Topic Models with Logical Constraints on Words

Hayato Kobayashi **Hiromi Wakaki** **Tomohiro Yamasaki** **Masaru Suzuki**
Research and Development Center,
Toshiba Corporation, Japan
{hayato.kobayashi, hiromi.wakaki,
tomohiro2.yamasaki, masaru1.suzuki}@toshiba.co.jp

Abstract

This paper describes a simple method to achieve logical constraints on words for topic models based on a recently developed topic modeling framework with Dirichlet forest priors (LDA-DF). Logical constraints mean logical expressions of pairwise constraints, *Must-links* and *Cannot-Links*, used in the literature of constrained clustering. Our method can not only cover the original constraints of the existing work, but also allow us easily to add new customized constraints. We discuss the validity of our method by defining its asymptotic behaviors. We verify the effectiveness of our method with comparative studies on a synthetic corpus and interactive topic analysis on a real corpus.

1 Introduction

Topic models such as Latent Dirichlet Allocation or LDA (Blei et al., 2003) are widely used to capture hidden topics in a corpus. When we have domain knowledge of a target corpus, incorporating the knowledge into topic models would be useful in a practical sense. Thus there have been many studies of semi-supervised extensions of topic models (Andrzejewski et al., 2007; Toutanova and Johnson, 2008; Andrzejewski et al., 2009; Andrzejewski and Zhu, 2009), although topic models are often regarded as unsupervised learning. Recently, (Andrzejewski et al., 2009) developed a novel topic modeling framework, LDA with Dirichlet Forest priors (LDA-DF), which achieves two links *Must-Link* (*ML*) and *Cannot-Link* (*CL*) in the constrained clustering literature (Basu et al., 2008). For given words A and B, *ML*(A, B) and *CL*(A, B) are soft constraints that A and B must appear in the same topic, and that A and B cannot appear in the same topic, respectively.

Let us consider topic analysis of a corpus with movie reviews for illustrative purposes. We know that two words ‘jackie’ (means Jackie Chan) and ‘kung-fu’ should appear in the same topic, while ‘dicaprio’ (means Leonardo DiCaprio) and ‘kung-fu’ should not appear in the same topic. In this case, we can add constraints *ML*(‘jackie’, ‘kung-fu’) and *CL*(‘dicaprio’, ‘kung-fu’) to smoothly conduct analysis. However, what if there is a word ‘bruce’ (means Bruce Lee) in the corpus, and we want to distinguish between ‘jackie’ and ‘bruce’? Our full knowledge among ‘kung-fu’, ‘jackie’, and ‘bruce’ should be (*ML*(‘kung-fu’, ‘jackie’) $\vee$ *ML*(‘kung-fu’, ‘bruce’)) $\wedge$ *CL*(‘bruce’, ‘jackie’), although the original framework does not allow a disjunction ($\vee$) of links. In this paper, we address such logical expressions of links on LDA-DF framework.

Combination between a probabilistic model and logical knowledge expressions such as Markov Logic Network (MLN) is recently getting a lot of attention (Riedel and Meza-Ruiz, 2008; Yu et al., 2008; Meza-Ruiz and Riedel, 2009; Yoshikawa et al., 2009; Poon and Domingos, 2009), and our work can be regarded as on this research line. At least, to our knowledge, our method is the first one that can directly incorporate logical knowledge into a prior for topic models without MLN. This means the complexity of the inference in our method is essentially the same as in the original LDA-DF, despite that our method can broaden knowledge expressions.

2 LDA with Dirichlet Forest Priors

We briefly review LDA-DF. Let $\mathbf{w} := w_1 \dots w_n$ be a corpus consisting of D documents, where n is the total number of words in the documents. Let d_i and z_i be the document that includes the i-th word w_i and the hidden topic that is assigned to w_i, respectively. Let T be the number of topics.

Robust Unsupervised and Semi-Supervised Methods in Natural Language Processing, pages 33–40,
Hissar, Bulgaria, 15 September 2011.

As in LDA, we assume a probabilistic language model that generates a corpus as a mixture of hidden topics and infer two parameters: a document-topic probability θ that represents a mixture rate of topics in each document, and a topic-word probability ϕ that represents an occurrence rate of words in each topic. The model is defined as

$$\begin{aligned} \theta_{d_i} &\sim \textit{Dirichlet}(\alpha), \\ z_i|\theta_{d_i} &\sim \textit{Multinomial}(\theta_{d_i}), \\ \mathbf{q} &\sim \textit{DirichletForest}(\beta, \eta), \\ \phi_{z_i} &\sim \textit{DirichletTree}(\mathbf{q}), \\ w_i|z_i, \phi_{z_i} &\sim \textit{Multinomial}(\phi_{z_i}), \end{aligned}$$

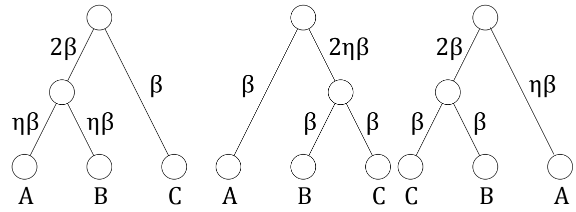
where α and (β, η) are hyper parameters for θ and ϕ, respectively. The only difference between LDA and LDA-DF is that ϕ is chosen not from the Dirichlet distribution, but from the Dirichlet tree distribution (Dennis III, 1991), which is a generalization of the Dirichlet distribution. The Dirichlet forest distribution assigns one tree to each topic from a set of Dirichlet trees, into which we encode domain knowledge. The trees assigned to topics $\mathbf{z}$ are denoted as $\mathbf{q}$.

In the framework, *ML*(A, B) is achieved by the Dirichlet tree in Fig. 1(a), which equalizes the occurrence probabilities of A and B in a topic when η is large. This tree generates probabilities with *Dirichlet*$(2\beta, \beta)$ and redistributes the probability for "2β" with *Dirichlet*$(\eta\beta, \eta\beta)$.

In the case of *CL*s, we use the following algorithm.

1. Consider a undirected graph regarding words as vertices and links *CL*(A, B) as edges between A and B.
2. Divide the graph into connected components.
3. Extract the maximal independent sets of each component.
4. Create Dirichlet trees to raise the occurrence probabilities of words corresponding to each maximal independent set.

For examples, the algorithm creates the two trees in Fig. 1(b) for the constraint *CL*$(A, B) \wedge$ *CL*(A, C). The constraint is achieved when η is large, since words in each topic are chosen from the distribution of either the left tree that zeros the occurrence probability of A, or the right tree that zeros those of B and C.



(a) *ML*(A, B) (b) *CL*$(A, B) \wedge$ *CL*(A, C)

Figure 1: Dirichlet trees for two constraints of (a) *ML*(A, B) and (b) *CL*$(A, B) \wedge$ *CL*(A, C).

Inference of ϕ and θ is achieved by alternately sampling topic z_i for each word w_i and Dirichlet tree q_z for each topic z. Since the Dirichlet tree distribution is conjugate to the multinomial distribution, the sampling equation of z_i is easily derived like LDA as follows:

$$p(z_i = z \mid \mathbf{z}_{-i}, \mathbf{q}, \mathbf{w}) \quad \propto \\ (n_{-i,z}^{(d_i)} + \alpha) \prod_s^{I_z(\uparrow i)} \frac{\gamma_z^{(C_z(s\downarrow i))} + n_{-i}^{(C_z(s\downarrow i))}}{\sum_k^{C_z(s)} \left(\gamma_z^{(k)} + n_{-i,z}^{(k)}\right)},$$

where $n_{-i,z}^{(d)}$ represents the number of words (excluding w_i) assigning topic z in document d. $n_{-i,z}^{(k)}$ represents the number of words (excluding w_i) assigning topic z in the subtree rooted at node k in tree q_z. $I_z(\uparrow i)$ and $C_z(s \downarrow i)$ represents the set of internal nodes and the immediate child of node s, respectively, on the path from the root to leaf w_i in tree q_z. $C_z(s)$ represents the set of children of node s in tree q_z. $\gamma_z^{(k)}$ represents a weight of the edge to node k in tree q_z. Additionally, we define $\sum_s^S := \sum_{s \in S}$.

Sampling of tree q_z is achieved by sequentially sampling subtree $q_z^{(r)}$ corresponding to the r-th connected component by using the following equation:

$$p(q_z^{(r)} = q' \mid \mathbf{z}, \mathbf{q}_{-z}, q_z^{(-r)}, \mathbf{w}) \quad \propto \quad |M_{r,q'}| \times \\ \prod_s^{I_{z,r}^{(q')}} \left(\frac{\Gamma\left(\sum_k^{C_z(s)} \gamma_z^{(k)}\right) \prod_k^{C_z(s)} \Gamma\left(\gamma_z^{(k)} + n_z^{(k)}\right)}{\Gamma\left(\sum_k^{C_z(s)} (\gamma_z^{(k)} + n_z^{(k)})\right) \prod_k^{C_z(s)} \Gamma\left(\gamma_z^{(k)}\right)} \right),$$

where $I_{z,r}^{(q')}$ represents the set of internal nodes in the subtree q' corresponding to the r-th connected component for tree q_z. $|M_{r,q'}|$ represents the size of the maximal independent set corresponding to the subtree q' for r-th connected component.

After sufficiently sampling z_i and q_z, we can infer posterior probabilities $\hat{\phi}$ and $\hat{\theta}$ using the last sampled $\mathbf{z}$ and $\mathbf{q}$, in a similar manner to the standard LDA as follows.

$$\hat{\theta}_z^{(d)} = \frac{n_z^{(d)} + \alpha}{\sum_{z'=1}^{T} \left(n_{z'}^{(d)} + \alpha \right)}$$

$$\hat{\phi}_z^{(w)} = \prod_{s}^{I_z(\uparrow w)} \frac{\gamma_z^{(C_z(s\downarrow w))} + n_z^{(C_z(s\downarrow w))}}{\sum_k^{C_z(s)} \left(\gamma_z^{(k)} + n_z^{(k)} \right)}$$

3 Logical Constraints on Words

In this section, we address logical expressions of two links using disjunctions ($\vee$) and negations ($\neg$), as well as conjunctions ($\wedge$), e.g., $\neg \mathit{ML}(A,B) \vee \mathit{ML}(A,C)$. We denote it as ($\wedge$,$\vee$,$\neg$)-expressions. Since each negation can be removed in a preprocessing stage, we focus only on ($\wedge$,$\vee$)-expressions. Interpretation of negations is discussed in Sec. 3.4.

3.1 ($\wedge$,$\vee$)-expressions of Links

We propose a simple method that simultaneously achieves conjunctions and disjunctions of links, where the existing method can only treat conjunctions of links. The key observation is that any Dirichlet trees constructed by *ML*s and *CL*s are essentially based only on two primitives. One is $\mathit{Ep}(A,B)$ that equalizes the occurrence probabilities of A and B in a topic as in Fig. 1(a), and the other is $\mathit{Np}(A)$ that zeros the occurrence probability of A in a topic as in the left tree of Fig. 1(b). The right tree of Fig. 1(b) is created by $\mathit{Np}(B) \wedge \mathit{Np}(C)$. Thus, we can substitute *ML* and *CL* with *Ep* and *Np* as follows:

$$\begin{aligned} \mathit{ML}(A,B) &= \mathit{Ep}(A,B) \\ \mathit{CL}(A,B) &= \mathit{Np}(A) \vee \mathit{Np}(B) \end{aligned}$$

Using this substitution, we can compile a ($\wedge$, $\vee$)-expression of links to the corresponding Dirichlet trees with the following algorithm.

1. Substitute all links (*ML* and *CL*) with the corresponding primitives (*Ep* and *Np*).

2. Calculate the minimum DNF of the primitives.

3. Construct Dirichlet trees corresponding to the (monotone) monomials of the DNF.

Let us consider three words A = 'kung-fu', B = 'jackie', and C = 'bruce' in Sec. 1. We want to constrain them with $(\mathit{ML}(A,B) \vee \mathit{ML}(A,C)) \wedge \mathit{CL}(B,C)$. In this case, the algorithm calculates the minimum DNF of primitives as

$$\begin{aligned} &(\mathit{ML}(A,B) \vee \mathit{ML}(A,C)) \ \wedge \ \mathit{CL}(B,C) \\ &= (\mathit{Ep}(A,B) \vee \mathit{Ep}(A,C)) \ \wedge \ (\mathit{Np}(B) \vee Np(C)) \\ &= (\mathit{Ep}(A,B) \wedge \mathit{Np}(B)) \ \vee \ (\mathit{Ep}(A,B) \wedge \mathit{Np}(C)) \\ &\ \vee \ (\mathit{Ep}(A,C) \wedge \mathit{Np}(B)) \ \vee \ (\mathit{Ep}(A,C) \wedge \mathit{Np}(C)) \end{aligned}$$

and constructs four Dirichlet trees corresponding to the four monomials $\mathit{Ep}(A,B) \wedge \mathit{Np}(B)$, $\mathit{Ep}(A,B) \wedge \mathit{Np}(C)$, $\mathit{Ep}(A,C) \wedge \mathit{Np}(B)$, and $\mathit{Ep}(A,C) \wedge \mathit{Np}(C)$ in the last equation.

Considering only ($\wedge$)-expressions of links, our method is equivalent to the existing method in the original framework in terms of an asymptotic behavior of Dirichlet trees. We define asymptotic behavior as *Asymptotic Topic Family (ATF)* as follows.

Definition 1 (Asymptotic Topic Family)**.** *For any ($\wedge$, $\vee$)-expression f of primitives and any set $\mathcal{W}$ of words, we define* the asymptotic topic family of f with respect to $\mathcal{W}$ *as a family f^* calculated by the following rules: Given ($\wedge$, $\vee$)-expressions f_1 and f_2 of primitives and words $A, B \in \mathcal{W}$,*

(i) $(f_1 \vee f_2)^* \ := \ f_1^* \cup f_2^*$,

(ii) $(f_1 \wedge f_2)^* \ := \ f_1^* \cap f_2^*$,

(iii) $\mathit{Ep}^*(A,B) \ := \ \{\emptyset, \{A,B\}\} \otimes 2^{\mathcal{W}-\{A,B\}}$,

(iv) $\mathit{Np}^*(A) \ := \ 2^{\mathcal{W}-\{A\}}$.

Here, notation $\otimes$ is defined as $X \otimes Y := \{x \cup y \mid x \in X, y \in Y\}$ for given two sets X and Y. ATF expresses all combinations of words that can occur in a topic when η is large. In the above example, the ATF of its expression with respect to $\mathcal{W} = \{A,B,C\}$ is calculated as

$$\begin{aligned} &((\mathit{ML}(A,B) \vee \mathit{ML}(A,C)) \ \wedge \ \mathit{CL}(B,C))^* \\ &= (\mathit{Ep}(A,B) \vee \mathit{Ep}(A,C)) \ \wedge \ (\mathit{Np}(B) \vee \mathit{Np}(C))^* \\ &= \begin{pmatrix} \{\emptyset, \{A,B\}\} \otimes 2^{\mathcal{W}-\{A,B\}} \\ \cup \{\emptyset, \{A,C\}\} \otimes 2^{\mathcal{W}-\{A,C\}} \end{pmatrix} \\ &\quad \cap \left(2^{\mathcal{W}-\{B\}} \ \cup \ 2^{\mathcal{W}-\{C\}} \right) \\ &= \{\emptyset, \{B\}, \{C\}, \{A,B\}, \{A,C\}\}. \end{aligned}$$

As we expected, the ATF of the last equation indicates such a constraint that either A and B or A and C must appear in the same topic, and B and C cannot appear in the same topic. Note that the part of $\{B\}$ satisfies $ML(A, C) \wedge CL(B, C)$. If you want to remove $\{B\}$ and $\{C\}$, you can use exclusive disjunctions. For the sake of simplicity, we omit descriptions about $\mathcal{W}$ when its instance is arbitrary or obvious from now on.

The next theorem gives the guarantee of asymptotic equivalency between our method and the existing method. Let $MIS(G)$ be the set of maximal independent sets of graph G. We define $\mathcal{L} := \{\{w, w'\} \mid w, w' \in \mathcal{W}, w \neq w'\}$. We consider *CL*s only, since the asymptotic equivalency including *ML*s is obvious by identifying all vertices connected by *ML*s.

Theorem 2. *For any ($\wedge$)-expression of CLs represented by* $\bigwedge_{\{x,y\}\in\ell:\ell\subseteq\mathcal{L}} CL(x, y)$*, the ATF of the corresponding minimum DNF of primitives represented by* $\bigvee_{X\in\mathcal{X}:\mathcal{X}\subseteq 2^{\mathcal{W}}}(\bigwedge_{x\in X} Np(x))$ *is equivalent to the union of the power sets of every maximal independent set* $S \in MIS(G)$ *of a graph* $G := (\mathcal{W}, \ell)$*, that is,* $\bigcup_{X\in\mathcal{X}}\left(\bigcap_{x\in X} Np^*(x)\right) = \bigcup_{S\in MIS(G)} 2^S$.

Proof. For any ($\wedge$)-expressions of links characterized by $\ell \subseteq \mathcal{L}$, we denote f_ℓ and G_ℓ as the corresponding minimum DNF and graph, respectively. We define $\mathcal{U}_\ell := \bigcup_{S\in MIS(G_\ell)} 2^S$. When $|\ell| = 1$, $f_\ell^* = \mathcal{U}_\ell$ is trivial. Assuming $f_\ell^* = \mathcal{U}_\ell$ when $|\ell| > 1$, for any set $\ell' := \ell \cup \{\{A, B\}\}$ with an additional link characterized by $\{A, B\} \in \mathcal{L}$, we obtain

$$
\begin{aligned}
f_{\ell'}^* &= ((Np(A) \vee Np(B)) \wedge f_\ell)^* \\
&= (2^{\mathcal{W}-\{A\}} \cup 2^{\mathcal{W}-\{B\}}) \cap \mathcal{U}_\ell \\
&= \bigcup_{S\in MIS(G_\ell)} \begin{pmatrix} (2^{\mathcal{W}-\{A\}} \cap 2^S) \\ \cup(2^{\mathcal{W}-\{B\}} \cap 2^S) \end{pmatrix} \\
&= \bigcup_{S\in MIS(G_\ell)} (2^{S-\{A\}} \cup 2^{S-\{B\}}) \\
&= \bigcup_{S\in MIS(G_{\ell'})} 2^S \quad = \quad \mathcal{U}_{\ell'}
\end{aligned}
$$

This proves the theorem by induction. In the last line of the above deformation, we used $\bigcup_{S\in MIS(G)} 2^S = \bigcup_{S\in IS(G)} 2^S$ and $MIS(G_{\ell'}) \subseteq \bigcup_{S\in MIS(G_\ell)}((S - \{A\}) \cup (S - \{B\})) \subseteq IS(G_{\ell'})$, where $IS(G)$ represents the set of all independent sets on graph G. □

In the above theorem, $\bigcup_{X\in\mathcal{X}}\left(\bigcap_{x\in X} Np^*(x)\right)$ represents asymptotic behaviors of our method, while $\bigcup_{S\in MIS(G)} 2^S$ represents those of the existing method. By using a similar argument to the proof, we can prove the elements of the two sets are completely the same, i.e., $\bigcap_{x\in X} Np^*(x) = \{2^S \mid S \in MIS(G)\}$. This interestingly means that for any logical expression characterized by *CL*s, calculating its minimum DNF is the same as calculating the maximal independent sets of the corresponding graph, or the maximal cliques of its complement graph.

3.2 Shrinking Dirichlet Forests

Focusing on asymptotic behaviors, we can reduce the number of Dirichlet trees, which means the performance improvement of Gibbs sampling for Dirichlet trees. This is achieved just by minimizing DNF on *asymptotic equivalence relation* defined as follows.

Definition 3 (Asymptotic Equivalence Relation)**.** *Given two ($\wedge$, $\vee$)-expressions f_1, f_2, we say that f_1* is asymptotically equivalent to *f_2, if and only if $f_1^* = f_2^*$. We denote the relation as notation $\asymp$, that is, $f_1 \asymp f_2 \Leftrightarrow f_1^* = f_2^*$.*

The next proposition gives an intuitive understanding of why asymptotic equivalence relation can shrink Dirichlet forests.

Proposition 4. *For any two words $A, B \in \mathcal{W}$,*

(a) $Ep(A, B) \vee (Np(A) \wedge Np(B)) \asymp Ep(A, B)$,

(b) $Ep(A, B) \wedge Np(A) \quad \asymp \quad Np(A) \wedge Np(B)$.

Proof. We prove (a) only.

$$
\begin{aligned}
&Ep^*(A, B) \cup (Np^*(A) \cap Np^*(B)) \\
&\quad = \{\emptyset, \{A, B\}\} \otimes 2^{\mathcal{W}-\{A,B\}} \\
&\qquad \cup (2^{\mathcal{W}-\{A\}} \cap 2^{\mathcal{W}-\{B\}}) \\
&\quad = (\{\emptyset, \{A, B\}\} \cup (\{\emptyset, \{B\}\} \cap \{\emptyset, \{A\}\})) \\
&\qquad \otimes 2^{\mathcal{W}-\{A,B\}} \\
&\quad = \{\emptyset, \{A, B\}\} \otimes 2^{\mathcal{W}-\{A,B\}} \quad = \quad Ep^*(A, B)
\end{aligned}
$$

□

In the above proposition, Eq. (a) directly reduces the number of Dirichlet trees since a disjunction ($\vee$) disappears, while Eq. (b) indirectly reduces since $(Np(A) \wedge Np(B)) \vee Np(B) = Np(B)$.

We conduct an experiment to clarify how many trees can be reduced by asymptotic equivalency. In the experiment, we prepare conjunctions of random links of *ML*s and *CL*s when $|\mathcal{W}| = 10$, and compare the average numbers of Dirichlet trees compiled by minimum DNF (M-DNF) and asymptotic minimum DNF (AM-DNF) in 100 trials. The experimental result shown in Tab. 1

Table 1: The average numbers of Dirichlet trees compiled by minimum DNF (M-DNF) and asymptotic minimum DNF (AM-DNF) in terms of the number of random links. Each value is the average of 100 trials.

# of links	1	2	4	8	16
M-DNF	1	2.08	3.43	6.18	10.35
AM-DNF	1	2.08	3.23	4.24	4.07

indicates that asymptotic equivalency effectively reduces the number of Dirichlet trees especially when the number of links is large.

3.3 Customizing New Links

Two primitives *Ep* and *Np* allow us to easily customize new links without changing the algorithm. Let us consider *Imply-Link*(A, B) or *IL*(A, B), which is a constraint that B must appear if A appears in a topic (informally, $A \to B$). In this case, the setting

$$\mathit{IL}(A, B) \quad = \quad \mathit{Ep}(A, B) \vee \mathit{Np}(A)$$

is acceptable, since the ATF of *IL*(A, B) with respect to $\mathcal{W} = \{A, B\}$ is $\{\emptyset, \{A, B\}, \{B\}\}$. *IL*$(A, B)$ is effective when B has multiple meanings as mentioned later in Sec. 4.

Informally regarding *IL*(A, B) as $A \to B$ and *ML*(A, B) as $A \Leftrightarrow B$, *ML*(A, B) seems to be the same meaning of *IL*$(A, B) \wedge$ *IL*(B, A). However, this anticipation is wrong on the normal equivalency, i.e., *ML*$(A, B) \neq$ *IL*$(A, B) \wedge$ *IL*(B, A). The asymptotic equivalency can fulfill the anticipation with the next proposition. This simultaneously suggests that our definition is semantically valid.

Proposition 5. *For any two words* $A, B \in \mathcal{W}$,

$$\mathit{IL}(A, B) \wedge \mathit{IL}(B, A) \quad \asymp \quad \mathit{ML}(A, B)$$

Proof. From Proposition 4,

$$\begin{aligned} & \mathit{IL}(A, B) \wedge \mathit{IL}(B, A) \\ = \quad & (\mathit{Ep}(A, B) \vee \mathit{Np}(A)) \wedge (\mathit{Ep}(B, A) \vee \mathit{Np}(B)) \\ = \quad & \mathit{Ep}(A, B) \vee (\mathit{Ep}(A, B) \wedge \mathit{Np}(A)) \\ & \vee (\mathit{Ep}(A, B) \wedge \mathit{Np}(B)) \vee (\mathit{Np}(A) \wedge \mathit{Np}(B)) \\ \asymp \quad & \mathit{Ep}(A, B) \vee (\mathit{Np}(A) \wedge \mathit{Np}(B)) \\ \asymp \quad & \mathit{Ep}(A, B) \ = \ \mathit{ML}(A, B) \end{aligned}$$

$\square$

Further, we can construct *XIL*$(X_1, \cdots, X_n, Y)$ as an extended version of *IL*(A, B), which allows us to use multiple conditions like Horn clauses. This informally means $\bigwedge_{i=1}^{n} X_i \to Y$ as an extension of $A \to B$. In this case, we set

$$\mathit{XIL}(X_1, \cdots, X_n, Y) = \bigwedge_{i=1}^{n} \mathit{Ep}(X_i, Y) \vee \bigvee_{i=1}^{n} \mathit{Np}(X_i).$$

When we want to isolate unnecessary words (i.e., stop words), we can use *Isolate-Link* (*ISL*) defined as

$$\mathit{ISL}(X_1, \cdots, X_n) \quad = \quad \bigwedge_{i=1}^{n} \mathit{Np}(X_i).$$

This is easier than considering *CL*s between high-frequency words and unnecessary words as described in (Andrzejewski et al., 2009).

3.4 Negation of Links

There are two types of interpretation for negation of links. One is *strong negation*, which regards $\neg$*ML*(A, B) as "A and B must not appear in the same topic", and the other is *weak negation*, which regards it as "A and B need not appear in the same topic". We set $\neg$*ML*$(A, B) \asymp$ *CL*(A, B) for strong negation, while we just remove $\neg$*ML*(A, B) for weak negation. We consider the strong negation in this study.

According to Def. 1, the ATF of the negation $\neg f$ of primitive f seems to be defined as $(\neg f)^* := 2^{\mathcal{W}} - f^*$. However, this definition is not fit in strong negation, since $\neg$*ML*$(A, B) \not\asymp$ *CL*(A, B) on the definition. Thus we define it to be fit in strong negation as follows.

Definition 6 (ATF of strong negation of links)**.** *Given a link L with arguments $X_1, \cdots, X_n$, letting f_L be the primitives of L, we define the ATF of the negation of L as* $(\neg L(X_1, \cdots, X_n))^* := (2^{\mathcal{W}} - f_L^*(X_1, \cdots, X_n)) \cup 2^{\mathcal{W} - \{X_1, \cdots, X_n\}}$.

Note that the definition is used not for primitives but for links. Actually, the similar definition for primitives is not fit in strong negation, and so we must remove all negations in a preprocessing stage.

The next proposition gives the way to remove the negation of each link treated in this study. We define no constraint condition as ϵ for the result of *ISL*.

Proposition 7. *For any words* $A, B, X_1, \cdots, X_n, Y \in \mathcal{W}$,

(a) $\neg ML(A,B) \asymp CL(A,B)$,

(b) $\neg CL(A,B) \asymp ML(A,B)$,

(c) $\neg IL(A,B) \asymp Np(B)$,

(d) $\neg XIL(X_1,\cdots,X_n,Y) \asymp \bigwedge_{i=1}^{n-1} Ep(X_i,X_n) \wedge Np(Y)$,

(e) $\neg ISL(X_1,\cdots,X_n) \asymp \epsilon$.

Proof. We prove (a) only.

$$\begin{aligned}
&(\neg ML(A,B))^* \\
&= (2^{\mathcal{W}} - Ep^*(A,B)) \cup 2^{\mathcal{W}-\{A,B\}} \\
&= (2^{\{A,B\}} - \{\emptyset,\{A,B\}\}) \otimes 2^{\mathcal{W}-\{A,B\}} \\
&\qquad \cup 2^{\mathcal{W}-\{A,B\}} \\
&= \{\emptyset,\{A\},\{B\}\} \otimes 2^{\mathcal{W}-\{A,B\}} \\
&= 2^{\mathcal{W}-\{A\}} \cup 2^{\mathcal{W}-\{B\}} \\
&= Np^*(A) \cup Np^*(B) \quad = \quad (CL(A,B))^*
\end{aligned}$$

□

4 Comparison on a Synthetic Corpus

We experiment using a synthetic corpus $\{ABAB, ACAC\} \times 2$ with vocabulary $\mathcal{W} = \{A,B,C\}$ to clarify the property of our method in the same way as in the existing work (Andrzejewski et al., 2009). We set topic size as $T = 2$. The goal of this experiment is to obtain two topics: a topic where A and B frequently occur and a topic where A and C frequently occur. We abbreviate the grouping type as $AB|AC$. In preliminary experiments, LDA yielded almost four grouping types: $AB|AC$, $AB|C$, $AC|B$, and $A|BC$. Thus, we naively classify a grouping type of each result into the four types. Concretely speaking, for any two topic-word probabilities $\hat{\phi}$ and $\hat{\phi}'$, we calculate the average of Euclidian distances between each vector component of $\hat{\phi}$ and the corresponding one of $\hat{\phi}'$, ignoring the difference of topic labels, and regard them as the same type if the average is less than 0.1.

Fig. 2 shows the occurrence rates of grouping types on 1,000 results after 1,000 iterations by LDA-DF with six constraints (1) no constraint, (2) $ML(A,B)$, (3) $CL(B,C)$, (4) $ML(A,B) \wedge CL(B,C)$, (5) $IL(B,A)$, and (6) $ML(A,B) \vee ML(A,C)$. In the experiment, we set $\alpha = 1$, $\beta = 0.01$, and $\eta = 100$. In the figure, the higher rate of the objective type $AB|AC$ (open bar) is better. The results of (1-4) can be achieved even by the existing method, and those of (5-6) can be achieved only by our method. Roughly speaking, the figure shows that our method is clearly better than the existing method, since our method can obtain almost 100% as the rate of $AB|AC$, which is the best of all results, while the existing methods can only obtain about 60%, which is the best of the results of (1-4).

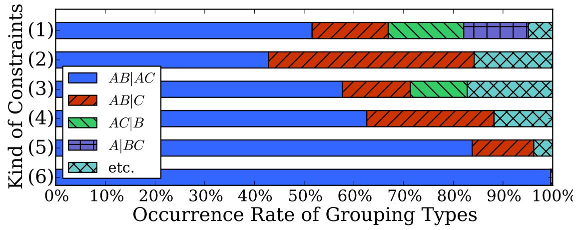


Figure 2: Rates of Grouping types in the 1,000 results on synthetic corpus $\{ABAB, ACAC\} \times 2$ with six constraints: (1) no constraint, (2) $ML(A,B)$, (3) $CL(B,C)$, (4) $ML(A,B) \wedge CL(B,C)$, (5) $IL(B,A)$, and (6) $ML(A,B) \vee ML(A,C)$.

The result of (1) is the same result as LDA, because of no constraints. In the result, the rate of $AB|AC$ is only about 50%, since each of $AB|C$, $AC|B$, and $A|BC$ remains at a high 15%. As we expected, the result of (2) shows that $ML(A,B)$ cannot remove $AB|C$ although it can remove $AC|B$ and $A|BC$, while the result of (3) shows that $CL(B,C)$ cannot remove $AB|C$ and $AC|B$ although it can remove $A|BC$. The result of (4) indicates that $ML(A,B) \wedge CL(B,C)$ is the best of knowledge expressions in the existing method. Note that $ML(A,B) \wedge ML(A,C)$ implies $ML(B,C)$ by transitive law and is inconsistent with all of the four types. The result (80%) of (5) $IL(B,A)$ is interestingly better than that (60%) of (4), despite that (5) has less primitives than (4). The reason is that (5) allows A to appear with C, while (4) does not. In the result of (6) $ML(A,B) \vee ML(A,C)$, the constraint achieves almost 100%, which is the best of knowledge expressions in our method. Of course, the constraint of $(ML(A,B) \vee ML(A,C)) \wedge CL(B,C)$ can also achieve almost 100%.

5 Interactive Topic Analysis

We demonstrate advantages of our method via interactive topic analysis on a real corpus, which consists of stemmed, down-cased 1,000 (positive) movie reviews used in (Pang and Lee, 2004). In this experiment, the parameters are set as $\alpha = 1$, $\beta = 0.01$, $\eta = 1000$, and $T = 20$.

We first ran LDA-DF with 1,000 iterations without any constraints and noticed that most topics have stop words (e.g., 'have' and 'not') and corpus-specific, unnecessary words (e.g., 'film', 'movie'), as in the first block in Tab. 2. To remove them, we added $\mathit{ISL}(\text{'film'}, \text{'movie'}, \text{'have'}, \text{'not'}, \text{'n't'})$ to the constraint of LDA-DF, which is compiled to one Dirichlet tree. After the second run of LDA-DF with the isolate-link, we specified most topics such as Comedy, Disney, and Family, since cumbersome words are isolated, and so we noticed that two topics about Star Wars and Star Trek are merged, as in the second block. Each topic label is determined by looking carefully at high-frequency words in the topic. To split the merged two topics, we added $\mathit{CL}(\text{'jedi'}, \text{'trek'})$ to the constraint, which is compiled to two Dirichlet trees. However, after the third run of LDA-DF, we noticed that there is no topic only about Star Trek, since 'star' appears only in the Star Wars topic, as in the third block. Note that the topic including 'trek' had other topics such as a topic about comedy film Big Lebowski. We finally added $\mathit{ML}(\text{'star'}, \text{'jedi'}) \lor \mathit{ML}(\text{'star'}, \text{'trek'})$ to the constraint, which is compiled to four Dirichlet trees, to split the two topics considering polysemy of 'star'. After the fourth run of LDA-DF, we appropriately obtained two topics about Star Wars and Star Trek as in the fourth block. Note that our solution is not ad-hoc, and we can easily apply it to similar problems.

Table 2: Characteristic topics obtained in the experiment on the real corpus. Four blocks in the table corresponds to the results of the four constraints ϵ, $\mathit{ISL}(\cdots)$, $\mathit{CL}(\text{'jedi'}, \text{'trek'}) \land \mathit{ISL}(\cdots)$, and $(\mathit{ML}(\text{'jedi'}, \text{'trek'}) \lor \mathit{ML}(\text{'star'}, \text{'trek'})) \land \mathit{CL}(\text{'jedi'}, \text{'trek'}) \land \mathit{ISL}(\cdots)$, respectively.

Topic	High frequency words in each topic
?	**have** give night **film** turn performance
?	**not** life **have** own first only family tell
?	**movie have n't** get good **not** see
?	**have** black scene tom death die joe
?	**film have n't not** make out well see
Isolated	**have film movie not** good make **n't**
?	**star war trek** planet effect special
Comedy	comedy funny laugh school hilarious
Disney	disney voice mulan animated song
Family	life love family mother woman father
Isolated	**have film movie not** make good **n't**
StarWars	**star war** lucas effect **jedi** special
?	science world **trek** fiction lebowski
Comedy	funny comedy laugh get hilarious
Disney	disney truman voice toy show
Family	family father mother boy child son
Isolated	**have film movie not** make good **n't**
StarWars	**star war** toy **jedi** menace phantom
StarTrek	alien effect **star** science special **trek**
Comedy	comedy funny laugh hilarious joke
Disney	disney voice animated mulan
Family	life love family man story child

6 Conclusions

We proposed a simple method to achieve topic models with logical constraints on words. Our method compiles a given constraint to the prior of LDA-DF, which is a recently developed semi-supervised extension of LDA with Dirichlet forest priors. As well as covering the constraints in the original LDA-DF, our method allows us to construct new customized constraints without changing the algorithm. We proved that our method is asymptotically the same as the existing method for any constraints with conjunctive expressions, and showed that asymptotic equivalency can shrink a constructed Dirichlet forest. In the comparative study on a synthetic corpus, we clarified the property of our method, and in the interactive topic analysis on a movie review corpus, we demonstrated its effectiveness. In the future, we intend to address detail comparative studies on real corpora and consider a simple method integrating negations into a whole, although we removed them in a preprocessing stage in this study.

References

David Andrzejewski and Xiaojin Zhu. 2009. Latent Dirichlet Allocation with Topic-in-Set Knowledge. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, pages 43–48.

David Andrzejewski, Anne Mulhern, Ben Liblit, and Xiaojin Zhu. 2007. Statistical Debugging Using Latent Topic Models. In *Proceedings of the 18th European Conference on Machine Learning (ECML 2007)*, pages 6–17. Springer-Verlag.

David Andrzejewski, Xiaojin Zhu, and Mark Craven. 2009. Incorporating domain knowledge into topic modeling via Dirichlet Forest priors. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML 2009)*, pages 25–32. ACM.

Sugato Basu, Ian Davidson, and Kiri Wagstaff. 2008. *Constrained Clustering: Advances in Algorithms, Theory, and Applications*. Chapman & Hall/CRC, 1 edition.

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

Samuel Y. Dennis III. 1991. On the hyper-Dirichlet type 1 and hyper-Liouville distributions. *Communications in Statistics — Theory and Methods*, 20(12):4069–4081.

Ivan Meza-Ruiz and Sebastian Riedel. 2009. Jointly Identifying Predicates, Arguments and Senses using Markov Logic. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2009)*, pages 155–163. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL 2004)*, pages 271–278.

Hoifung Poon and Pedro Domingos. 2009. Unsupervised Semantic Parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, pages 1–10. Association for Computational Linguistics.

Sebastian Riedel and Ivan Meza-Ruiz. 2008. Collective Semantic Role Labelling with Markov Logic. In *Proceedings of the 12th Conference on Computational Natural Language Learning (CoNLL 2008)*, pages 193–197. Association for Computational Linguistics.

Kristina Toutanova and Mark Johnson. 2008. A Bayesian LDA-based model for semi-supervised part-of-speech tagging. In *Advances in Neural Information Processing Systems 20*, pages 1521–1528. MIT Press.

Katsumasa Yoshikawa, Sebastian Riedel, Masayuki Asahara, and Yuji Matsumoto. 2009. Jointly Identifying Temporal Relations with Markov Logic. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2009)*, pages 405–413. Association for Computational Linguistics.

Xiaofeng Yu, Wai Lam, and Shing-Kit Chan. 2008. A Framework Based on Graphical Models with Logic for Chinese Named Entity Recognition. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP 2008)*, pages 335–342.