Review Topic Discovery with Phrases using the Pólya Urn Model

Geli Fei
Department of Computer Science, University of Illinois at Chicago, Chicago, USA
gfei2@uic.edu

Zhiyuan Chen
Department of Computer Science, University of Illinois at Chicago, Chicago, USA
czyuanacm@gmail.com

Bing Liu
Department of Computer Science, University of Illinois at Chicago, Chicago, USA
liub@cs.uic.edu

Abstract

Topic modelling has been popularly used to discover latent topics from text documents. Most existing models work on individual words. That is, they treat each topic as a distribution over words. However, using only individual words has several shortcomings. First, it increases the co-occurrences of words which may be incorrect because a phrase with two words is not equivalent to two separate words. These extra and often incorrect co-occurrences result in poorer output topics. A multi-word phrase should be treated as one term by itself. Second, individual words are often difficult to use in practice because the meaning of a word in a phrase and the meaning of a word in isolation can be quite different. Third, topics as a list of individual words are also difficult to understand by users who are not domain experts and do not have any knowledge of topic models. In this paper, we aim to solve these problems by considering phrases in their natural form. One simple way to include phrases in topic modelling is to treat each phrase as a single term. However, this method is not ideal because the meaning of a phrase is often related to its composite words. That information is lost. This paper proposes to use the generalized Pólya Urn (GPU) model to solve the problem, which gives superior results. GPU enables the connection of a phrase with its content words naturally. Our experimental results using 32 review datasets show that the proposed approach is highly effective.

1 Introduction

Topic models such as LDA (Blei et al., 2003) and pSLA (Hofmann 1999) and their extensions have been popularly used to find topics in text documents. These models are mostly governed by the phenomenon called "higher-order co-occurrence" (Heinrich 2009), i.e., how often terms co-occur in different contexts. Word w_1 co-occurring with word w_2 which in turn co-occurs with word w_3 denotes a second-order co-occurrence between w_1 and w_3. Almost all these models regard each topic as a distribution over words. The words under each topic are often sorted according to their associated probabilities. Those top ranked words are used to represent the topic. However, this representation of topics as a list of individual words has some [1]major shortcomings:

- Topics are often difficult to understand or interpret by users unless they are domain experts and also knowledgeable about topic models. In most real-life situations, these are not the case. In some of our applications, we show users several good topics, but they have no idea what they are because many domain phrases cannot be split to individual words. For example, "battery" and "life" are put under the same topic, which is not bad. But the users wondered why "battery" and "life" are the same because they thought words under a topic should somehow have similar meanings. We had to explain that it is due to "battery life." As another example, sentences such as "This hotel has a very nice sandy beach" may cause a topic model to put "hotel" and "sandy" in a topic, which is not wrong but again it is hard to understand by a user who may not be able to connect the two words. Thus in order to interpret topics well, the user must know the phrases (they are split into individual words) that may

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 667–676, Dublin, Ireland, August 23-29 2014.

be used in a domain and how words may be associated with each other. To make the matters worse, in most cases, the topics generated from a topic model are not perfect. There are some wrong words under a topic, which make the interpretation even harder.

- Individual words are difficult to use in practice because in some cases a word under a topic may not have its intended meaning for the topic in a particular sentence context. This can cause many mistakes. For example, in sentiment analysis of product reviews, a topic is often regarded as a set of words indicating a product feature or attribute. This is not true in many cases. For example, if "battery" and "life" are put in one topic, when the system sees "life," it assumes it is related to "battery." But in the sentence "The life expectancy of the machine is about 2 years," this "life" has nothing to do with battery or battery life. This causes an error. If the system can directly use phrases, "battery life" and "life expectancy," the error will not occur.
- Splitting phrases into multiple individual words causes extra co-occurrences that may result in poor or wrong topics involving other words. For example, due to sentences like "Beach staffs are rude" and "The hotel has a nice sandy beach," a topic model may put "staff" and "sandy" under a topic for staff and/or put "beach" and "rude" together under the topic of beach views.

Based on our experiences in opinion mining and social media mining, these are major issues with topic models. We believe that they must be dealt with before wide spread adaptation of topic models in real-life applications. In this paper, we make an attempt to solve this problem. We will use *term* to represent both word and phrase, and use *word* or *phrase* when we want to distinguish them.

One obvious way to consider phrases is to use a natural language parser to find all phrases and then treat each phrase as one term, e.g., "battery life," "sandy beach" and "beach staff." However, the problem with this approach is that it may lose the connection of many related words or phrases in a topic. For example, under the topic for beach, we may not find "sandy beach" because there is no co-occurrence of "sandy beach" and "beach" if we treat "sandy beach" as a single term. This is clearly not a good solution as it may miss a lot of topical terms (words or phrases) for a topic. It can also result in poor topics due to the loss of co-occurrences.

Another obvious solution is to use individual words as they are, but add an extra term representing the phrase. For example, we can turn the sentence "This hotel has a nice sandy beach" to "This hotel has a nice sandy beach <sandy beach>." This solution helps deal with the problem of losing co-occurrences to some extent, but because the words are still treated individually, the three problems discussed above still exist, although the phrase "sandy beach" now can show up in some topics. However, due to the fact that phrases are obviously less frequent than individual words, they may be ranked very low, which make little difference to solving the three problems.

In this paper, we propose a novel approach to solve the problem, which is based on the generalized Pólya urn (GPU) model (Mahmoud 2008). GPU was first introduced into LDA in (Mimno et al., 2011) to concentrate words with high co-document frequency. However, Mimno et al. (2011) and other researchers Chen et al., (2013) still use them in the framework of individual words. In the GPU model, we can deal with the problems above by treating phrases as individual terms and allowing their component words to have some connections or co-occurrences with them. Furthermore, we can push phrases up in a topic as phrases are important for understanding but are usually less frequent than individual words and ranked low in a topic. The intuition here is that when we see a phrase, we also see a small fraction of their component words; and when we see each individual word, we also see a small fraction of its related phrases. Further, in a phrase not all words are equally important. For example, in "hotel staff", "staff" is more important as it is the head noun, which represents the semantic category of the phrase.

Our experiments are conducted using online review collections from 32 domains. We will see that the proposed method produces significantly better results both quantitatively based on the statistical measure of topic coherence and qualitatively based on human labeling of topics and topical terms.

In summary, this paper makes the following contributions:

1. It proposes to consider phrases in topic models, which as we have explained above, is important for accurate topic generation, the use of the resulting topics and human interpretation. As we will see in Section 2, although some prior works exist, they are based on n-grams (Mukherjee and Liu, 2013). They are different from our approach. N-grams can generate many non-understandable phrases. Furthermore, due to infrequency of n-grams (much less frequent than individual words),

typically a huge amount of data is needed in order to produce reasonable topics, which many applications simply do not have.
2. It proposes to use the generalized Pólya Urn (GPU) model to deal with the problems arising in considering phrases. To the best of our knowledge, the GPU model has not been used in the context of phrases. This model not only generates better topics, but also rank phrases relatively high in their topics, which greatly helps understanding of the generated topics.
3. Comprehensive experiments conducted using product and service review collections from 32 domains demonstrate the effectiveness of the proposed model.

2 Related Work

GPU was first introduced to topic modelling in (Mimno et al., 2011), in which GPU is used to concentrate words with high co-document frequency based on corpus-specific co-occurrence statistics. Chen et al. (2013) applied GPU to deal with the adverse effect of using prior domain knowledge in topic modeling by increasing the counts of rare words in the knowledge sets. However, these works still use only individual words.

Topics in most topic models like LDA are unigram distributions over words and assume words to be exchangeable at the word level. However, there exists some work that tries to take word order into consideration by including n-gram language models. Wallach (2006) proposed the Bigram Topic Model (BTM) which integrates bigram statistics with topic-based approaches to document modeling. Wang et al. (2007) proposed the Topical N-gram Model (TNG), which is a generalization of the BTM. It generates words in their textual order by first sampling a topic, then sampling its status as a unigram or bigram, and then sampling the word from a topic-specific unigram or bigram distribution. Although the "bag-of-words" assumption does not always hold in real-life applications, it offers a great computational advantage over more complex models taking word order into account for discovering significant n-grams. Our approach is different from these works in two ways. First, we still follow the "bag-of-words" or rather "bag-of-terms" assumption. Second, we find actual phrases rather than just n-grams. Most n-grams are still hard to understand because they are not natural phrases.

Blei and Lafferty (2009), Liu et al. (2010) and Zhao et al. (2011) also try to extract keyphrases from texts. Their methods, however, are very different because they identify multi-word phrases using relevance and likelihood scores in the post-processing step based on the discovered topical unigrams.

Mukherjee and Liu (2013) and Mukherjee et al. (2013) all try to include n-grams to enhance the expressiveness of their models while preserving the advantages of "bag-of-words" assumption, which has a similar idea as our paper. However, as we point out in the introduction, this way of including phrases/n-grams suffers from several shortcomings. Solving these problems is the goal of our paper.

Finally, since we use product reviews as our datasets, our work is also related to opinion mining using topic models, e.g. (Mei et al., 2007; Lu and Zhai, 2008; Titov and McDonald, 2008; Zhao et al., 2010; Li et al., 2010; Sauper and Barzilay, 2013; Lin and He, 2009; Jo and Oh, 2011). However, none of these models uses phrases.

3 Proposed Model

We start by briefly reviewing the Latent Dirichlet Allocation (LDA) model (Blei et al., 2003). Then we describe the simple Pólya urn (SPU) model, which is embedded in LDA. After that, we present the generalized Pólya urn (GPU) model and discuss how it can be applied to our context. The proposed model uses GPU for its inference. It shares the same graphical model as LDA. However, the GPU inference mechanism is very different from that of LDA, which cannot be reflected in the graphical model or the generative process as it only helps to infer more desirable posterior distributions of topic models.

3.1 Latent Dirichlet Allocation

LDA is a generative probabilistic model for a document collection. It assumes that documents are represented as a mixture of latent topics, and each latent topic is characterized by a distribution over terms. In order to generate a term $w_n^{(d)}$ in document d, where n is its position, we first draw a discrete topic assignment $z_n^{(d)}$ from a document-specific distribution over T topics θ_d, which is drawn from a prior Dirichlet distribution with hyperparameter α. Then we draw a term from the topic-specific distribution

over the vocabulary $\phi_{z_n^{(d)}}$, which is drawn from a prior Dirichlet distribution with hyperparameter β.

For inference, instead of directly estimating θ and ϕ, Gibbs sampling is used to approximate them based on the posterior estimates of latent topic assignment $\boldsymbol{z}$. The Gibbs sampling procedure considers each term in the documents in turn, and estimates the probability of assigning the current term to each topic, conditioned on the topic assignments to all other terms. Griffiths and Steyvers (2004) showed this could be calculated by:

$$p\left(z_n^{(d)} = t \middle| \boldsymbol{z}_{-d,n}, W, \alpha, \beta\right) \propto \frac{C_{t|d} + \alpha}{C_d + T\alpha} \times \frac{N_{w_n^{(d)}|t} + \beta}{N_t + V\beta} \quad (1)$$

where $z_n^{(d)} = t$ represents the topic assignment of term $w_n^{(d)}$ to topic t, and $\boldsymbol{z}_{-d,n}$ refers to the topic assignments of all other terms. W denotes all terms in the document collection, V denotes the size of vocabulary of the collection, T is the number of topics in the corpus, $N_{w|t}$ is the count of term w under topic t, $N_t = \sum_{w'} N_{w'|t}$, and $C_{t|d}$ refers the count of topic t being assigned to some terms in document d, $C_d = \sum_{t'} C_{t'|d}$. All these counts exclude the current term.

3.2 Simple Pólya Urn Model

Traditionally, the *Pólya urn* model is designed in the context of colored balls and urns. In the context of topic models, a term can be seen as a ball of a certain color and the urn contains a mixture of balls with various colors. The classic topic-word (or topic-term) distribution can be reflected by the color proportion of balls in the urn. LDA follows the *simple Pólya urn* (SPU) model, which works as follows: when a ball of a particular color is drawn from an urn, that ball is put back to the urn along with another ball of the same color. This process corresponds to assigning a topic to a term in the Gibbs sampler of LDA. Based on the topic-specific "collapsed" probability of a term w given topic t, $\frac{N_{w|t}+\beta}{N_t+V\beta}$, which is essentially the second ratio in (1), drawing a term w will only increase the probability of seeing w in the future sampling process. This self-reinforcing property is known as "the rich get richer". In the next subsection, we will introduce the *generalized Pólya urn* (GPU) model, which increases the probability of seeing certain other terms when we sample a term.

3.3 Generalized Pólya Urn Model

The *generalized Pólya urn* (GPU) model differs from SPU in that, when a ball of a certain color is drawn, two balls of that color is put back along with a certain number of balls of some other colors. Unlike SPU, GPU sampling not only allows us to see a ball of the same color again with higher probability, but also increases the probability of seeing balls with certain other colors. These additional balls of certain other colors added to the urn increase their proportions in the urn. We call this the *promotion* of these colored balls. Applying the idea, there are two directions of promotion in our application (Note that in each sentence, we need to identify each phrase, but do not need to add any extra information):

1. Word to phrase: When an individual word is assigned to a topic (analogous to drawing a ball of a certain color), each phrase containing the word will be promoted, meaning that the phrase will be added to the same topic with a small count. That is, a fraction of the phrase will be assigned to the topic. This is justified because it is reasonable to assume that the phrase is related to the word to some extent in meaning.
2. Phrase to word: When a phrase is assigned to a topic, each component word in it is also promoted with a certain small count. That is, each word is also assigned the topic by a certain amount. In most cases, the head nouns are more important. Thus, we promote the head nouns more. For example, in "hotel staff", "staff" is the head noun that determines the category of the noun phrase. The rationale of this promotion is similar to that above.

Let $w_n^{(d)}$ be a word and p_w be the word itself or a phrase containing the word $w_n^{(d)}$. v represents a term, and p_v indicates all the related terms of v. The new GPU sampling is as follows:

$$p\left(z_n^{(d)} = t \middle| \boldsymbol{z}_{-d,n}, W, \alpha, \beta, A\right) \propto \frac{C_{t|d} + \alpha}{C_d + T\alpha} \times \frac{\sum_{p\_w} N_{p\_w|t} A_{p\_w, w_n^{(d)}} + \beta}{\sum_v \sum_{p\_v} N_{p\_v|t} A_{p\_v,v} + V\beta} \quad (2)$$

where A is a $V \times V$ real-value matrix, each cell of which contains a real value *virtualcount*, indicating the amount of promotion of a term under a topic when assigning this topic to another term. V is size of all terms. The new model retains the document-topic component of standard LDA, which is the first ratio in (1), but replaces the usual Pólya urn topic-word (topic-term) component, the second ratio in (1), with a generalized Pólya urn framework (Mahmoud 2008; Mimno et al., 2011). The simple Pólya urn model is a simplified version of GPU in which matrix A is an identity matrix. In this paper, A is an asymmetric matrix because the main goal of using GPU is to promote the less frequent phrases in the documents.

4 EXPERIMENTS

In this section, we evaluate the proposed method of considering phrases in topic discovery, and compare it with three baselines. The first baseline discovers topics using LDA in a traditional way without considering phrases, i.e., using only individual words. We refer to this baseline as *LDA*(*w*). The second baseline considers phrases by treating each whole phrase as a separate term in the corpus. We refer to this baseline as *LDA*(*p*). The third baseline considers phrases by keeping individual component words in the phrases as they are, but also adding phrases as extra terms. We refer to this baseline as *LDA*(*w_p*). We refer to our proposed method as *LDA*(*p_GPU*). Note that for those words that are not in any phrases, they are treated as individual words (or unigrams).

Data Set: We use product reviews from 30 sub-categories (types of product) in the electronics domain from Amazon.com. The sub-categories are "Camera", "Mouse", "Cellphone," etc (see the whole list below Figure 1). Each domain contains 1,000 reviews. Besides, we also use a collection of hotel reviews and a collection of restaurant reviews from TripAdvisor.com and Yelp.com. The hotel review data contains 101,234 reviews, and the restaurant review data contains 25,459 reviews. We thus have a total of 32 domains. We ran the Stanford Parser to perform sentence detection, lemmatization and POS tagging. Punctuations, stopwords, numbers and words appearing less than 5 times in each dataset are removed. Domain names are also removed, e.g., word "camera" for the domain Camera, since it co-occurs with most words in the dataset, leading to high similarity among topics/aspects.

Sentences as Documents: As noted in (Titov and McDonald, 2008), when standard topic models are applied to reviews as documents, they tend to produce topics that correspond to global properties of products (e.g., product brand name), but cannot separate different product aspects or features well. The reason is that all reviews of the same product type basically evaluate the same aspects of the product type. Only the brand names and product names are different. Thus, using individual reviews for modelling is ineffective for finding product aspects or features, which are our topics. Although there are approaches which model sentences (Jo and Oh, 2011; Zhao et al., 2010; Titov and McDonald, 2008), we take the approach in (Brody and Elhadad, 2010; Chen et al., 2013), dividing each review into sentences and treating each sentence as an independent document.

Noun Phrase Detection: Although there are different types of phrases, in this first work we focus only on noun phrases as they are more representative of topics in online reviews. We will deal with other types of phrases in the future. Our first step is thus to obtain all noun phrases from each domain. Due to the efficiency issue of full natural language parser with a huge number of reviews, instead of applying the Stanford Parser to recognize noun phrases, we design a rule-based approach to recognize noun phrases as consecutive nouns based on POS tags of sentences. Although the Stanford Parser may give us better noun phrases, our simple method serves the purpose and gives us very good results. In fact, based on our initial experiments, the Stanford Parser also gives many wrong phrases.

Parameter Settings: In all our experiments, the posterior inference was drawn after 2000 Gibbs sampling iterations with a burn-in of 400 iterations. Following (Griffiths and Steyvers, 2004), we fix the Dirichlet priors as follows: for all document-topic distributions, we set α=50/K, where K is the number of topics. And for all topic-term distributions, we set β=0.1. We also experimented with other settings of these priors and did not notice much difference.

Setting the number of topics/aspects in topic models is often tricky as it is difficult to know the exact number of topics that a corpus has. While non-parametric Bayesian approaches (Teh et al., 2005) do exist for estimating the number of topics, it's not the focus of this paper. We empirically set the number of topics to 15. Although 15 may not be optimum, since all models use the same number, there is no bias against any model.

In Section 3.3, we introduced the *promotion* concept for the GPU model. When we sample a topic for a word, we add *virtualcount* of topic assignment to all its related phrases. However, not all words in a phrase are equally important. For example, in phrase "hotel staff", "staff" is more important, and we call such words the head nouns. In this work, we apply a simple method used in (Wang et al., 2007), which is to always assume that the last word in a noun phrase is the head noun. Although we are aware of the potential harm to our model when we promote a wrong word, we will leave it as our future work. Again, because we want to connect phrases with their component words and promote the rank of phrases in their topics, we add less virtual counts to individual words. Thus, we add 0.5 * *virtualcount* to the last word in a phrase and add 0.25 * *virtualcount* to all other words. We set *virtualcount* = 0.1 in our experiments empirically.

Based on the discovered topics, we conduct statistical evaluation using topic coherence, human evaluation and also a case study to quantitatively and qualitatively show the superiority of the proposed method in terms of both interpretability and topic wellness.

4.1 Statistical Evaluation

Perplexity and KL-divergence are often used to evaluate topic models statistically. However, researchers have found that perplexity on held-out documents is not always a good predictor of human judgments of topics (Chang et al., 2009). In our application, we are not concerned with the test on future data using the hold-out set. KL-divergence measures the difference of distributions, and thus can be used to measure the distinctiveness of topics. However, distinctiveness of topics does not necessarily mean human agreeable topics. Recently, Mimno et al. (2011) proposed a new measure called topic coherence, which has been shown to correlate with human judgments of topic quality quite well. Higher topic coherence score indicates higher quality of topics, i.e., better topic coherence. Topic coherence is computed as below.

$$TC\left(t; V^{(t)}\right) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} log \frac{D\left(v_m^{(t)}, v_l^{(t)}\right) + 1}{D\left(v_l^{(t)}\right)} \tag{3}$$

in which $D(v)$ is the document frequency of term v (i.e., the number of documents with at least one term v) and $D(v, v')$ is the co-document frequency of term v and term v' (i.e., the number of documents containing both term v and term v'). Also, $V^{(t)} = \left(v_1^{(t)}, \ldots, v_M^{(t)}\right)$ is the list of M most probable terms in topic t. 1 is added as a smoothing count to avoid taking the logarithm of zero.

We thus use this measure to score all four experiments. Figure 1 and Figure 2 show the topic coherence using top 15 terms and top 30 terms respectively on the 32 different domains. Notice the topic coherence is a negative value, and a smaller absolute value is better than a larger one. Firstly, we can see from both charts that our proposed model *LDA*(*p_GPU*) is better than all other three baselines by a large margin. Secondly, the performance of the other three baselines are quite similar. In general, *LDA*(*p*) is slightly worse than the other two baselines. It is because replacing many words with phrases decreases the number of co-occurrences in the corpus. In contrast, *LDA*(*w_p*) is slightly better than the other two baselines on most domains because some frequent phrases add more reliable co-occurrences in the corpus. However, as we point out in the introduction, some problems still exist. Firstly, it does not solve the problem of phrases and their component words having different meanings, and thus artificially creating such wrong co-occurrences may damage the overall performance. Secondly, even if the number of co-occurrences increases, most of the phrases are still too infrequent to be ranked high in their associated topics to be useful in helping users understand the topic.

In order to test the significance of the improvement, we conduct paired *t*-tests on the topic coherence results. Using both 15 top terms and 30 top terms, statistical tests show that our proposed method, *LDA*(*p_GPU*), outperforms all three baselines significantly ($p < 0.01$). However, there's no significant improvement between any pair of the three baselines.

4.2 Manual Evaluation

Although several statistical measures, such as perplexity, KL-divergence and topic coherence, have been used to statistically evaluate topic models, since topic models are mostly (including ours) unsupervised,

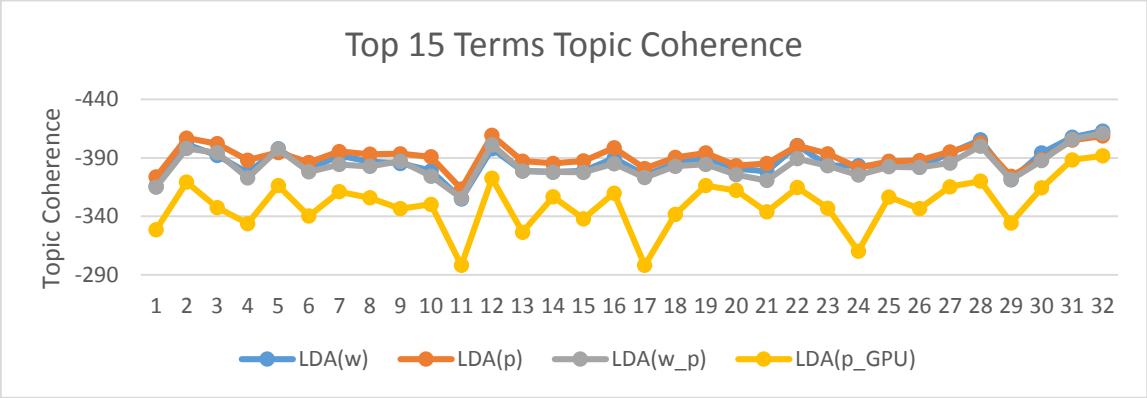

Figure 1: Topic coherence of the top 15 terms of each model on each of the 32 datasets. Notice that since topic coherence is a negative value, a smaller absolute value is better than a larger one.

Domain/dataset names are listed as follows (1:Amplifier; 2:BluRayPlayer; 3:Camera; 4:CellPhone; 5:Computer; 6:DVDPlayer; 7:GPS; 8:HardDrive; 9:Headphone; 10:Keyboard; 11:Kindle; 12:MediaPlayer; 13:Microphone; 14:Monitor; 15:Mouse; 16:MP3Player; 17:NetworkAdapter; 18:Printer; 19:Projector; 20:RadarDetector; 21:RemoteControl; 22:Scanner; 23:Speaker; 24:Subwoofer; 25:Tablet; 26:TV; 27:VideoPlayer; 28:VideoRecorder; 29:Watch; 30:WirelessRouter; 31:Hotel; 32:Restaurant).

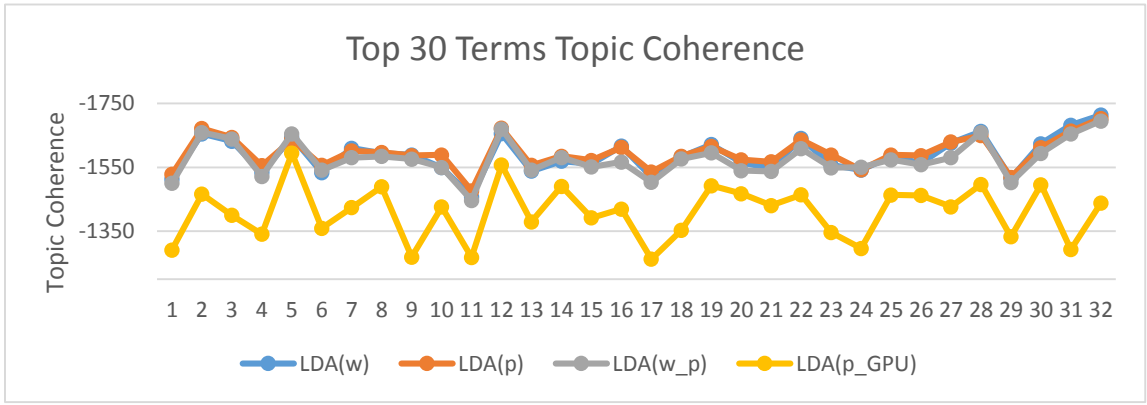

Figure 2: Topic coherence of the top 30 terms of each model on each dataset. Notice again that since topic coherence is a negative value, a smaller absolute value is better than a larger one. X-axis indicates the domain id numbers, whose names are listed below Figure 1.

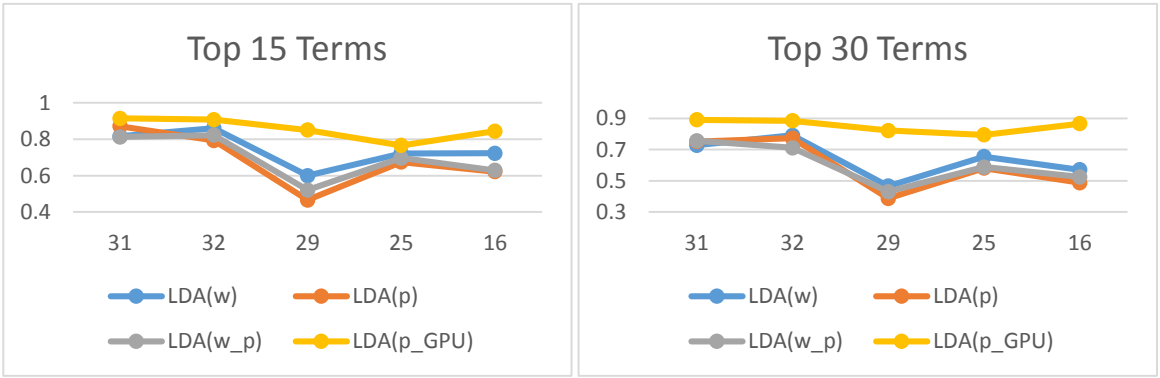

Figure 3: Human evaluation on five domains using top 15 and top 30 terms. X-axis indicates the domain id numbers, whose corresponding domain names are listed below Figure 1. Y-axis indicates the ratio of correct topic terms.

statistical measures may not always correlate with human interpretations or judgments. Thus, in this sub-section, we perform a manual evaluation through manual labeling of topics and topical terms.

Manual labeling was done by two annotators, who are familiar with reviews and topic models. The labeling was carried out in two stages sequentially: (1) labeling of topics and (2) labeling of topical terms in each topic. After the first stage, an annotator agreement is computed and then the two annotators discuss about the disagreed topics to reach a consensus. Then, they move on to the next stage to label the top ranked topical terms in each topic (based on their probabilities in the topic). For the annotator

Table 1: Example topics discovered by LDA(w) and LDA(p_GPU)

Hotel		**Restaurant**		**Watch**	
LDA(w)	LDA(p_GPU)	LDA(w)	LDA(p_GPU)	LDA(w)	LDA(p_GPU)
bed	clean	service	service	hand	big
comfortable	comfortable	star	friendly	minute	hand
small	quiet	staff	server	hour	minute
sleep	sleep	atmosphere	staff	beautiful	cheap
size	large	friendly	atmosphere	casual	hour
large	spacious	server	waiter	christmas	automatic
tv	size	waiter	attentive	setting	seconds
pillow	king size bed	attentive	star	condition	line
king	pillow	reason	service staff	worth	hour hand
chair	queen size bed	decor	star service	weight	durable
table	bed size	quick	customer service	red	analog hand
mattress	bed nd pillow	customer	table service	press	hand move
clean	bed sheet	waitress	delivery service	gift	hand line
double	bed linen	tip	rush hour service	run	seconds hand
big	sofa bed	pleasant	service attitude	functionality	hand sweep

Tablet		**MP3Player**	
LDA(w)	LDA(p_GPU)	LDA(w)	LDA(p_GPU)
screen	screen	battery	battery
touch	size	headphone	hour
software	easier	life	battery life
hard	pro	media	price
pad	touch screen	car	worth
option	bigger	windows	charge
version	area	hour	replacement
website	inch	decent	free
angle	screen protector	reason	market
car	screen size	xp	aaa battery
charger	inch screen	program	aa battery
ipod	draw	aaa	purchase
worth	home screen	window	hour battery
gb	screen look	set	aaa
drive	line	pair	life

agreement, we compute Kappa scores. The Kappa score for topic labeling is 0.838, and the Kappa score for topical terms labeling is 0.846. Both scores indicate strong agreement in the labeling.

Evaluation measure. A commonly used evaluation measure in human evaluation is *precision@n* (or *P@n* for short), which is the precision at a particular rank position *n* in a topic. For example, *Precision*@5 means the precision of the top ranked 5 terms for a topic. To be consistent with the automatic evaluation, we use *Precision*@15 and 30. Top 15 terms is usually sufficient to represent the topic. However, since we include phrases in our experiments which may lead to some other terms ranked lower than using only words, we labeled up to top 30 terms. The *Precision*@n measure is also used in (Zhao et al., 2010) and some others, e.g., (Chen et al., 2013).

In our experiments, we labeled four results for each domain, i.e., those of *LDA*(*w*), *LDA*(*p*), *LDA*(*w_p*) and *LDA*(*p_GPU*). Due to the large amount of human labeling effort, we only labeled 5 domains. We find that it is sometimes hard to figure out what some of the topics are about and whether some terms are related to a topic or not, so we give the results to our human evaluators together with the phrases in each domain extracted by our rules in order to let them be familiar with the domain vocabulary. The human evaluation results are shown in Figure 3.

Results and Discussions. Again, we conduct paired *t*-tests on the human evaluation results of top 15 and 30 terms. Statistical tests show that our proposed method, *LDA*(*p_GPU*), outperforms all other three methods significantly ($p < 0.05$) using both top 15 and top 30 terms. However, there's no significant improvement between any pair of the three baselines.

4.3 Case Study

In order to illustrate the importance of phrases in enhancing human readability, we conduct case study using one topic from each of the five manually labeled domains. Due to space limitations, we only compare the results of our model *LDA*(*p_GPU*) with *LDA*(*w*).

In the above table, we notice that with phrases, the topics are much more interpretable than only reading individual words given by *LDA*(*w*). For example, "hand" in "Watch" domain given by *LDA*(*w*) is quite confusing at first, but in *LDA*(*p_GPU*), "hour hand" makes it more understandable. Another example is "aaa" in "MP3Player" domain. It is quite confusing at first, but "aaa battery" should make it more interpretable by an application user who is not familiar with topic models or does not have extensive domain knowledge. Also, due to wrong co-occurrences created by individual words in a phrase, the *LDA*(*w*) results contain much more noise than those of *LDA*(*p_GPU*).

5 CONCLUSION

This paper proposed a new method to consider phrases in discovering topics using topic models. The method is based on the generalized Pólya urn (GPU) model, which allows us to connect phrases with their component words during the inference and rank phrases higher in their related topics. Our method preserves the advantages of "bag-of-words" assumption while preventing the side effects that traditional methods have when considering phrases. We tested our method against three baselines across 32 different domains, and demonstrated the superiority of our method in improving the topic quality and human interpretability both quantitatively and qualitatively.

References

David M. Blei and John D. Lafferty. 2009. "Visualizing Topics with Multi-Word Expressions." Tech. Report. (arXiv:0907.1013).

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." Journal of Machine Learning Research 993-1022.

Samuel Brody and Noemie Elhadad. 2010. "An Unsupervised Aspect-Sentiment Model for Online Reviews." NAACL. Los Angeles, California: ACL.

Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. "Reading Tea Leaves: How Humans Interpret Topic Models." Neural Information Processing Systems.

Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2013. "Exploiting Domain Knowledge in Aspect Extraction." EMNLP.

Thomas L. Griffiths, and Mark Steyvers. 2004. "Finding scientific topics." Proceedings of National Academy of Sciences.

Gregor Heinrich. 2009. "A Generic Approach to Topic Models." ECML PKDD. ACM. Pages 517 - 532.

Thomas Hofmann. 1999. "Probabilistic latent semantic analysis." UAI.

Yohan Jo and Alice Oh. 2011. "Aspect and Sentiment Unification Model for Online Review Analysis." WSDM. Hong Kong, China: ACM.

Chenghua Lin and Yulan He. 2009. "Joint Sentiment/Topic Model for Sentiment Analysis". CIKM. Hong Kong, China.

Fangtao Li, Minlie Huang, Xiaoyan Zhu. 2010. "Sentiment Analysis with Global Topics and Local Dependency". AAAI

Yue Lu and Chengxiang Zhai. 2008. "Opinion Integration Through Semi-supervised Topic Modeling." WWW. 2008, Beijing, China: ACM.

Zhiyuan Liu, Wenyi Huang, Yabin Zheng, and Maosong Sun. 2010. "Automatic Keyphrase Extraction via Topic Decomposition." EMNLP.

Arjun Mukherjee and Bing Liu. 2013. "Discovering User Interactions in Ideological Discussions." ACL.

Arjun Mukherjee, Vivek Venkataraman, Bing Liu, and Sharon Meraz. 2013. "Public Dialogue: Analysis of Tolerance in Online Discussions." ACL.

David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. "Optimizing Semantic Coherence in Topic Models." EMNLP. Edinburgh, Scotland, UK: ACL.

Hosan Mahmoud. 2008. Polya Urn Models. Chapman & Hall/CRC Texts in Statistical Science.

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. 2007. "Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs." WWW. Banff, Alberta, Canada: ACM.

Christina Sauper and Regina Barzilay. 2013. "Automatic Aggregation by Joint Modeling of Aspects and Values". Journal of Artificial Intelligence Research 46 (2013) 89-127

Ivan Titov and Ryan McDonald. 2008. "Modeling Online Reviews with Multi-grain Topic Models." WWW. 2008, Beijing, China: ACM.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2005. "Hierarchical Dirichlet Processes." Journal of the American Statistical Association.

Hanna M. Wallach. 2006. "Topic Modeling: Beyond Bag-of-Words." ICML. Pittsburgh, PA: ACM.

Xuerui Wang, Andrew McCallum, and Xing Wei. 2007. "Topical N-grams: Phrase and Topic Discovery, with an Application to Information Retrieval." ICDM.

Wayne Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. 2010. "Jointly Modeling Aspects and Opinions with a MaxEnt-LDA Hybrid." EMNLP. Massachusetts, USA: ACL.

Wayne Xin Zhao, Jing Jiang, Jing He, Yang Song, Palakorn Achananuparp, Ee-Peng Lim, and Xiaoming Li. 2011. "Topical Keyphrase Extraction from Twitter." ACL.