Incorporating Word Correlation Knowledge into Topic Modeling

Pengtao Xie
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
pengtaox@cs.cmu.edu

Diyi Yang
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
diyiy@cs.cmu.edu

Eric P. Xing
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
epxing@cs.cmu.edu

Abstract

This paper studies how to incorporate the external word correlation knowledge to improve the coherence of topic modeling. Existing topic models assume words are generated independently and lack the mechanism to utilize the rich similarity relationships among words to learn coherent topics. To solve this problem, we build a Markov Random Field (MRF) regularized Latent Dirichlet Allocation (LDA) model, which defines a MRF on the latent topic layer of LDA to encourage words labeled as similar to share the same topic label. Under our model, the topic assignment of each word is not independent, but rather affected by the topic labels of its correlated words. Similar words have better chance to be put into the same topic due to the regularization of MRF, hence the coherence of topics can be boosted. In addition, our model can accommodate the subtlety that whether two words are similar depends on which topic they appear in, which allows word with multiple senses to be put into different topics properly. We derive a variational inference method to infer the posterior probabilities and learn model parameters and present techniques to deal with the hard-to-compute partition function in MRF. Experiments on two datasets demonstrate the effectiveness of our model.

1 Introduction

Probabilistic topic models (PTM), such as probabilistic latent semantic indexing(PLSI) (Hofmann, 1999) and latent Dirichlet allocation(LDA) (Blei et al., 2003) have shown great success in documents modeling and analysis. Topic models posit document collection exhibits multiple latent semantic topics where each topic is represented as a multinomial distribution over a given vocabulary and each document is a mixture of hidden topics. To generate a document d, PTM first samples a topic proportion vector, then for each word w in d, samples a topic indicator z and generates w from the topic-word multinomial corresponding to topic z.

A key limitation of the existing PTMs is that words are assumed to be uncorrelated and generated independently. The topic assignment for each word is irrelevant to all other words. While this assumption facilitates computational efficiency, it loses the rich correlations between words. In many applications, users have external knowledge regarding word correlation, which can be taken into account to improve the semantic coherence of topic modeling. For example, WordNet (Miller, 1995a) presents a large amount of synonym relationships between words, Wikipedia[1] provides a knowledge graph by linking correlated concepts together and named entity recognizer identifies the categories of entity mentions. All of these external knowledge can be leveraged to learn more coherent topics if we can design a mechanism to encourage similar words, correlated concepts, entities of the same category to be assigned to the same topic.

Many approaches (Andrzejewski et al., 2009; Petterson et al., 2010; Newman et al., 2011) have attempted to solve this problem by enforcing hard and topic-independent rules that similar words should have similar probabilities in all topics, which is

[1] https://www.wikipedia.org/

questionable in that two words with similar representativeness of one topic are not necessarily of equal importance for another topic. For example, in the *fruit* topic, the words *apple* and *orange* have similar representativeness, while in an *IT company* topic, *apple* has much higher importance than *orange*. As another example, $church$ and $bible$ are similarly relevant to a $religion$ topic, whereas their relevance to an $architecture$ topic are vastly different. Exiting approaches are unable to differentiate the subtleties of word sense across topics and would falsely put irrelevant words into the same topic. For instance, since *orange* and *microsoft* are both labeled as similar to *apple* and are required to have similar probabilities in all topics as *apple* has, in the end, they will be unreasonably allocated to the same topic.

The existing approaches fail to properly use the word correlation knowledge, which is usually a list of word pairs labeled as *similar*. The similarity is computed based on statistics such as co-occurrence which are unable to accommodate the subtlety that whether two words labeled as similar are truly similar depends on which topic they appear in, as explained by the aforementioned examples. Ideally, the knowledge would be word A and B are similar under topic C. However, in reality, we only know two words are similar, but not under which topic. In this paper, we aim to abridge this gap. Gaining insights from (Verbeek and Triggs, 2007; Zhao et al., 2010; Zhu and Xing, 2010), we design a Markov Random Field regularized LDA model (MRF-LDA) which utilizes the external knowledge in a soft and topic-dependent manner to improve the coherence of topic modeling. We define a MRF on the latent topic layer of LDA to encode word correlations. Within a document, if two words are labeled as similar according to the external knowledge, their latent topic nodes will be connected by an undirected edge and a binary potential function is defined to encourage them to share the same topic label. This mechanism gives correlated words a better chance to be put into the same topic, thereby, improves the coherence of the learned topics. Our model provides a mechanism to automatically decide under which topic, two words labeled as similar are truly similar. We encourage words labeled as similar to share the same topic label, but do not specify which topic label they should share, and leave this to be decided by data. In the above mentioned *apple*, *orange*, *microsoft* example, we encourage *apple* and *orange* to share the same topic label A and try to push *apple* and *microsoft* to the same topic B. But A and B are not necessarily the same and they will be inferred according to the fitness of data. Different from the existing approaches which directly use the word similarities to control the topic-word distributions in a hard and topic-independent way, our method imposes constraints on the latent topic layer by which the topic-word multinomials are influenced indirectly and softly and are topic-aware.

The rest of the paper is organized as follows. In Section 2, we introduce related work. In Section 3, we propose the MRF-LDA model and present the variational inference method. Section 4 gives experimental results. Section 5 concludes the paper.

2 Related Work

Different from purely unsupervised topics models that often result in incoherent topics, knowledge based topic models enable us to take prior knowledge into account to produce more meaningful topics. Various approaches have been proposed to exploit the correlations and similarities among words to improve topic modeling instead of purely relying on how often words co-occur in different contexts (Heinrich, 2009). For instance, Andrzejewski et al. (2009) imposes Dirichlet Forest prior over the topic-word multinomials to encode the Must-Links and Cannot-Links between words. Words with Must-Links are encouraged to have similar probabilities within all topics while those with Cannot-Links are disallowed to simultaneously have large probabilities within any topic. Similarly, Petterson et al. (2010) adopted word information as features rather than as explicit constraints and defined a prior over the topic-word multinomials such that similar words share similar topic distributions. Newman et al. (2011) proposed a quadratic regularizer and a convolved Dirichlet regularizer over topic-word multinomials to incorporate the correlation between words. All of these methods directly incorporate the word correlation knowledge into the topic-word distributions in a hard and topic-independent way, which ignore the fact that whether two words are

correlated depends on which topic they appear in.

There are several works utilizing knowledge with more complex structure to improve topic modeling. Boyd-Graber et al. (2007) incorporate the synset structure in WordNet (Miller, 1995b) into LDA for word sense disambiguation, where each topic is a random process defined over the synsets. Hu et al. (2011) proposed interactive topic modeling, which allows users to iteratively refine the discovered topics by adding constraints such as certain set of words must appear together in the same topic. Andrzejewski et al. (2011) proposed a general framework which uses first order logic to encode various domain knowledge regarding documents, topics and side information into LDA. The vast generality and expressivity of this model makes its inference to be very hard. Chen et al. (2013) proposed a topic model to model multi-domain knowledge, where each document is an admixture of latent topics and each topic is a probability distribution over domain knowledge. Jagarlamudi et al. (2012) proposed to guide topic modeling by setting a set of seed words in the beginning that user believes could represent certain topics. While these knowledge are rich in structure, they are hard to acquire in the real world applications. In this paper, we focus on pairwise word correlation knowledge which are widely attainable in many scenarios.

In the domain of computer vision, the idea of using MRF to enforce topical coherence between neighboring patches or superpixels has been exploited by several works. Verbeek and Triggs (2007) proposed Markov field aspect model where each image patch is modeled using PLSA (Hofmann, 1999) and a Potts model is imposed on the hidden topic layer to enforce spatial coherence. Zhao et al. (2010) proposed topic random field model where each superpixel is modeled using a combination of LDA and mixture of Gaussian model and a Potts model is defined on the topic layer to encourage neighboring superpixels to share the same topic. Similarly, Zhu and Xing (2010) proposed a conditional topic random field to incorporate features about words and documents into topic modeling. In their model, the MRF is restricted to be a linear chain, which can only capture the dependencies between neighboring words and is unable to incorporate long range word correlations. Different from these works, the MRF in our model is not restricted to Potts or chain structure. Instead, its structure is decided by the word correlation knowledge and can be arbitrary.

3 Markov Random Field Regularized Latent Dirichlet Allocation

In this section, we present the MRF-LDA model and the variational inference technique.

3.1 MRF-LDA

We propose the MRF-LDA model to incorporate word similarities into topic modeling. As shown in Figure 1, MRF-LDA extends the standard LDA model by imposing a Markov Random Field on the latent topic layer. Similar to LDA, we assume a document possesses a topic proportion vector θ sampled from a Dirichlet distribution. Each topic $\boldsymbol{\beta}_k$ is a multinomial distribution over words. Each word w has a topic label z indicating which topic w belongs to.

In many scenarios, we have access to external knowledge regarding the correlations between words, such as *apple* and *orange* are similar, *church* and *bible* are semantically related. These similarity relationships among words can be leveraged to improve the coherence of learned topics. To do this, we define a Markov Random Field over the latent topic layer. Given a document d containing N words $\{w_i\}_{i=1}^N$, we examine each word pair (w_i, w_j). If they are correlated according to the external knowledge, we create an undirected edge between their topic labels (z_i, z_j). In the end, we obtain an undirected graph G where the nodes are latent topic labels $\{z_i\}_{i=1}^N$ and edges connect topic labels of correlated words. In the example shown in Figure 1, G contains five nodes z_1, z_2, z_3, z_4, z_5 and four edges connecting $(z_1, z_3), (z_2, z_5), (z_3, z_4), (z_3, z_5)$.

Given the undirected graph G, we can turn it into a Markov Random Field by defining unary potentials over nodes and binary potentials over edges. We define the unary potential for z_i as $p(z_i|\boldsymbol{\theta})$, which is a multinomial distribution parameterized by $\boldsymbol{\theta}$. In standard LDA, this is how a topic is sampled from the topic proportion vector. For binary potential, with the goal to encourage similar words to have similar topic assignments, we define the edge potential between (z_i, z_j) as $\exp\{\mathbb{I}(z_i = z_j)\}$, where $\mathbb{I}(\cdot)$ is the indicator function. This potential func-

tion yields a larger value if the two topic labels are the same and a smaller value if the two topic labels are different. Hence, it encourages similar words to be assigned to the same topic. Under the MRF model, the joint probability of all topic assignments $\mathbf{z} = \{z_i\}_{i=1}^N$ can be written as

$$p(\mathbf{z}|\boldsymbol{\theta}, \lambda) = \frac{1}{A(\boldsymbol{\theta},\lambda)} \prod_{i=1}^{N} p(z_i|\boldsymbol{\theta}) \exp\{\lambda \sum_{(m,n)\in\mathcal{P}} \mathbb{I}(z_m = z_n)\} \tag{1}$$

where $\mathcal{P}$ denotes the edges in G and $A(\boldsymbol{\theta}, \lambda)$ is the partition function

$$A(\boldsymbol{\theta}) = \sum_{\mathbf{z}} \prod_{i=1}^{N} p(z_i|\boldsymbol{\theta}) \exp\{\lambda \sum_{(m,n)\in\mathcal{P}} \mathbb{I}(z_m = z_n)\} \tag{2}$$

We introduce $\lambda \geq 0$ as a trade-off parameter between unary potential and binary potential. In standard LDA, topic label z_i only depends on topic proportion vector $\boldsymbol{\theta}$. In MRF-LDA, z_i not only depends on $\boldsymbol{\theta}$, but also depends on the topic labels of similar words. If γ is set to zero, the correlation between words is ignored and MRF-LDA is reduced to LDA. Given the topic labels, the generation of words is the same as LDA. w_i is generated from the topic-words multinomial distribution $\boldsymbol{\beta}_{z_i}$ corresponding to z_i.

In MRF-LDA, the generative process of a document is summarized as follows:

- Draw a topic proportion vector $\boldsymbol{\theta} \sim Dir(\boldsymbol{\alpha})$
- Draw topic labels $\mathbf{z}$ for all words from the joint distribution defined in Eq.(1)
- For each word w_i, drawn $w_i \sim multi(\boldsymbol{\beta}_{z_i})$

Accordingly, the joint distribution of $\boldsymbol{\theta}, \mathbf{z}$ and $\mathbf{w}$ can be written as

$$p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda) = p(\boldsymbol{\theta}|\boldsymbol{\alpha}) p(\mathbf{z}|\boldsymbol{\theta}, \lambda) \prod_{i=1}^{N} p(w_i|z_i, \boldsymbol{\beta}) \tag{3}$$

3.2 Variational Inference and Parameter Learning

The key inference problem we need to solve in MRF-LDA is to compute the posterior $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w})$ of latent variables $\boldsymbol{\theta}, \mathbf{z}$ given observed data $\mathbf{w}$. As in LDA (Blei et al., 2003), exact computation is intractable. What makes things even challenging in MRF-LDA is that, an undirected MRF is coupled with a directed LDA and the hard-to-compute partition function of MRF makes the posterior inference and parameter learning very difficult. To solve this problem, we resort to variational inference (Wainwright and Jordan, 2008), which uses a easy-to-handle variational distribution to approximate the true posterior of latent variables. To deal with the partition function in MRF, we seek lower bound of the variational lower bound to achieve tractability. We introduce a variational distribution

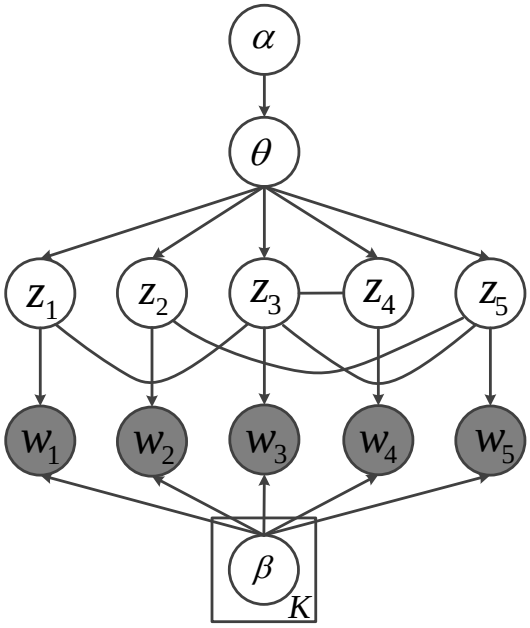

Figure 1: Markov Random Field Regularized Latent Dirichlet Allocation Model

$$q(\boldsymbol{\theta}, \mathbf{z}) = q(\boldsymbol{\theta}|\boldsymbol{\eta}) \prod_{i=1}^{N} q(z_i|\boldsymbol{\phi}_i) \tag{4}$$

where Dirichlet parameter $\boldsymbol{\eta}$ and multinomial parameters $\{\boldsymbol{\phi}_i\}_{i=1}^N$ are free variational parameters. Using Jensen's inequality (Wainwright and Jordan, 2008), we can obtain a variational lower bound

$$\begin{aligned} \mathcal{L} &= \mathbb{E}_q[\log p(\boldsymbol{\theta}|\boldsymbol{\alpha})] + \mathbb{E}_q[\log p(\mathbf{z}|\boldsymbol{\theta}, \lambda)] \\ &+ \mathbb{E}_q[\log \prod_{i=1}^{N} p(w_i|z_i, \boldsymbol{\beta})] - \mathbb{E}_q[\log q(\boldsymbol{\theta}|\boldsymbol{\eta})] \\ &- \mathbb{E}_q[\log \prod_{i=1}^{N} q(z_i|\boldsymbol{\phi}_i)] \end{aligned} \tag{5}$$

in which $\mathbb{E}_q[\log p(\mathbf{z}|\boldsymbol{\theta}, \lambda)]$ can be expanded as

$$\begin{aligned}&\mathbb{E}_q[\log p(\mathbf{z}|\boldsymbol{\theta}, \lambda)]\\&= -\mathbb{E}_q[\log A(\boldsymbol{\theta}, \lambda)] + \lambda \sum_{(m,n)\in\mathcal{P}} \sum_{k=1}^{K} \phi_{mk}\phi_{nk}\\&+ \sum_{i=1}^{N}\sum_{k=1}^{K} \phi_{ik}(\Psi(\eta_k) - \Psi(\sum_{j=1}^{K}\eta_j))\end{aligned} \tag{6}$$

The item $\mathbb{E}_q[\log A(\boldsymbol{\theta}, \lambda)]$ involves the hard-to-compute partition function, which has no analytical expressions. We discuss how to deal with it in the sequel. With Taylor expansion, we can obtain an upper bound of $\mathbb{E}_q[\log A(\boldsymbol{\theta}, \lambda)]$

$$\mathbb{E}_q[\log A(\boldsymbol{\theta}, \lambda)] \leq c^{-1}\mathbb{E}_q[A(\boldsymbol{\theta}, \lambda)] - 1 + \log c \tag{7}$$

where $c \geq 0$ is a new variational parameter. $\mathbb{E}_q[A(\boldsymbol{\theta}, \lambda)]$ can be further upper bounded as

$$\begin{aligned}&\mathbb{E}_q[\log A(\boldsymbol{\theta}, \lambda)] \leq \exp\{\sum_{(m,n)\in\mathcal{P}} \lambda\}\\&\sum_{n_1,n_2,\cdots,n_K} \mathbb{E}_q[\prod_{k=1}^{K} \boldsymbol{\theta}^{n_k}]\end{aligned} \tag{8}$$

where n_k denotes the number of words assigned with topic label k and $\sum_{k=1}^{K} n_k = N$. We further bound $\sum_{n_1,n_2,\cdots,n_K} \mathbb{E}_q[\prod_{k=1}^{K} \boldsymbol{\theta}^{n_k}]$ as follows

$$\begin{aligned}&\sum_{n_1,n_2,\cdots,n_K} \mathbb{E}_q[\prod_{k=1}^{K} \boldsymbol{\theta}^{n_k}]\\&= \sum_{n_1,n_2,\cdots,n_K} \frac{\Gamma(\sum_{k=1}^{K}\eta_k)}{\prod_{k=1}^{K}\Gamma(\eta_k)} \int \prod_{k=1}^{K} \theta^{n_k+\eta_k-1}\mathrm{d}\boldsymbol{\theta}\\&= \sum_{n_1,n_2,\cdots,n_K} \frac{\Gamma(\sum_{k=1}^{K}\eta_k)}{\prod_{k=1}^{K}\Gamma(\eta_k)} \frac{\prod_{k=1}^{K}\Gamma(n_k+\eta_k)}{\Gamma(\sum_{k=1}^{K} n_k+\eta_k)}\\&= \sum_{n_1,n_2,\cdots,n_K} \frac{\prod_{k=1}^{K}(\eta_k)_{n_k}}{(\sum_{k=1}^{K}\eta_k)_N}\\&\leq \sum_{n_1,n_2,\cdots,n_K} \frac{\prod_{k=1}^{K}(n_k)!}{(N)!}\end{aligned} \tag{9}$$

where $(a)_n$ denotes the Pochhammer symbol, which is defined as $(a)_n = a(a+1)\ldots(a+n-1)$ and $\sum_{n_1,n_2,\cdots,n_K} \frac{\prod_{k=1}^{K}(n_k)!}{(N)!}$ is a constant. Setting $c = c/\sum_{n_1,n_2,\cdots,n_K} \frac{\prod_{k=1}^{K}(n_k)!}{(N)!}$, we get

$$\mathbb{E}_q[\log A(\boldsymbol{\theta}, \lambda)] \leq c^{-1}\exp\{\sum_{(i,j)\in\mathcal{P}} \lambda\} - 1 + \log c \tag{10}$$

Given this upper bound, we can obtain a lower bound of the variational lower bound defined in Eq.(5). Variational parameters and model parameters can be learned by maximizing the lower bound using iterative EM algorithm. In E-step, we fix the model parameters and compute the variational parameters by setting the derivatives of the lower bound w.r.t the variational parameters to zero

$$\eta_k = \alpha_k + \sum_{i=1}^{N}\phi_{ik}, c = \exp\{\sum_{(m,n)\in\mathcal{P}} \lambda\} \tag{11}$$

$$\begin{aligned}&\phi_{ik} \propto \exp\{\Psi(\eta_k) - \Psi(\sum_{j=1}^{K}\eta_j) + \lambda \sum_{j\in\mathcal{N}(i)} \phi_{jk}\\&+ \sum_{v=1}^{V} w_{iv}\log\beta_{kv}\}\end{aligned} \tag{12}$$

In Eq.(12), $\mathcal{N}(i)$ denotes the words that are labeled to be similar to i. As can be seen from this equation, the probability ϕ_{ik} that word i is assigned to topic k depends on the probability ϕ_{jk} of i's correlated words j. This explains how our model can incorporate word correlations in topic assignments. In M-step, we fix the variational parameters and update the model parameters by maximizing the lower bound defined on the set of documents $\{\mathbf{w}_d\}_{d=1}^{D}$

$$\beta_{kv} \propto \sum_{d=1}^{D}\sum_{i=1}^{N_d} \phi_{d,i,k} w_{d,i,v} \tag{13}$$

$$\lambda = \frac{1}{|P|}\log \frac{\sum_{d=1}^{D}\sum_{(m,n)\in P_d}\sum_{k=1}^{K} \phi_{d,m,k}\phi_{d,n,k}}{|P|\sum_{d=1}^{D}\frac{1}{c_d}} \tag{14}$$

4 Experiment

In this section, we corroborate the effectiveness of our model by comparing it with three baseline methods on two datasets.

dataset	20-Newsgroups	NIPS
# documents	18846	1500
# words	40343	12419

Table 1: Dataset Statistics

4.1 Experiment Setup

- **Dataset**: We use two datasets in the experiments: 20-Newsgroups[2] and NIPS[3]. Their statistics are summarized in Table 1.

- **External Knowledge**: We extract word correlation knowledge from Web Eigenwords[4], where each word has a real-valued vector capturing the semantic meaning of this word based on distributional similarity. Two words are regarded as correlated if their representation vectors are similar enough. It is worth mentioning that, other sources of external word correlation knowledge, such as Word2Vec (Mikolov et al., 2013) and Glove (Pennington et al., 2014), can be readily incorporated into MRF-LDA.

- **Baselines**: We compare our model with three baseline methods: LDA (Blei et al., 2003), DF-LDA (Andrzejewski et al., 2009) and Quad-LDA (Newman et al., 2011). LDA is the most widely used topic model, but it is unable to incorporate external knowledge. DF-LDA and Quad-LDA are two models designed to incorporate word correlation to improve topic modeling. DF-LDA puts a Dirichlet Forest prior over the topic-word multinomials to encode the Must-Links and Cannot-Links between words. Quad-LDA regularizes the topic-word distributions with a structured prior to incorporate word relation.

- **Parameter Settings:** For all methods, we learn 100 topics. LDA parameters are set to their default settings in (Andrzejewski et al., 2009). For DF-LDA, we set its parameters as $\alpha = 1$, $\beta = 0.01$ and $\eta = 100$. The Must/Cannot links between words are generated based on the cosine similarity of words' vector representations in Web Eigenwords. Word pairs with similarity higher than 0.99 are set as Must-Links, and pairs with similarity lower than 0.1 are put into Cannot-Link set. For Quad-LDA, β is set as 0.01; α is defined as $\frac{0.05 \cdot N}{D \cdot T}$, where N is the total occurrences of all words in all documents, D is the number of documents and T is topic number. For MRF-LDA, word pairs with similarity higher than 0.99 are labeled as correlated.

[2]http://qwone.com/ jason/20Newsgroups/

[3]http://archive.ics.uci.edu/ml/datasets/Bag+of+Words

[4]http://www.cis.upenn.edu/ ungar/eigenwords/

4.2 Results

We compare our model with the baseline methods both qualitatively and quantitatively.

4.2.1 Qualitative Evaluation

Table 2 shows some exemplar topics learned by the four methods on the 20-Newsgroups dataset. Each topic is visualized by the top ten words. Words that are noisy and lack representativeness are highlighted with bold font. Topic 1 is about crime and guns. Topic 2 is about sex. Topic 3 is about sports and topic 4 is about health insurance. As can be seen from the table, our method MRF-LDA can learn more coherent topics with fewer noisy and meaningless words than the baseline methods. LDA lacks the mechanism to incorporate word correlation knowledge and generates the words independently. The similarity relationships among words cannot be utilized to imporve the coherence of topic modeling. Consequently, noise words such as *will*, *year*, *used* which cannot effectively represent a topic, show up due to their high frequency. DF-LDA and Quad-LDA proposed to use word correlations to enhance the coherence of learned topics. However, they improperly enforce words labeled as similar to have similar probabilities in all topics, which violates the fact that whether two words are similar depend on which topic they appear in. As a consequence, the topics extracted by these two methods are unsatisfactory. For example, topic 2 learned by DF-LDA mixed up a *sex* topic and a *reading* topic. Less relevant words such as *columbia*, *year*, *write* show up in the health insurance topic (topic 4) learned by Quad-LDA. Our method MRF-LDA incorporates the word correlation knowledge by imposing a MRF over the latent topic layer to encourage correlated words to share the same topic label, hence similar words have better chance to be put into the same topic. Conse-

Table 2: Topics Learned from 20-Newsgroups Dataset

LDA				DF-LDA			
Topic 1 (Crime)	Topic 2 (Sex)	Topic 3 (Sports)	Topic 4 (Health)	Topic 1 (Crime)	Topic 2 (Sex)	Topic 3 (Sports)	Topic 4 (Health)
gun	sex	team	government	gun	**book**	game	money
guns	men	game	money	police	men	games	pay
weapons	homosexuality	hockey	private	carry	**books**	players	insurance
control	homosexual	season	people	kill	homosexual	hockey	policy
firearms	gay	**will**	**will**	killed	homosexuality	baseball	tax
crime	sexual	**year**	health	weapon	**reference**	fan	companies
police	**com**	play	tax	cops	gay	league	**today**
com	homosexuals	nhl	care	warrant	**read**	played	plan
weapon	people	games	insurance	deaths	male	season	health
used	**cramer**	teams	program	control	homosexuals	ball	jobs
Quad-LDA				MRF-LDA			
Topic 1 (Crime)	Topic 2 (Sex)	Topic 3 (Sports)	Topic 4 (Health)	Topic 1 (Crime)	Topic 2 (Sex)	Topic 3 (Sports)	Topic 4 (Health)
gun	homosexuality	game	money	gun	men	game	care
guns	sex	team	insurance	guns	sex	team	insurance
crime	homosexual	play	**columbia**	weapons	women	hockey	private
police	sin	games	pay	child	homosexual	players	cost
weapons	marriage	hockey	health	police	homosexuality	play	health
firearms	**context**	season	tax	control	child	player	costs
criminal	people	**rom**	**year**	kill	ass	fans	company
criminals	sexual	**period**	private	deaths	sexual	teams	companies
people	gay	goal	care	death	gay	fan	tax
law	homosexuals	player	**write**	**people**	homosexuals	best	public

quently, the learned topics are of high coherence. As shown in Table 2, the topics learned by our method are largely better than those learned by the baseline methods. The topics are of high coherence and contain fewer noise and irrelevant words.

Our method provides a mechanism to automatically decide under which topic, two words labeled as similar are truly similar. The decision is made flexibly by data according to their fitness to the model, rather than by a hard rule adopted by DF-LDA and Quad-LDA. For instance, according to the external knowledge, the word *child* is correlated with *gun* and with *men* simultaneously. Under a *crime* topic, *child* and *gun* are truly correlated because they co-occur a lot in youth crime news, whereas, *child* and *men* are less correlated in this topic. Under a *sex* topic, *child* and *men* are truly correlated whereas *child* and *gun* are not. Our method can differentiate this subtlety and successfully put *child* and *gun* into the *crime* topic and put *child* and *men* into the *sex* topic. This is because our method encourages *child* and *gun* to be put into the same topic A and encourages *child* and *men* to be put into the same topic B, but does not require A and B to be the same. A and B are freely decided by data.

Table 3 shows some topics learned on NIPS dataset. The four topics correspond to vision, neural network, speech recognition and electronic circuits respectively. From this table, we observe that the topics learned by our method are better in coherence than those learned from the baseline methods, which again demonstrates the effectiveness of our model.

4.2.2 Quantitative Evaluation

We also evaluate our method in a quantitative manner. Similar to (Xie and Xing, 2013), we use the coherence measure (CM) to assess how coherent the learned topics are. For each topic, we pick up the top 10 candidate words and ask human annotators to judge whether they are relevant to the topic. First, annotators needs to judge whether a topic is interpretable or not. If not, the ten candidate words in this

Table 3: Topics Learned from NIPS Dataset

LDA				DF-LDA			
Topic 1 (Vision)	Topic 2 (Neural Net)	Topic 3 (Speech)	Topic 4 (Circuits)	Topic 1 (Vision)	Topic 2 (Neural Net)	Topic 3 (Speech)	Topic 4 (Circuits)
image	network	hmm	chip	images	network	speech	analog
images	neural	**mlp**	analog	pixel	system	context	chip
pixel	feedforward	hidden	**weight**	view	connection	speaker	vlsi
vision	architecture	context	digital	recognition	application	frame	implement
segment	**research**	model	**neural**	face	artificial	continuous	digital
visual	**general**	recognition	hardware	ica	input	processing	hardware
scene	**applied**	probabilities	bit	vision	**obtained**	**number**	voltage
texture	**vol**	training	**neuron**	system	**department**	**dependent**	bit
contour	**paper**	markov	implement	**natural**	**fixed**	frames	transistor
edge	**introduction**	system	vlsi	faces	techniques	spectral	design
Quad-LDA				MRF-LDA			
Topic 1 (Vision)	Topic 2 (Neural Net)	Topic 3 (Speech)	Topic 4 (Circuits)	Topic 1 (Vision)	Topic 2 (Neural Net)	Topic 3 (Speech)	Topic 4 (Circuits)
image	training	speech	circuit	image	network	hmm	chip
images	set	hmm	analog	images	model	speech	**synapse**
pixel	network	speaker	chip	pixel	learning	acoustic	digital
region	learning	acoustic	voltage	disparity	function	context	analog
vision	net	phonetic	current	color	input	word	board
scene	number	vocabulary	vlsi	intensity	neural	phonetic	charge
surface	algorithm	phone	**neuron**	stereo	set	frames	**synaptic**
texture	**class**	utterance	gate	scene	algorithm	speaker	hardware
local	input	utterances	**input**	camera	system	phone	vlsi
contour	**examples**	frames	transistor	**detector**	data	vocabulary	programmable

topic are automatically labeled as irrelevant. Otherwise, annotators are asked to identify words that are relevant to this topic. Coherence measure (CM) is defined as the ratio between the number of relevant words and total number of candidate words. In our experiments, four graduate students participated the labeling. For each dataset and each method, 10% of topics were randomly chosen for labeling.

Table 4 and 5 summarize the coherence measure of topics learned on 20-Newsgroups dataset and NIPS dataset respectively. As shown in the table, our method significantly outperforms the baseline methods with a large margin. On the 20-Newsgroups dataset, our method achieves an average coherence measure of 60.8%, which is two times better than LDA. On the NIPS dataset, our method is also much better than the baselines. In summary, we conclude that MRF-LDA produces much better results on both datasets compared to baselines, which demonstrates the effectiveness of our model in exploiting word correlation knowledge to improve the quality of topic modeling. To assess the consistency of the labelings made by different annotators, we computed the intraclass correlation coefficient (ICC). The ICCs on 20-Newsgroups and NIPS dataset are 0.925 and 0.725 respectively, which indicate good agreement between different annotators.

5 Conclusion

In this paper, we propose a MRF-LDA model, aiming to incorporate word correlation knowledge to improve topic modeling. Our model defines a MRF over the latent topic layer of LDA, to encourage correlated words to be put into the same topic. Our model provides the flexibility to enable a word to be similar to different words under different topics, which is more plausible and allows a word to show up in multiple topics properly. We evaluate our model on two datasets and corroborate its effectiveness both qualitatively and quantitatively.

Method	Annotator1	Annotator2	Annotator3	Annotator4	Mean	Standard Deviation
LDA	30	33	22	29	28.5	4.7
DF-LDA	35	41	35	27	36.8	2.9
Quad-LDA	32	36	33	26	31.8	4.2
MRF-LDA	**60**	**60**	**63**	**60**	**60.8**	**1.5**

Table 4: CM (%) on 20-Newsgroups Dataset

Method	Annotator1	Annotator2	Annotator3	Annotator4	Mean	Standard Deviation
LDA	75	74	74	69	73	2.7
DF-LDA	65	74	72	47	66	9.5
Quad-LDA	40	40	38	25	35.8	7.2
MRF-LDA	**86**	**85**	**87**	**84**	**85.8**	**1.0**

Table 5: CM (%) on NIPS Dataset

References

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*, pages 25–32. ACM.

David Andrzejewski, Xiaojin Zhu, Mark Craven, and Benjamin Recht. 2011. A framework for incorporating general domain knowledge into latent dirichlet allocation using first-order logic. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Two*, pages 1171–1177. AAAI Press.

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

Jordan L Boyd-Graber, David M Blei, and Xiaojin Zhu. 2007. A topic model for word sense disambiguation. In *EMNLP-CoNLL*, pages 1024–1033.

Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2013. Leveraging multi-domain prior knowledge in topic models. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI'13, pages 2071–2077.

Gregor Heinrich. 2009. A generic approach to topic models. In *Machine Learning and Knowledge Discovery in Databases*, pages 517–532. Springer.

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. ACM.

Yuening Hu, Jordan L Boyd-Graber, and Brianna Satinoff. 2011. Interactive topic modeling. In *ACL*, pages 248–257.

Jagadeesh Jagarlamudi, Hal Daumé, III, and Raghavendra Udupa. 2012. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 204–213.

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

George A Miller. 1995a. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

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

David Newman, Edwin V Bonilla, and Wray Buntine. 2011. Improving topic coherence with regularized topic models. In *Advances in Neural Information Processing Systems*, pages 496–504.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 12.

James Petterson, Wray Buntine, Shravan M Narayanamurthy, Tibério S Caetano, and Alex J Smola. 2010. Word features for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 1921–1929.

Jakob Verbeek and Bill Triggs. 2007. Region classification with markov field aspect models. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE.

Martin J Wainwright and Michael I Jordan. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305.

Pengtao Xie and Eric P Xing. 2013. Integrating document clustering and topic modeling. *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*.

Bin Zhao, Li Fei-Fei, and Eric Xing. 2010. Image segmentation with topic random field. *Computer Vision–ECCV 2010*, pages 785–798.

Jun Zhu and Eric P Xing. 2010. Conditional topic random fields. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 1239–1246.