Confirmatory Latent Dirichlet Allocation

Table of Contents

1. Introduction

1.1. Problem Definition

Because of the unsupervised nature of LDA, topics should be manually named after inference. As a result, it is sometimes difficult to interpret some of unclarified topic distributions. In addition, it is often denoted that it is difficult to integrate prior knowledge into LDA which could be used for performance improvement. To tackle these issues, we propose a novel topic modeling approach, called Confirmatory Latent Dirichlet Allocation (CLDA). We extend functionality of LDA from unsupervised learning to semi supervised learning with text sumarization function by incorporating prior knowledge in shape of seed words.

1.2. Contribution

1. Developing variational inference for short text
2. Integrating prior knowledge in shape of seed words to LDA
3. Propose a novel text summarization method

1.3 Abstract on ICDM submission

Latent Dirichlet Allocation (LDA) is one of the most popular topic modeling method which can be used to find latent topic representation of a given corpus. However, it is difficult to interpret topics in clarified conceptions due to that LDA often produces incoherent words in each topic dimension. Even if we have clearly identified topics, those may not be main interest of users. Those limitations are more critical in the case of short texts. To tackle this issue, we propose a novel topic modeling approach, called Confirmatory Latent Dirichlet Allocation (CLDA). We integrate prior knowledge into each topic cluster in shape of seed words with asymmetric dirichlet priors. By its construction, CLDA is more likely to generate coherent word tokens within the top list of each topic cluster. Those coherent words help us identify a single topic cluster and distinguish one from another. Consequently, users get cohesive topic distribution on which corpus is summarized more easily. Experiments with real datasets show that our proposed methodology is better at achieving coherent topic distributions than other LDA-based methods.

2. Methodology

2.1. Latent Dirichlet Allocation with Multinomial Mixture assumption

2.1.1. Generative Process

1. For each topic $k = 1, \cdots, k = K$, choose $\beta_k \sim \text{Dir}(\eta)$
2. Choose $\theta \sim \text{Dir}(\alpha)$
3. For each word $n = 1, \cdots, n = N_d$ in document d
   - Select topic of word, i.e., $z_d \sim \text{Multi}(\theta)$
   - Select word, i.e., $w_{d,n} \sim \text{Multi}(\beta_{z_d})$

2.1.2. Model Assumption

$$
\begin{aligned}
w_{di} &\sim \prod_w \prod_k \beta_{kw}^{\mathrm{I}(z_d=k)\mathrm{I}(w_{di}=w)} \\
\beta_k &\sim \text{Dir}(\eta) \\
z_d &\sim \prod_k \theta_k^{\mathrm{I}(z_d=k)} \\
\theta &\sim \text{Dir}(\alpha) \\
q(z_d = k) &= \phi_{dk} \\
q(\theta) &= \text{Dir}(\theta; \gamma) \\
q(\beta_k) &= \text{Dir}(\beta_k; \lambda_k)
\end{aligned}
\tag{1}
$$

2.1.3. Objective function

$$
\begin{aligned}
\mathcal{L} &= E_q[\log p(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{z}, \mathbf{w} \mid \alpha, \eta)] - E_q[\log q(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{z})] \\
&= \sum_d [E_q[\log p(w_d \mid \theta_d, z_d, \boldsymbol{\beta})] + E_q[\log p(z_d \mid \theta_d)] \\
&- E_q[\log q(z_d)] + E_q[\log p(\theta_d \mid \alpha)] - E_q[\log q(\theta_d)] \\
&+ (E_q[\log p(\boldsymbol{\beta} \mid \eta)] - E_q[\log q(\boldsymbol{\beta})])/D]
\end{aligned}
\tag{2}
$$

2.1.4 Variational Inference

Algorithm 1: Variational Inference for LDA-MM

Input Document-term matrix X_{ij}; Number of latent topics K
Output Document topic distribution ϕ_{dk}; Topic word distribution λ_{kw}
while *Relative change in* $\mathcal{L} > 0.0001$ **do**
  E-Step
  for $d \leftarrow 1$ **to** D **do**
    $\phi_{dk} \propto \exp\{E[\log \theta_k] + \sum_w n_{dw} E[\log \beta_{kw}]\}$
  end
  M-Step
  $\gamma_k = \alpha + \sum_d \phi_{dk}$
  $\lambda_{kw} = \eta + \sum_d n_{dw} \phi_{dk}$
end

2.2. Confirmatory Latent Dirichlet Allocation

2.2.1. Setting seed words

1. Determine topic dimension on which text will be summarized
   - Ex) Food, Service, price, drink for restaurant review data

2. Find seed words related with topics using pretrained word vector model.

    - criteria of relateness: Cosine similarity between two word vectors

2.2.2. How to integrate seed words into LDA

Original LDA assumes symmetric dirichlet prior whereas we assume asymmetric dirichlet prior categorized by seed words. In other words, assuming topic-word distribution as $\beta_k \sim \text{Dir}(\eta_k)$ and seed words for topic k as S_k, we set

$$\eta_{kw} > \eta_{kw'},\ w \in S_k,\ w' \notin S_k$$

2.2.3. Variational Inference

Algorithm 2: Variational Inference for CLDA

Input Document-term matrix $X_{i,j}$; Number of latent topics K; Seed words for each topic S_k
Output Document topic distribution ϕ_{dk}; Topic word distribution λ_{kw}

```
1. Setting prior
for k ∈ {1, ··· , K} do
    for w ∈ {1, ··· , V} do
        if w ∈ S_k then
            η_kw ∝ A_kw
        end
        else if w ∈ S_{k'∈{1,···,K}\k} then
            η_kw ∝ 1 / A_kw
        end
        else
            η_kw = α
        end
    end
end
2. Inferring Variational Parameters
while Relative change in L > 0.0001 do
    E-Step
    for d ← 1 to D do
        φ_dk ∝ exp{E[log θ_k] + Σ_w n_dw E[log β_kw]}
    end
    M-Step
    γ_k = α + Σ_d φ_dk
    λ_kw = η_kw + Σ_d n_dw φ_dk
end
```

3. Experiments

3.1. Metric

3.1.1. Perplexity

In topic model literature, perplexity is usually used for data likelihood. In other words, we can measure how well our model fits to data comparing other competitive models. Note that high data likelihood, i.e., low perplexity, does not mean coherent topic distributions. Therefore, we consider other metric to measure topic coherence.

3.1.2. Within Topic Coherence (WTC)

We use cosine similarity between word vector to measure coherence within topic.

$$
\begin{aligned}
WC &= \frac{1}{K}\sum_{k=1}^{K} WC_k \\
&= \frac{1}{K}\sum_{k=1}^{K}\sum_{j=1}^{N}\sum_{i\in\{1,\cdots,N\}\setminus j} \mathrm{sim}(w_{kj}, w_{ki})
\end{aligned}
\tag{3}
$$

where $\mathrm{sim}(w_{kj}, w_{ki})$ denotes similarity measure between jth word and ith word in topic k. However, this metric cannot measure semantic distance **between** topic. Therefore, we consider another metric.

3.1.3. Between Topic Coherence (BTC)

For wth word in topic k, i.e., w_{kw}, BTC is defined as

$$
BTC_{w_{kw}} = \begin{cases} 1, & \text{if } B \leq A \\ 0, & \text{if } A \leq B \end{cases} \tag{4}
$$

where

$$
\begin{aligned}
A &= \sum_{i\in\{1,\cdots,N\}\setminus w} \mathrm{sim}(w_{kw}, w_{ki}) \\
B &= \max_{k'\in\{1,\cdots,K\}\setminus k} \left\{ \sum_{i=1}^{N} \mathrm{sim}(w_{kw}, w_{k'i}) \right\}
\end{aligned}
\tag{5}
$$

First, A denotes within topic coherence for topic k. Next, B denotes coherence between w_{kw} and top N words in other topics, $k' \in \{1, \cdots, K\} \setminus k$ and we take maximum among k'. Therefore, if $BTC_{w_{kw}} = 1$, we can claim that w_{kw} is semantically closer in topic k than other topics k'. Finally, BTC for topic model is the summation of all $BTC_{w_{kw}}$,

$$
BC = \sum_{k=1}^{K}\sum_{w=1}^{N} BC_{w_{kw}} \tag{6}
$$

3.2. Set up

We compare CLDA and LDA and SourceLDA with multinomial mixture. We choose to comparable method as SourceLDA because the way it integrates prior knowledge is similar to ours.

3.3. Result

TABLE IV: Quantitative result for top 20 words

Dataset	Model	Perplexity	WTC	BTC
Hotel review data	LDA-MM	**3307**	30	4
	SourceLDA	4391	29	6
	CLDA	4247	**35**	**38**
Restaurant review data	LDA-MM	**15781**	8.5	18
	SourceLDA	27745	8.2	15
	CLDA	23074	**9.1**	**52**
Car review data	LDA-MM	**4740**	8.5	1
	SourceLDA	6156	**8.9**	3
	CLDA	5050	8.3	**19**

3.3.1. Perplexity

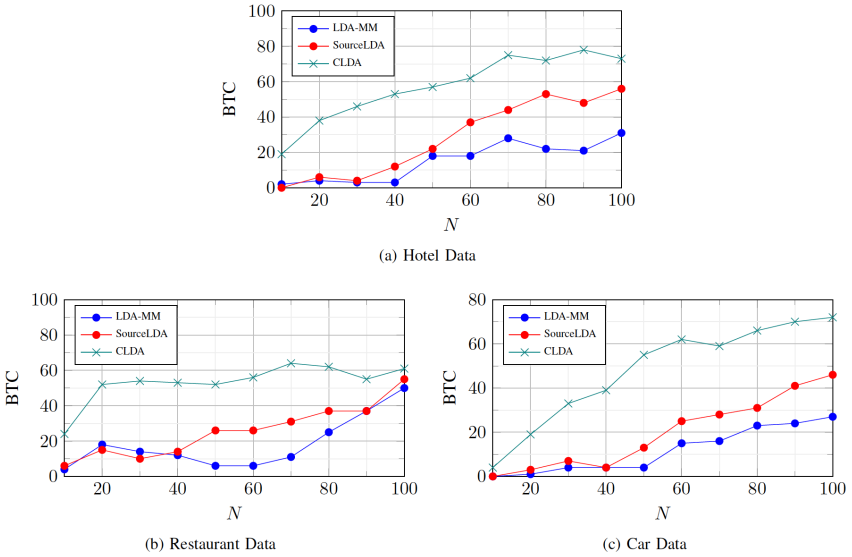
As expected, perplexity of CLDA is higher than original LDA in any dataset, which means CLDA lacks data likelihood comparing original LDA. However, as noted before, perplexity does not gurantee coherent topic distributions which will be verified by other metric

3.3.2. WTC and BTC

Because BTC complements disadvantage of WTC, we see the results of BTC only. For each dataset, BTC of CLDA is highest comparing other models. Therefore, we can conclude that CLDA produces more coherent topics than other models.

3.3.3. Details of BTC

Above table shows results of top **20 words** for each topic. It is necessary to check whether we get consistent results for other number of top N words.



(a) Hotel Data

(b) Restaurant Data

(c) Car Data

Fig. 2: (a),(b),(c) - BTC results in each dataset

Above figure shows change in BTC varying top N words in each dataset. From this figure, it is verified that performance of CLDA is consistently better than other models in each dataset, varying top N words.

4. Future work

4.1. Developing algorithms for selecting seed words

In this work, we choose seed words based on cosine similarity between word vector. We plan to modify this procedure algorithmically

4.2. Integrating other forms of prior knowledge

In this work, we integrate prior knowledge in shape of seed words. We plan to try various forms of prior knowledge such as word vectors or markov random fields.
By integrating word vectors to topic model, we can relax the independence assumption between words in topic model. Also, it is worth to incorporate prior knowledge of documents. We plan to do this with markov random field.