al., 2007), for example each sentence is assumed to be assigned to one topic, however this assumption is too general, in many cases, different words should be assigned different topics even in short text. Our model is more refined to distinguish key-phrase and non-phrase. In addition, if two or more key-phrases exist in the same short text, they are probably assigned different topics.

The graphical representation of PTM is illustrated in Figure 1. α and β are hyper-parameters, which are experienced tuned. φ is corpus-level parameter, while θ is document-level parameter. The hidden variables consist of $z_{m,n}$ and $\delta_{m,s}$. The generative process of phrase topic model is presented as follows.

- For each topic $k \in [1, K]$
  - draw a topic-specific word distribution $\varphi_k \sim Dir(\beta)$
- For each document $m \in [1, M]$
  - draw a topic distribution $\theta_m \sim Dir(\alpha)$
  - For each key-phrase $n \in [1, N_m]$
    - draw topic assignment $z_{m,n} \sim Multi(\theta_m)$
    - For each word $l \in [1, N_{m,n}]$
      - draw $w_{m,n,l} \sim Multi(\varphi_{z_{m,n}})$
  - For each non-phrase word $s \in [1, S_m]$
    - draw a topic assignment $\delta_{m,s} \sim Uniform(z_{m,1}, \ldots, z_{m,N_m})$
    - draw word $o_{m,s} \sim Multi(\varphi_{\delta_{m,s}})$

From this process, we can see the generation of key-phrases and non-phrases are distinguished and non-phrase's generation is based on the topic assignment of key-phrases in the same document.

[1] http://crfpp.googlecode.com/svn/trunk/doc/index.html

2.2 Inference By Gibbs Sampling

Similarly with LDA, collapsed Gibbs sampling (Griffiths and Steyvers, 2004) can be utilized to perform approximate inference. In our model, the hidden variables are key-phrase's topic assignment $\boldsymbol{z}$ and non-phrase word's topic assignment $\boldsymbol{\delta}$. To perform Gibbs sampling, we first randomly initialize the hidden variables. Then we sample the topic assignment based on the conditional distribution $p(z_{m,n} = k|\mathbf{z}_{\neg(m,n)}, \mathbf{w}, \mathbf{o}, \delta)$ and $p(\delta_{m,s} = k|\mathbf{z}, \mathbf{w}, \mathbf{o}, \delta_{\neg(m,s)})$.

We can derive the conditional probability for $z_{m,n}$ following Equation 1, where $n^{k}_{m,\neg(m,n)}$ denotes the number of key-phrases whose topic assignment are k in document m without consideration of key-phrase $\{m, n\}$, which is similar to $n^{k'}_{m,\neg(m,n)}$. $n^{w_{m,n,l}}_{k,\neg(m,n)}$ denotes the number of times key-phrase term $w_{m,n,l}$ assigned to topic k without consideration of key-phrase $\{m, n\}$, which is similar to $n^{w}_{k,\neg(m,n)}$. $n^{o_{m,s}}_{k,\neg m}$ denotes the number of times non-phrase term $o_{m,s}$ assigned to topic k without consideration of document m, which is similar to $n^{w}_{k,\neg m}$.

Similarly, we can derive the conditional probability for $\delta_{m,s}$ following Equation 2, where $n^{o_{m,s}}_{k,\neg(m,s)}$ denotes the number of times non-phrase term $o_{m,s}$ assigned to topic k without consideration of non-phrase term $\{m, s\}$, which is similar to $n^{w}_{k,\neg(m,s)}$. L_m denotes the number of topics assigned to key-phrases in document m.

Finally, we can easily estimate the topic distribution $\theta_{m,k}$ and topic-word distribution $\varphi_{k,w}$ following Equation 3 and 4.

$$\theta_{m,k} = \frac{n^{k}_{m} + \alpha}{\sum_{k'=1}^{K} n^{k'}_{m} + K\alpha} \tag{3}$$

$$\varphi_{k,w} = \frac{n^{w}_{k} + \beta}{\sum_{w'=1}^{V} n^{w'}_{k} + V\beta} \tag{4}$$

3 Experiments and Results

Online reviews dataset (Chen et al., 2013), which consists of four domains, is utilized to evaluate our model, where each domain collection contains 500 reviews. Each review's average length is 20.42. The statistics of each domain are presented in Table 1. It's worth noting that the ***Phrase*** is auto-identified by the key-phrase extraction method. And the ***Word*** represents the whole distinct words for those identified key-phrases.

In our paper, we assumed each domain has a single topic model. For different domain, we think the semantic space is quite different. So we performed the proposed topic model with respect to different domain. The number of topics is usually determined by experience, in our experiment, each domain collection contains 500 reviews, we think the number of topics ranging from 2 to 20 is appropriate, and these reviews are sufficient to train a topic model.

Table 1: Statistic information of the dataset.

Dataset	***Phrase***	***Word***	Vocabulary
Computer	1439	1423	5109
Cellphone	1110	1109	4184
Camera	2962	2620	8366
Food	1235	1350	4488

Recent research (Chang et al., 2009; Newman et al., 2010) shows that the models which achieve better predictive perplexity often have less interpretable latent spaces. So the *Topic Coherence Metric* (Mimno et al., 2011) is utilized to assess topic quality, which is consistent with human labeling.

We compare our model with four baseline models: non-knowledgeable model LDA, self-contained knowledgeable model BTM, external knowledge-based model GK-LDA (Chen et al., 2013) and DF-LDA (Andrzejewski et al., 2009). Those identified key-phrases are used as *must-links* in DF-LDA and *LR-sets* in GK-LDA. This can ensure the incorporated knowledge upon different models are equal.

Table 2 illustrates the auto-identified phrases from cellphone dataset. From this result, we can see key-phrase extraction method can efficiently identify mostly phrases. More than one phrase, for example *warranty service* and *android phone*, may appear in a single sentence, and their topic assignments are probably different. Our proposed phrase topic model(PTM) can well handle this case, which is more well-defined than the assumption of all words within a sentence share one topic. Our phrase topic model assumes non-phrase term's topic assignment should depend on that of key-phrases in the same text. This assumption can be clearly confirmed by Table 2, for example, *Nokia N97 mini* is semantic dependent *US-*

$$p(z_{m,n}=k|\mathbf{z}_{\neg(m,n)},\mathbf{w},\mathbf{o},\delta) = \frac{n^{k}_{m,\neg(m,n)}+\alpha}{\sum_{k'=1}^{K} n^{k'}_{m,\neg(m,n)}+K\alpha} \cdot \frac{\prod_{l=1}^{N_{m,n}}(n^{w_{m,n,l}}_{k,\neg(m,n)}+\beta)}{\prod_{l=1}^{N_{m,n}}(\sum_{w=1}^{V} n^{w}_{k,\neg(m,n)}+V\beta)} \cdot \frac{\prod_{s=1}^{S_m}(n^{o_{m,s}}_{k,\neg m}+\beta)}{\prod_{s=1}^{S_m}(\sum_{w=1}^{V} n^{w}_{k,\neg m}+V\beta)} \tag{1}$$

$$p(\delta_{m,s}=k|\mathbf{z},\mathbf{w},\mathbf{o},\delta_{\neg(m,s)}) = \frac{n^{o_{m,s}}_{k,\neg(m,s)}+\beta}{\sum_{w=1}^{V} n^{w}_{k,\neg(m,s)}+V\beta} \cdot \frac{1}{L_m} \tag{2}$$

B charge cable, the same as *company* and *warranty service*.

For all models, posterior inference was drawn after 1000 Gibbs iterations with an initial burn-in of 800 iterations. For all models, we set the hyperparameters $\alpha = 2$ and $\beta = 0.5$.

The evaluation results over *Topic Coherence Metric* are presented in Figure 2 and Figure 3. This figure indicates our model and BTM can get higher topic coherence score than GK-LDA and DF-LDA, which means the self-defined knowledge and the mechanism of knowledge incorporation are effective to topic model. LDA's performance is acceptable but not stable. Our model performs better than BTM, which is probably because the rough assumption of BTM on generated biterms. From the above analysis, we can see our proposed model can get the best performance.

T-test results show that the performance improvement of our model over baselines is statistically significant on *Topic Coherence Metric*. All p-values for t-test are less than 0.00001.

Figure 4 presents the fluctuation of topic coherence when tuning the hyper-parameter α and β. We can see that the performance fluctuates within a limited range as we vary α and β. The topic coherence fluctuates between -550 and -950 other than food dataset, which gets less fluctuation range.

Table 3 shows example topics for each domain, where inconsistent words are highlighted in red. From this results, we can see the number of errors in phrase topic model(PTM) is significantly less than LDA, which indicates our proposed topic model is more suitable than LDA for short text.

4 Conclusions and Future Work

In this paper, we present a topic model to achieve *STU* task starting from key-phrases. The terms in key-phrases identified from the short texts are supposed to share a common topic respectively. And those key-phrases are assumed to be the central focus in the generative process of documents. In the future work, the self-contained knowledge, such as those identified key-phrases, and the external knowledge-base should be integrated to guide topic modeling.

Acknowledgements

This work is supported by the National Natural Science Foundation of China No.61103099, the Fundamental Research Funds for the Central Universities(2014QNA5008), Chinese Knowledge Center of Engineering Science and Technology(CKCEST) and Specialized Research Fund for the Doctoral Program of Higher Education(SRFDP)(No.20130101110136).

References

[Andrzejewski et al.2009] David Andrzejewski, Xiaojin Zhu, and Mark Craven. 2009. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009*, pages 25–32.

[Blei et al.2003] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

[Carlson et al.2010] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr.,

Table 2: Identified Key-phrase in Cellphone dataset

[1] Both sites list compatible devices including the *Samsung Galaxy Tab*.
[2] My Dell Streak needed more power than any normal *USB car adapter* could give me.
[3] This actually comes with a micro *USB charge cable* which fits and works perfectly for my Nokia N97 mini.
[4] I contacted the company for *warranty service*. On my *android phone* I paired ...
[5] Everything from pulling it out of the box to syncing it with both my IPhone and *Ipod touch* 4g were effortless.

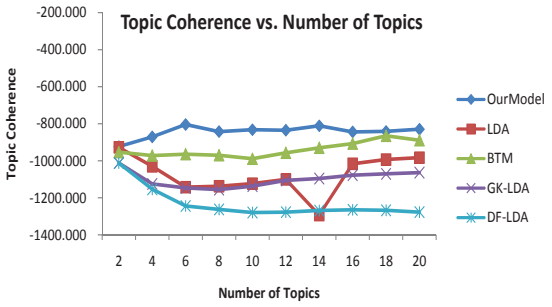

Figure 2: Average Topic Coherence score of each model.

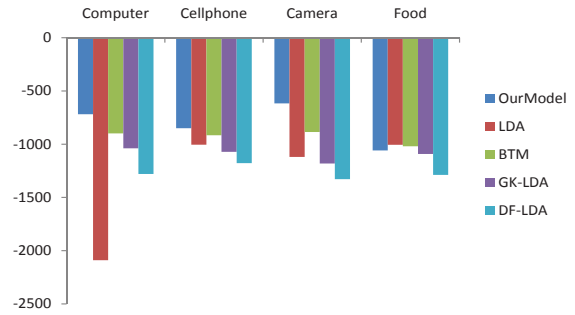

Figure 3: Detailed Topic Coherence score of $T = 15$.

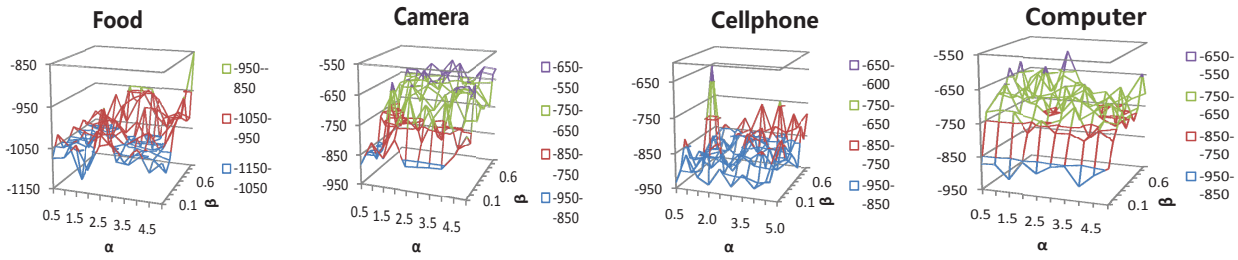

Figure 4: Parameter influences with fixed topic number $T = 15$.

Table 3: Example topics. First row: domain, Second row: inferred topic tag. Errors are highlighted in red.

Cellphone		Computer		Food		Camera	
music		game		dinner		buy	
LDA	PTM	LDA	PTM	LDA	PTM	LDA	PTM
phone	phone	*buy*	fps	coffee	soup	camera	camera
music	music	*make*	disruption	*product*	good	bought	bought
iphone	car	games	*dips*	*found*	bread	wanted	pictures
calls	radio	*time*	playable	*love*	mix	*year*	video
play	device	fast	wars	*amazon*	popcorn	*time*	sony
bluetooth	sound	play	age	bread	taste	*happy*	*back*
hear	quality	people	*pretty*	popcorn	great	purchase	canon
free	iphone	*thing*	update	eating	*bag*	*day*	battery
cell	volume	card	star	*ordered*	flavor	*month*	*time*
listen	bluetooth	*full*	empires	*bought*	make	love	price
hands	*easy*	*product*	laggy	good	coffee	*ago*	lens
charge	good	*read*	unplayable	taste	eat	*week*	quality

and Tom M. Mitchell. 2010. Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010*.

[Chang et al.2009] Jonathan Chang, Jordan L. Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *23rd Annual Conference on Neural Information Processing Systems 2009*, pages 288–296.

[Chen et al.2013] Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2013. Discovering coherent topics using general knowledge. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 209–218. ACM.

[Etzioni et al.2005] Oren Etzioni, Michael J. Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. 2005. Unsupervised named-entity extraction from the web: An experimental study. *Artif. Intell.*, 165(1):91–134.

[Gabrilovich and Markovitch2007] Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611.

[Griffiths and Steyvers2004] Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235.

[Gruber et al.2007] Amit Gruber, Yair Weiss, and Michal Rosen-Zvi. 2007. Hidden topic markov models. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, AISTATS 2007, San Juan, Puerto Rico, March 21-24, 2007*, pages 163–170.

[Hofmann1999] Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57.

[Hu et al.2011] Yuening Hu, Jordan L. Boyd-Graber, and Brianna Satinoff. 2011. Interactive topic modeling. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 248–257.

[Hua et al.2015] Wen Hua, Zhongyuan Wang, Haixun Wang, Kai Zheng, and Xiaofang Zhou. 2015. Short text understanding through lexical-semantic analysis. In *International Conference on Data Engineering (ICDE)*.

[Jagarlamudi et al.2012] Jagadeesh Jagarlamudi, Hal Daum III, and Raghavendra Udupa. 2012. Incorporating lexical priors into topic models. In *13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213.

[Miller1995] George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

[Mimno et al.2011] David M. Mimno, Hanna M. Wallach, Edmund M. Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011*, pages 262–272.

[Mukherjee and Liu2012] Arjun Mukherjee and Bing Liu. 2012. Aspect extraction through semi-supervised modeling. In *The 50th Annual Meeting of the Association for Computational Linguistics*, pages 339–348.

[Newman et al.2010] David Newman, Youn Noh, Edmund M. Talley, Sarvnaz Karimi, and Timothy Baldwin. 2010. Evaluating topic models for digital libraries. In *Proceedings of the 2010 Joint International Conference on Digital Libraries*, pages 215–224.

[Suchanek et al.2007] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007*, pages 697–706.

[Toutanova et al.2003] Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *HLT-NAACL 2003, Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

[Wu et al.2012] Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Qili Zhu. 2012. Probase: a probabilistic taxonomy for text understanding. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2012, Scottsdale, AZ, USA, May 20-24, 2012*, pages 481–492.

[Yan et al.2013] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *22nd International World Wide Web Conference*, pages 1445–1456.