Leveraging Multi-Domain Prior Knowledge in Topic Models

Zhiyuan Chen[†], Arjun Mukherjee[†], Bing Liu[†], Meichun Hsu[‡], Malu Castellanos[‡], Riddhiman Ghosh[‡]
[†]University of Illinois at Chicago, [‡]HP Labs
{czyuanacm, arjun4787}@gmail.com, liub@cs.uic.edu,
{meichun.hsu, malu.castellanos, riddhiman.ghosh}@hp.com

Abstract

Topic models have been widely used to identify topics in text corpora. It is also known that purely unsupervised models often result in topics that are not comprehensible in applications. In recent years, a number of knowledge-based models have been proposed, which allow the user to input prior knowledge of the domain to produce more coherent and meaningful topics. In this paper, we go one step further to study how the prior knowledge from other domains can be exploited to help topic modeling in the new domain. This problem setting is important from both the application and the learning perspectives because knowledge is inherently accumulative. We human beings gain knowledge gradually and use the old knowledge to help solve new problems. To achieve this objective, existing models have some major difficulties. In this paper, we propose a novel knowledge-based model, called MDK-LDA, which is capable of using prior knowledge from multiple domains. Our evaluation results will demonstrate its effectiveness.

1 Introduction

Topic models, such as pLSA [Hofmann, 1999] and LDA [Blei *et al*., 2003] provide a powerful framework for extracting latent topics in text documents [Bart, 2011; Mei *et al*., 2007; Waltinger *et al*., 2011; Wang *et al*., 2010]. Topic generation in these models is based on what [Heinrich, 2009] refers to as "higher-order co-occurrence", i.e., how often words co-occur in different contexts. In recent years, researchers also found that these unsupervised models may not produce topics that conform to the user's existing knowledge [Mimno *et al*., 2011]. One key reason is that the objective functions of topic models (e.g., LDA) often do not correlate well with human judgments [Chang *et al*., 2009].

To deal with this problem, several *knowledge-based models* have been proposed, which use domain knowledge provided by the user to guide the modeling process. For example, the DF-LDA model in [Andrzejewski *et al*., 2009] can incorporate such knowledge in the form of must-links and cannot-links. A must-link states that two words should belong to the same topic, while a cannot-link states that two words should not be in the same topic. In [Andrzejewski *et al*., 2011], more general knowledge using first-order logic can be specified. In [Burns *et al*., 2012; Jagarlamudi *et al*., 2012; Lu *et al*., 2011; Mukherjee and Liu, 2012], seeded models were also proposed, which allow the user to specify some prior seed words in some topics. A seed set, which is a set of seed words, can also be expressed as must-links. The model in [Hu *et al*., 2011] further enables the user to provide guidance interactively. In this paper, we use *s-set* (semantic-set) to refer to a set of words sharing the *same semantic meaning* in a domain, similar to must-link. We don't use the existing terminology due to its restricted meaning which is not suitable for the proposed framework. This will become clear shortly. This work does not use cannot-links.

One characteristic of the existing models is that for each domain new knowledge is needed because different domains have different knowledge. This is, however, undesirable because knowledge should be accumulative. Human beings gain new knowledge gradually and old knowledge is not discarded but used to help solve new problems. For topic modeling, we want to achieve a similar effect since different domains do share some knowledge. This paper proposes a new topic model which can exploit knowledge from multiple past domains in the form of s-sets provided by the user (details in Section 3.1) to produce coherent topics in the new domain. The s-sets from multiple domains are combined to serve as knowledge for the new domain.

Problem definition: Given a set *S* of s-sets (prior knowledge) from multiple past domains, we want to leverage *S* to help topic modeling in the new domain.

There are, however, two key challenges for this to work:

Multiple senses: A word typically has multiple meanings or senses. In a particular domain, it may take one sense and the user inputs an s-set in that sense. However, in the new domain, the word may take another sense, which means that the existing s-set is not appropriate for the new domain. To make matters worse, even in the same domain, a word may take multiple senses, e.g., *light*, which can mean "*of little weight*" or "*something that makes things visible.*" Existing models cannot handle multiple senses except the model in [Jagarlamudi *et al*., 2012] (but it has its own problems, detailed below). DF-LDA cannot handle multiple senses because its definition of must-link is transitive, which is the same in [Andrzejewski *et al*., 2011]. That is, if A and B

form a must-link, and B and C form a must-link, it implies a must-link between A and C, indicating A, B and C should be in the same topic. This is also the case for the models in [Petterson *et al*., 2010] and [Mukherjee and Liu, 2012]. Although the model in [Jagarlamudi *et al*., 2012] allows multiple senses for a word, it requires that each topic has at most one seed set, which means that the number of seed sets must not exceed the number of topics. This is undesirable since the number of s-sets should be flexible (allowing for whatever knowledge). Our model can deal with arbitrary number of s-sets from multiple domains.

Adverse effect of must-links: In handling must-links (or seed sets), existing models basically bring the words in a must-link or a seed set close together by ensuring that they have similar probability under the same topic. This causes the following problem. If a must-link consists of a frequent word and an infrequent word, in bringing them closer, the probability for the frequent word in a topic will decrease while probability for the infrequent word will increase. This can harm the final topics as the reduction of probability of the frequent (often important) word under a topic can result in some irrelevant words being ranked higher (due to redistribution of word probability masses under the topic). We observed this undesirable phenomenon in DF-LDA. This issue becomes a more serious concern when multi-domain knowledge is applied, as the frequencies of words in s-sets across different domains can vary greatly.

This paper proposes a new model, called MDK-LDA (LDA with Multi-Domain Knowledge). To deal with multiple senses, we add a new latent variable s in LDA to model s-sets. Each document is an admixture of latent topics while each topic is a probability distribution over s-sets. MDK-LDA is able to handle multiple senses because the new latent variable s enables the model to choose the right sense represented by an s-set. For example, the word *light* can have two s-sets with distinct senses, i.e., {light, heavy, weight} and {light, bright, luminance}. *Light* in a document can be correctly assigned to the s-set {light, heavy, weight} if it co-occurs with *heavy* or *weight*. If *light* co-occurs with *bright* or *luminance* in some other documents, then the s-set {light, bright, luminance} will be assigned to it in those documents. Note that if a word w is not provided with any s-set, it is treated as a singleton s-set $\{w\}$. Clearly, a word can have multiple s-sets corresponding to its senses used in one domain or multiple domains. If two s-sets share the same sense of a word, either s-set can be assigned to the word if the sense is correct for this word in the domain.

However, this base model, which we call MDK-LDA(b), is insufficient due to the *adverse effect* issue above, which also occurs in existing models. For example, we have an s-set {price, cost}. The word *price* is very frequent in the camera review domain but *cost* is infrequent. MDK-LDA(b) will redistribute the probability masses, i.e., it will increase the mass for *cost* but decrease the mass for *price* to bring them closer in a topic. However, a more appropriate situation would be that as *price* being very frequent, it should be correct in its own topic and the other words in its s-set should be promoted as well. In order to achieve this objective, we use the *generalized Pólya urn* (GPU) *model* [Mahmoud, 2008] to promote the s-set as a whole. The GPU model was first introduced in LDA in [Mimno *et al*., 2011] to concentrate words with high co-document frequency, but [Mimno *et al*., 2011] does not use any prior knowledge in their model. Using GPU, we can control the effect of promoting s-sets, which gives us the final model MDK-LDA.

Since we use prior knowledge from multiple domains to help topic modeling in the new domain, our work is essentially a form of *transfer learning*. However, to the best of our knowledge, there is still no work in the transfer learning literature that performs our task, although there are existing works that use topic models to help perform transfer learning in the supervised learning setting [Faisal *et al*., 2012; Pan and Yang, 2010; Xue *et al*., 2008]. Our focus is not supervised learning, but to generate coherent topics.

Our experimental results from online reviews in six different domains show the effectiveness of MDK-LDA, which outperforms several baseline methods by a large margin.

2 The Proposed MDK-LDA Model

2.1 Generative Process

We now introduce our base model MDK-LDA(b). Let M be the number of documents where each document m has N_m words. The vocabulary in the corpus is denoted by $\{1, \dots, V\}$. Since the words in an s-set share a similar semantic meaning, the model should redistribute the probability masses over words in the s-set to ensure that they have similar probability under the same topic. To incorporate this idea, we introduce a new latent variable s, which denotes the s-set assignment to each word. Assume that there are S s-sets in total. The generative process is given as follows:

1. For each topic $t \in \{1, \dots, T\}$
   i. Draw a per topic distribution over s-sets, $\varphi_t \sim Dir(\beta)$
   ii. For each s-set $s \in \{1, \dots, S\}$
      a) Draw a per topic, per s-set distribution over words, $\eta_{t,s} \sim Dir(\gamma)$
2. For each document $m \in \{1, \dots, M\}$
   i. Draw $\theta_m \sim Dir(\alpha)$
   ii. For each word $w_{m,n}$, where $n \in \{1, \dots, N_m\}$
      a) Draw a topic $z_{m,n} \sim Mult(\theta_m)$
      b) Draw an s-set $s_{m,n} \sim Mult(\varphi_{z_{m,n}})$
      c) Emit $w_{m,n} \sim Mult(\eta_{z_{m,n},s_{m,n}})$

The plate notation for MDK-LDA(b) is given in Figure 1. As we will see in Section 3, this simple framework is quite powerful (with the augmentations in Section 2.4). Note that

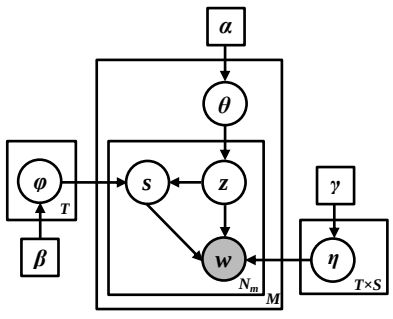

Figure 1. Plate notation of the proposed framework.

this framework has some resemblance to the LDAWN model in [Boyd-Graber *et al*., 2007]. Their synsets are similar to our s-sets. However, their task is word sense disambiguation. LDAWN assumes that a word is generated by a WordNet-Walk in WordNet [Miller, 1995], while in our case, a word is generated by its latent topic and s-set, where each topic is a multinomial distribution over s-sets.

2.2 Setting Concentration for Hyperparameter γ

Hyperparameters α and β are not very sensitive and the heuristic values suggested in [Griffiths and Steyvers, 2004] usually hold well [Wallach *et al*., 2009][1]. However, the hyperparameter γ is crucial as it governs the redistribution of the probability masses over words in each s-set. The intuition here is that in using an s-set from a different domain, we want to be conservative. When an s-set is large, some of its words may not be appropriate for the new domain. In this case, γ is used to concentrate the probability masses on a subset of words in the s-set that are likely to be correct in the new domain (i.e., we prefer sparser distributions). For a smaller s-set, we believe that most of its words are more likely to be correct (i.e., we prefer denser distributions). Thus, we set the hyperparameter γ based on the size of s-set. Based on the property of the Dirichlet concentration parameter that small/large concentration parameter results in sparse/dense distributions, we use the following function:

$$\gamma_s = \lambda \cdot e^{-|s|} \tag{1}$$

γ_s decreases exponentially with the increase in s-set size ($|s|$). We choose an exponential function to control the density of Dirichlet distribution because it suits the phenomenon that the probability of coherent semantic relation decreases exponentially with larger s-sets [Zipf, 1932]. The setting of coefficient λ will be discussed in Section 3.

2.3 Collapsed Gibbs Sampling

In topic models, collapsed Gibbs sampling [Griffiths and Steyvers, 2004] is a standard procedure for obtaining a Markov chain over the latent variables in the model. Given certain conditions, the stationary distribution of the Markov chain is the posterior [Neal, 1993]. In all MDK-LDA models (including MDK-LDA(b) and MDK-LDA), we jointly sample latent variables z and s, which gives us a blocked Gibbs sampler. An alternative way is to perform hierarchical sampling (sample z and then s). However, [Rosen-Zvi *et al*., 2010] argues that when the latent variables are highly related, blocked samplers improve convergence of the Markov chain and also reduce autocorrelation. Denoting the random variables $\{z, s, w\}$ by singular subscripts $\{z_i, s_i, w_i\}$, where i denotes the variable corresponding to each word in each document in the corpus, the Gibbs sampler is given by:

$$P(z_i = t, s_i = s \,|\, \boldsymbol{z}^{-i}, \boldsymbol{s}^{-i}, \boldsymbol{w}, \alpha, \beta, \gamma) \propto \frac{n_{m,t}^{-i} + \alpha}{\sum_{t'=1}^{T}(n_{m,t'}^{-i} + \alpha)} \times \frac{n_{t,s}^{-i} + \beta}{\sum_{s'=1}^{S}(n_{t,s'}^{-i} + \beta)} \times \frac{n_{t,s,w_i}^{-i} + \gamma_s}{\sum_{v'=1}^{V}(n_{t,s,v'}^{-i} + \gamma_s)} \tag{2}$$

[1] Although [Wallach *et al*., 2009] reports that optimizing α improves topics, in our preliminary experiments, we found that γ affects topics more than α. Hence, we focus on setting γ.

The superscript n^{-i} denotes the counts excluding the current assignment of z_i and s_i, i.e., $\boldsymbol{z}^{-i}$ and $\boldsymbol{s}^{-i}$. $n_{m,t}$ denotes the number of times that topic t was assigned to word tokens in document m. $n_{t,s}$ denotes the count that s-set s occurs under topic t. $n_{t,s,v}$ refers to the number of times that word v appears in s-set s under topic t.

In this work, we want to produce coherent topics, i.e., we are interested in the word distribution under each topic. Under the MDK-LDA framework, the word distribution under topic t, i.e., $\pi_t(w)$, can be computed as follows:

$$\pi_t(w) = \sum_{s=1}^{S}\big(\varphi_t(s) \cdot \eta_{t,s}(w)\big) \tag{3}$$

2.4 Generalized Pólya Urn Model

MDK-LDA(b) redistributes the probability masses over words in each s-set using the latent variable s, with a multinomial $\eta_{t,s}$, along with its functional concentration parameter. Due to the power-law characteristics of natural language [Zipf, 1932], most words are rare and will not co-occur with most other words regardless of their semantic similarity. If some rare words share the same s-set with some high-frequency words, the high-frequency words will be smoothed dramatically due to the hyperparameter γ which causes the adverse effect issue discussed in Section 1. For example, in the domain "Camera", the word *light* is an important word and its semantic meaning correlated with the domain. However, the words *brightness* and *luminousness* in the s-set {light, brightness, luminousness} can harm *light* due to their infrequency. Since words in the s-set are supposed to share some similar semantic meaning, if we see one of them, it is reasonable to expect higher probability of seeing any of the others. For the above s-set, if *brightness* is seen in topic t, there is a higher chance of seeing both *light* and *luminousness* under topic t. To encode this characteristic, we use the generalized Pólya urn (GPU) model [Mahmoud, 2008], where objects of interest are represented as colored balls in an urn. In a simple Pólya urn model, when a ball of a particular color is drawn, that ball is put back along with a new ball of the same color. GPU differs in that, when a ball is drawn, that ball is put back along with a certain number of balls of similar colors. In our case, the similarity of colors, which are words, is indicated by the fact that they are from the same s-set.

More formally, having drawn a ball of color c, some additional balls of each color $c' \in \{1, \dots, C\}$ are returned to the urn. In our case, each color c corresponds to each word $w \in \{1, \dots, V\}$. In order to promote an s-set upon observing any of its word, if a ball of color w is drawn, we put back $\mathbb{A}_{s,w',w}$ balls of each color $w' \in \{1, \dots, V\}$ where w and w' share s-set s. $\mathbb{A}_{s,w',w}$ is defined as:

$$\mathbb{A}_{s,w',w} = \begin{cases} 1 & w = w' \\ \sigma & w \in s, w' \in s, w \neq w' \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

As a result, we promote the entire s-set s. Note that the probabilities of the words under the s-set s do not change since the GPU model is only used to promote the entire s-set. Incorporating the GPU model to MDK-LDA(b) gives us the final MDK-LDA model, which addresses the adverse

effect issue discussed in Section 1. MDK-LDA still uses the same plate notation as MDK-LDA(b) in Figure 1 since GPU cannot be reflected in the plate.

Collapsed Gibbs sampling: The GPU model is non-exchangeable, meaning that the joint probability of the words in any given topic is not invariant to the permutation of those words. Inference of $\boldsymbol{z}$ and $\boldsymbol{s}$ can be computationally expensive due to the non-exchangeability of words, i.e., the sampling distribution for the word of interest depends on each possible value for the subsequent words along with their topic and s-set assignments. We take the approach of [Mimno *et al*., 2011] which approximates the true Gibbs sampling distribution by treating each word as if it were the last, ignoring implications of the subsequent words and their topic and s-set assignments. The approximate Gibbs sampler has the following conditional distribution:

$$P(z_i = t, s_i = s | \boldsymbol{z}^{-i}, \boldsymbol{s}^{-i}, \boldsymbol{w}, \alpha, \beta, \gamma, \mathbb{A}) \propto \frac{n_{m,t}^{-i} + \alpha}{\sum_{t'=1}^{T}(n_{m,t'}^{-i} + \alpha)} \times \frac{\sum_{w'=1}^{V}\sum_{v'=1}^{V}\mathbb{A}_{s,v',w'} \cdot n_{t,s,v'}^{-i} + \beta}{\sum_{s'=1}^{S}(\sum_{w'=1}^{V}\sum_{v'=1}^{V}\mathbb{A}_{s',v',w'} \cdot n_{t,s',v'}^{-i} + \beta)} \times \frac{n_{t,s,w_i}^{-i} + \gamma_s}{\sum_{v'=1}^{V}(n_{t,s,v'}^{-i} + \gamma_s)} \quad (5)$$

3 Experiments

We now evaluate the proposed models and compare them with three baselines: LDA [Blei *et al*., 2003], LDA with GPU (denoted as LDA_GPU) [Mimno *et al*., 2011] and DF-LDA [Andrzejewski *et al*., 2009]. LDA is the basic knowledge free unsupervised topic model. LDA_GPU applies GPU in LDA using co-document frequency. DF-LDA is perhaps the most well-known knowledge-based topic model which introduced must-links and cannot-links. It is also a natural fit for the proposed models as a must-link and an s-set share the similar notion, i.e., they both constrain the words in them to belong to the same topic.

3.1 Datasets and Settings

Datasets: We use product reviews from six domains from Amazon.com for evaluation. Each domain corpus consists of 500 reviews, as shown in Table 1 (columns 2 and 3). The domains are "Watch," "Camera," "Cellphone," "Computer," "Food" and "Care" (short for "Personal Care").

Pre-processing: Punctuations, stopwords[2], numbers and words appearing less than 5 times in each corpus were removed. The domain name was also removed, e.g., word *camera* in the domain "Camera", since it appears frequently and co-occurs with most words in the corpus, leading to high similarity among topics. We then ran the Stanford Parser[3] to perform sentence detection and lemmatization.

Sentences as documents: As noted in [Titov and McDonald, 2008], the main topics in every review for a particular type of products are basically the same. Thus, when LDA is applied to such a collection of reviews, it typically finds topics of mainly product or brand names. However, applying topic models to reviews mainly aims to find different aspects or features of products [Jo and Oh, 2011; Mukherjee and Liu, 2012; Titov and McDonald, 2008; Zhao *et al*., 2010]. Thus, using individual reviews for modeling is not effective [Titov and McDonald, 2008]. Although there are models dealing with sentences in complex ways [Jo and Oh, 2011; Titov and McDonald, 2008], we take a simple approach: We divide each review into sentences and each sentence is treated as an independent document. Sentences can be used by all three baselines without any change to their models. Although the relationship between sentences in the review is lost, the data is fair for all systems.

Domain	#Reviews	#Sentences	#s-sets
Watch	500	2712	91
Camera	500	5171	173
Cellphone	500	2527	61
Computer	500	2864	92
Food	500	2416	85
Care	500	3008	119
Average	500	3116	103

Table 1. Corpus statistics with #s-sets having at least two words.

Parameter settings: All models were trained using 1000 Gibbs iterations with an initial burn-in of 100 iterations. For all models, we set $\alpha = 1$ and $\beta = 0.1$. We found that small changes of α and β did not affect the results much, which was also found in [Jo and Oh, 2011] that also used online reviews. We use $T = 15$ for every domain. Note that it is difficult to know the exact number of topics. While non-parametric Bayesian approaches [Teh *et al*., 2006] aim to estimate T from the corpus, in this work the heuristic value obtained from our initial experiments produced good results.

For DF-LDA, we convert s-sets to must-links (we do not use cannot-links). Since the definition of must-links is transitive, meaning (must-link (u, v) and must-link (v, w) imply must-link (u, w)), we merge s-sets that require merging due to transitivity. We then ran DF-LDA (downloaded from its authors' website) while keeping the parameters as proposed in [Andrzejewski *et al*., 2009] (we also experimented with different parameter settings but they did not produce better results). For our new models, we set $\lambda = 2000$ after simulation experiments in *R* package (http://www.r-project.org/). For σ in equation 4, we empirically set it to 0.2. It is also interesting to study the sensitivity of λ and σ, which we defer to our future work.

Domain knowledge: User knowledge about a domain can vary a great deal, meaning different users may have very different knowledge. For our experiments, we want to reduce this variance. Instead of asking a human user to provide knowledge, we obtain the synonym sets and the antonym sets for each word that is a noun, verb, or adjective (as words of other parts-of-speech usually do not indicate topics) from WordNet [Miller, 1995] and manually remove those sets whose words should not be in a topic for the domain. The number of remaining s-sets is listed in the last column of Table 1. Duplicate s-sets have been removed.

Experiment sequence: First, we would like to see if the model can leverage user knowledge from multiple past domains to help topic modeling in the new domains. We choose the following experiment sequence of domains: Watch → Camera → Cellphone → Computer → Food → Care. Modeling in each domain uses the domain knowledge of all earlier domains. We choose this sequence because:

[2] http://jmlr.csail.mit.edu/papers/volume5/lewis04a/

[3] http://nlp.stanford.edu/software/corenlp.shtml

p@n	No Knowledge		PrK			DoK			PrK+DoK		
	LDA	LDA_GPU	DF-LDA	MDK-LDA(b)	MDK-LDA	DF-LDA	MDK-LDA(b)	MDK-LDA	DF-LDA	MDK-LDA(b)	MDK-LDA
p@5	0.77	0.60	0.85	0.78	**0.85**	0.84	0.87	**0.98**	0.62	0.78	**1.00**
p@10	0.68	0.70	0.66	0.73	**0.79**	0.79	0.80	**0.83**	0.62	0.75	**0.90**
p@15	0.64	0.53	0.54	0.71	**0.74**	0.65	0.73	**0.76**	0.52	0.66	**0.85**
p@20	0.58	0.55	0.50	0.64	**0.65**	0.56	0.64	**0.66**	0.47	0.59	**0.75**
Average	0.67	0.60	0.64	0.72	**0.76**	0.71	0.76	**0.81**	0.56	0.70	**0.88**

Table 2. *Precision @ n* (*p@n*) across good topics for each model with different knowledge sets for the domain "Computer".

	No Knowledge				PrK						DoK						PrK+DoK					
	LDA		LDA_GPU		DF-LDA		MDK-LDA(b)		MDK-LDA		DF-LDA		MDK-LDA(b)		MDK-LDA		DF-LDA		MDK-LDA(b)		MDK-LDA	
Domains	Prec	#T	Prec	#T	Prec	#T	Prec	#T	Prec	#T	Prec	#T	Prec	#T	Prec	#T	Prec	#T	Prec	#T	Prec	#T
Camera	0.80	11	0.50	3	0.61	8	0.82	11	**0.87**	**13**	0.64	9	0.79	12	**0.91**	**13**	0.67	6	0.81	12	**0.93**	**13**
Computer	0.67	6	0.60	4	0.64	8	0.72	10	**0.76**	**11**	0.71	9	0.76	11	**0.81**	**11**	0.56	7	0.70	11	**0.88**	**11**
Food	0.87	7	0.61	5	0.80	4	0.90	7	**0.92**	**7**	0.78	5	0.86	8	**0.93**	**9**	0.67	4	0.84	8	**0.91**	**8**
Care	0.81	9	0.64	3	0.77	8	0.80	10	**0.84**	**12**	0.73	10	0.88	11	**0.88**	**12**	0.72	7	0.92	10	**0.91**	**11**
Average	0.79	8.25	0.59	3.75	0.71	7	0.81	9.5	**0.85**	**10.75**	0.72	8.25	0.82	10.5	**0.88**	**11.25**	0.66	6	0.82	10.25	**0.91**	**10.75**

Table 3. Average precision of each domain and number of good topics of each model with different knowledge sets.

1. The first four domains have some similarities and we want to see whether s-sets from previous similar domains can help new domains.
2. The "Food" domain is very different from the first four domains. We want to see whether the previous s-sets can harm this domain. The "Care" domain has slight intersection of topics with other domains and we want to see whether the previous s-sets are useful.

We will give the topic discovery results of the domains: Camera, Computer, Food and Care. For each domain, we show and compare three types of topical results:

1. Those obtained using only knowledge (or s-sets) from previous domains and no knowledge of its own domain. We denote these results as PrK (Previous Knowledge).
2. Those obtained using only its own domain knowledge. We denote this set of results as DoK (its own Domain Knowledge).
3. Those using both its own knowledge and the knowledge accumulated from previous domains. We denote this set of results as PrK+DoK.

Section 3.2 follows the proposed sequence. Further, in Section 3.3, we report topic coherence [Mimno *et al*., 2011] for each domain using knowledge from all other domains (thus making the results invariant to the sequence of domains used for knowledge accumulation).

3.2 Topic Discovery Results

Quantitative Results

For evaluation of the discovered topics, we used two human judges fluent in English and with good knowledge of the domains to label every topic generated by each model. For each topic, we choose top 20 words ranked by per-topic word distribution (π_t in equation 3). Each word in a topic is considered correct if both judges agree; otherwise the word is labeled as incorrect. The models which generated the topics for labeling were oblivious to the judges. Computing Kappa is difficult because labeling involves both topics and topical words in each topic and they are also dependent on each other. We thus choose to base the labeling on consensus. Since topics from topic models are rankings based on word probability and we do not know the number of correct topical words, a natural way to evaluate these rankings is to use *Precision @ n* (or *p@n*) which was also used in [Mukherjee and Liu, 2012; Zhao *et al*., 2010], where *n* is the rank position. We give *p@n* for *n* = 5, 10, 15 and 20.

In general, for each model, some discovered topics are good or comprehensible and can be labeled with topic names, while others are semantically messy and no coherent topic can be identified. We define a topic as *good* (or comprehensible) if it meets the following three criteria:

1. There is an agreed topic label assigned to the topic by the judges.
2. At least one of *p@5*, *p@10*, *p@15* and *p@20* is greater than 50%.
3. When multiple topics have the same topic label (given by the judges), the topic that has the highest average precision (over all *p@n*) is assigned to this label. This is to take care of the case when a single semantic topic is split into multiple topics with many errors. In this case, we simply select the best topic for the label. The rest of the topics will be considered for other labels.

We now compare results using knowledge from previous domains (PrK), using knowledge from its own domain (DoK) and using both, PrK+DoK. Due to space constraints, we only show the detailed comparison results of one domain "Computer" in Table 2, which gives the average *precision @ n* over *all good topics for each model* under different sets of knowledge. For other domains, we show the average results of *precision @ 5, 10, 15 and 20* (the same as the last row of Table 2) in Table 3. Based on Table 2, we can make the following observations:

1. MDK-LDA consistently outperforms all other models across all three sets of knowledge, i.e. PrK, DoK, and PrK+DoK. It is also clear that DoK and PrK+DoK are better than PrK. PrK+DoK is also better than DoK alone, which shows that even with its own knowledge in the new domain, previous knowledge can be quite useful.
2. MDK-LDA(b) also improves over LDA, LDA_GPU and DF-LDA. But it does not perform as well as MDK-LDA due to the adverse effect issue discussed in Section 1.
3. DF-LDA does better than LDA given DoK. However, when combined with PrK (e.g., PrK+DoK), it does not perform well due to the two issues discussed Section 1.
4. LDA_GPU does not perform well in our review data due to its use of co-document frequency. As frequent words usually have high co-document frequency with many other words, the frequent words are ranked top in many topics. This shows that user knowledge is more effective than co-document frequency without knowledge.

Camera (Battery)		Computer (Price)		Food (Taste)		Care (Tooth)	
LDA	MDK	LDA	MDK	LDA	MDK	LDA	MDK
battery	extra	*acer*	cheap	taste	flavor	*price*	tooth
screen	charge	*power*	price	salt	sweet	tooth	gum
life	life	*base*	inexpensive	*almond*	sugar	*amazon*	dentist
lcd	replacement	*year*	money	*fresh*	salty	pen	dental
water	battery	*button*	expensive	*pack*	tasty	*shipping*	whitening
usb	charger	*amazon*	cost	tasty	tasting	gum	pen
cable	aa	*control*	dollar	*oil*	delicious	dentist	refill
e	power	price	buck	*roasted*	taste	whitening	*year*
charger	rechargeable	*color*	worth	pepper	salt	refill	*date*
hour	time	purchase	low	*easy*	spice	*worth*	*product*

Table 4. Example topics (MDK is short for MDK-LDA).

Table 3 shows the average precision (Prec.) result for each domain (the corresponding result of the domain "Computer" in the last row of Table 2), and the number of good topics discovered by each model (#T). On the precision results, we can see the same trends. MDK-LDA consistently performs better across different types of knowledge compared with other models. It has a higher average precision and also discovers more good topics. MDK-LDA using PrK+DoK is slightly inferior than using only DoK for domain "Food" and "Care" as for them previous domains and their knowledge (i.e. PrK) are quite different, which add some noise. MDK-LDA(b) performs well too, but not as well as MDK-LDA. From *precision @ n* in Table 3 over each setting of knowledge, MDK-LDA(b) performs significantly better than the three baseline models ($p < 0.02$). The improvements of MDK-LDA are also significant over baseline models ($p < 0.0001$) and MDK-LDA(b) ($p < 0.002$). A paired *t*-test was used for testing significance.

Qualitative Results

In this section, we show some qualitative results to give an intuitive feeling of the results from different models. There are a large number of topics that are dramatically improved by the proposed models. Due to the space limitations, we can only show some examples. To further focus, we will just show some results of LDA and MDK-LDA. The results from LDA_GPU and DF-LDA were inferior and they were even hard for the human judges to manually label topics and to match them with topics found by the other models. By no means do we say that LDA_GPU and DF-LDA are not effective. We are only saying that in our problem setting (of using knowledge from multiple past domains) and review data, these models do not generate as coherent topics as ours because they cannot effectively leverage prior knowledge.

Table 4 shows one example topic for each domain. "Camera (Battery)" means topic *Battery* in domain "Camera". The setting of knowledge in Table 4 is PrK+DoK. Wrong topical words are in italic and marked red (we tried to find the best possible topic matches for the models). We can see that MDK-LDA produces much better topics. For example, the s-set {price, cheap, expensive} in the domain "Camera" improved the topic *Price* in "Computer" (see Table 4 column "Computer (Price)"). LDA did not perform well here (reported is the best *Price* topic that we can find for LDA). Since the labeling of topics and topical words are subjective, we do not expect everyone to agree with us, but we tried our best to have the consensus of two human judges. We will also show the objective results below.

Domain	LDA	LDA_GPU	DF-LDA	MDK-LDA(b)	MDK-LDA
Watch	-1326.58	-1442.96	-1297.53	-1265.39	-1230.68
Camera	-1434.32	-1509.66	-1490.47	-1423.94	-1391.48
Cellphone	-1299.56	-1493.45	-1262.34	-1255.44	-1225.86
Computer	-1351.12	-1514.18	-1300.02	-1282.64	-1244.99
Food	-1228.91	-1500.98	-1250.61	-1193.92	-1188.64
Care	-1240.55	-1538.10	-1257.36	-1240.74	-1220.37
Average	-1313.51	-1499.89	-1309.72	-1277.01	-1250.34

Table 5. Topic Coherence across all domains and all models.

In summary, we can say that MDK-LDA produces much better results than the baselines, which demonstrates its effectiveness in exploiting knowledge from other domains.

3.3 Topic Coherence

Apart from quantitative and qualitative evaluation as above, topic models are often also evaluated using perplexity on held-out test data. However, as shown in [Chang *et al*., 2009; Newman *et al*., 2010], the perplexity measure does not reflect the semantic coherence of individual topics and can be contrary to human judgments. The *topic coherence* measure [Mimno *et al*., 2011] was proposed as a better alternative for assessing topic quality. The measure only relies upon word co-occurrence statistics within the documents, and does not depend on external resources or human labeling. It was also shown that topic coherence is highly consistent with human expert labeling [Mimno *et al*., 2011]. Higher topic coherence indicates higher quality of topics. We follow [Mimno *et al*., 2011] to calculate topic coherence. The experiment setting here is as follows: for each domain, we apply knowledge from all other 5 domains, e.g., the results for "Watch" used all knowledge except those of domain "Watch." This setting fully reflects how knowledge from other domains can aid modeling in the new domain where the knowledge accumulation is invariant to the sequence of domains. Table 5 shows the average topic coherence (over all 15 topics) for each domain. A paired *t*-test indicates that MDK-LDA(b) is significantly better than baseline models ($p < 0.02$), and MDK-LDA outperforms all models significantly ($p < 0.005$).

4 Conclusions

This paper proposed a novel framework to exploit prior knowledge from multiple past domains for producing better topics in the new domain. To the best of our knowledge, this has not been done before. To perform the task, the paper identified two key technical difficulties with existing knowledge-based models, i.e., multiple senses and adverse effect. A new model called MDK-LDA was then proposed to deal with the problems. Our evaluation results show that MDK-LDA outperforms the baselines significantly. We also believe that the proposed framework is not only useful in practice but also valuable to machine learning because human beings learn knowledge over time and past knowledge is not discarded but used to solve new problems.

Acknowledgments: This work was supported in part by a grant from National Science Foundation (NSF) under grant no. IIS-1111092, and a grant from HP Labs Innovation Research Program.

References

[Andrzejewski *et al*., 2009] David Andrzejewski, Xiaojin Zhu, and Mark Craven. Incorporating domain knowledge into topic modeling via Dirichlet Forest priors. In *ICML*, pages 25–32, 2009.

[Andrzejewski *et al*., 2011] David Andrzejewski, Xiaojin Zhu, Mark Craven, and Benjamin Recht. A framework for incorporating general domain knowledge into latent Dirichlet allocation using first-order logic. In *IJCAI*, pages 1171–1177, 2011.

[Bart, 2011] Evgeniy Bart. Improving performance of topic models by variable grouping. In *IJCAI*, pages 1178–1185, 2011.

[Blei *et al*., 2003] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3:993–1022, 2003.

[Boyd-Graber *et al*., 2007] Jordan Boyd-Graber, Blei David, and Zhu Xiaojin. A Topic Model for Word Sense Disambiguation. In *EMNLP*, pages 1024–1033, 2007.

[Burns *et al*., 2012] Nicola Burns, Yaxin Bi, Hui Wang, and Terry Anderson. Extended Twofold-LDA Model for Two Aspects in One Sentence. In *Advances in Computational Intelligence*, vol. 298, pages 265–275, 2012.

[Chang *et al*., 2009] Jonathan Chang, Jordan Boyd-Graber, Wang Chong, Sean Gerrish, and David M. Blei. Reading Tea Leaves: How Humans Interpret Topic Models. In *NIPS*, pages 288–296, 2009.

[Faisal *et al*., 2012] Ali Faisal, Jussi Gillberg, Jaakko Peltonen, Gayle Leen, and Samuel Kaski. Sparse Nonparametric Topic Model for Transfer Learning. In *ESANN*, 2012.

[Griffiths and Steyvers, 2004] Thomas L. Griffiths, and Mark Steyvers. Finding Scientific Topics. *PNAS* 101 Suppl:5228–5235, 2004.

[Heinrich, 2009] Gregor Heinrich. A Generic Approach to Topic Models. In *ECML PKDD*, pages 517 – 532, 2009

[Hofmann, 1999] Thomas Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57, 1999.

[Hu *et al*., 2011] Yuening Hu, Jordan Boyd-Graber, and Brianna Satinoff. Interactive Topic Modeling. In *ACL*, pages 248–257, 2011.

[Jagarlamudi *et al*., 2012] Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. Incorporating Lexical Priors into Topic Models. In *EACL*, pages 204–213, 2012.

[Jo and Oh, 2011] Yohan Jo, and Alice H. Oh. Aspect and sentiment unification model for online review analysis. In *WSDM*, pages 815–824, 2011.

[Lu *et al*., 2011] Bin Lu, Myle Ott, Claire Cardie, and Benjamin K. Tsou. Multi-aspect Sentiment Analysis with Topic Models. In *ICDM Workshops*, pages 81–88, 2011.

[Mahmoud, 2008] Hosam Mahmoud. *Polya Urn Models*. Chapman & Hall/CRC Texts in Statistical Science, 2008.

[Mei *et al*., 2007] Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *WWW*, pages 171–180, 2007.

[Miller, 1995] George A. Miller. WordNet: A Lexical Database for English. *Commun. ACM* 38(11):39–41, 1995

[Mimno *et al*., 2011] David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *EMNLP*, pages 262–272, 2011.

[Mukherjee and Liu, 2012] Arjun Mukherjee, and Bing Liu. Aspect Extraction through Semi-Supervised Modeling. In *ACL*, pages 339–348, 2012.

[Neal, 1993] R. M. Neal. *Probabilistic Inference Using Markov Chain Monte Carlo Methods*, 1993.

[Newman *et al*., 2010] David Newman, Youn Noh, Edmund Talley, Sarvnaz Karimi, and Timothy Baldwin. Evaluating topic models for digital libraries. In *JCDL*, Pages 215-224, 2010.

[Pan and Yang, 2010] Sinno Jialin Pan, and Qiang Yang. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* 22(10):1345–1359, 2010.

[Petterson *et al*., 2010] James Petterson, Alex Smola, Tibério Caetano, Wray Buntine, and Shravan Narayanamurthy. Word Features for Latent Dirichlet Allocation. In *NIPS*, pages 1921–1929, 2010.

[Rosen-Zvi *et al*., 2010] Michal Rosen-Zvi, Chaitanya Chemudugunta, Thomas Griffiths, Padhraic Smyth, and Mark Steyvers. Learning author-topic models from text corpora. *ACM Transactions on Information Systems* 28(1):1–38, 2010.

[Teh *et al*., 2006] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association* 1–30, 2006.

[Titov and McDonald, 2008] Ivan Titov, and Ryan McDonald. Modeling online reviews with multi-grain topic models. In *WWW*, pages 111–120, 2008.

[Wallach *et al*., 2009] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why Priors Matter. In *NIPS*, pages 1973–1981, 2009.

[Waltinger *et al*., 2011] Ulli Waltinger, Alexa Breuing, and Ipke Wachsmuth. Interfacing virtual agents with collaborative knowledge: Open domain question answering using wikipedia-based topic models. In *IJCAI*, pages 1896–1902, 2011.

[Wang *et al*., 2010] Hongning Wang, Yue Lu, and Chengxiang Zhai. Latent aspect rating analysis on review text data: a rating regression approach. In *KDD*, pages 783–792, 2010.

[Xue *et al*., 2008] Gui-Rong Xue, Wenyuan Dai, Qiang Yang, and Yong Yu. Topic-bridged PLSA for cross-domain text classification. In *SIGIR*, pages 627–634, 2008.

[Zhao *et al*., 2010] Wayne Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. Jointly Modeling Aspects and Opinions with a MaxEnt-LDA Hybrid. In *EMNLP*, pages 56–65, 2010.

[Zipf, 1932] George K. Zipf. *Selective Studies and the Principle of Relative Frequency in Language*. Harvard University Press, 1932.