Topic Models for Unsupervised Cluster Matching

Tomoharu Iwata, Tsutomu Hirao, and Naonori Ueda

Abstract—We propose topic models for unsupervised cluster matching, which is the task of finding matching between clusters in different domains without correspondence information. For example, the proposed model finds correspondence between document clusters in English and German without alignment information, such as dictionaries and parallel sentences/documents. The proposed model assumes that documents in all languages have a common latent topic structure, and there are potentially infinite number of topic proportion vectors in a latent topic space that is shared by all languages. Each document is generated using one of the topic proportion vectors and language-specific word distributions. By inferring a topic proportion vector used for each document, we can allocate documents in different languages into common clusters, where each cluster is associated with a topic proportion vector. Documents assigned into the same cluster are considered to be matched. We develop an efficient inference procedure for the proposed model based on collapsed Gibbs sampling. The effectiveness of the proposed model is demonstrated with real data sets including multilingual corpora of Wikipedia and product reviews.

Index Terms—topic modeling, unsupervised object matching, clustering

1 Introduction

There has been great interest in topic models for analyzing discrete data such as text documents [1], [2]. Topic models are successfully used in a wide variety of applications including information retrieval [3], collaborative filtering [4] and image analysis [5], [6].

In this paper, we propose a topic model for unsupervised object matching for bag-of-words data. Object matching is an important task for finding correspondence between objects in different domains. Examples of object matching include matching vocabulary in different languages [7], matching images and annotations [8], and matching user identifications in different databases [9]. When similarity measures between objects in different domains, or correspondence data for learning similarity measures, are given, we can find matching using them by using record linkage methods [10]. However, in some applications, similarity measures and correspondence data might be unavailable because of cost or privacy issues.

For this situation, a number of unsupervised object matching methods have been proposed recently, such as kernelized sorting [11] and matching canonical correlation analysis [12], which can find correspondence without alignment information. These methods find only one-to-one matching. However, some applications require many-to-many, or cluster-to-cluster, matching. For example, multiple English words with the same meaning (e.g. car, automobile, motorcar) correspond to multiple German words (e.g. wagen, automobil). We also might need to find correspondence between groups of people instead of individuals.

The proposed model is an unsupervised method for cluster matching, which is the task of finding matching between clusters in different domains, where correspondence and cluster information are unavailable. For example, the proposed model finds correspondence between document clusters in English and German without alignment information, such as dictionaries and parallel sentences/documents. Here, parallel sentences/documents mean that its German translation is attached to each sentence/document in English. A number of topic models for multilingual corpora have been proposed [13], [14], [15], [16]. However, these models require alignment information. To our knowledge, the proposed model is the first topic model that can find shared topics across different languages without alignment information. In real applications, we might not have alignment information. For example, there are no dictionaries between minor languages, creating parallel corpora requires high cost, and morphological similarities cannot be used for languages that use different alphabets. Another example is matching user clusters in different companies, where a user is represented by a set of products the user purchased. Since user and product identifications are different in different companies, there are no alignment information.

With the proposed model, a latent topic space is shared across all languages by considering that documents in all languages have a common latent topic structure. In the latent topic space, there are potentially infinite number of topic proportion vectors, and each document is generated using one of the topic proportion vectors and language-specific word distributions. By inferring a topic proportion vector used for each document, we can allocate documents in different languages into common clusters, where each cluster is associated with a topic proportion vector. Documents assigned into the same cluster are considered to be matched. Figure 1 shows the framework of the proposed model. We use Dirichlet processes, which enable us to determine the number of clusters in the inference, and we do not need to fix the number of clusters in advance. We develop an efficient inference procedure for the proposed model based on collapsed Gibbs sampling, where sampling of a topic proportion vector assignment for each document and sampling of a topic assignment for each word are alternately iterated.

The remainder of this paper is organized as follows. In Section 2, we review related work on unsupervised object

• *T. Iwata, T. Hirao, and N. Ueda are with NTT Communication Science Laboratories, 2-4 Hikaridai, Sorakugun, Seikacho, Kyoto, Japan 619-0237. E-mail: iwata.tomoharu,hirao.tsutomu,ueda.naonori@lab.ntt.co.jp*

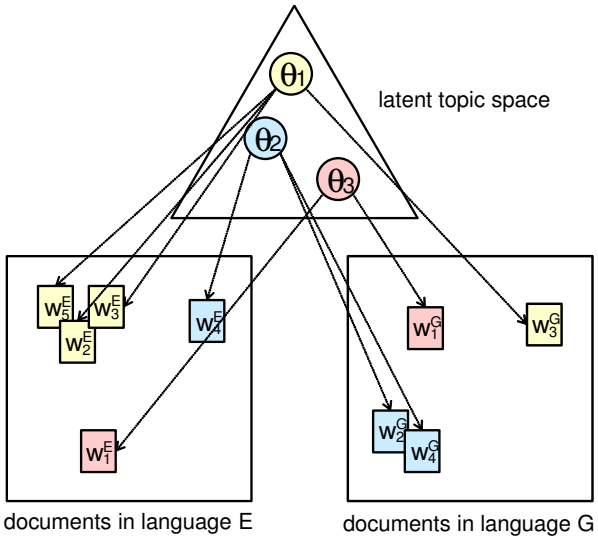


Fig. 1. The framework of the proposed model with bilingual data. The proposed model has a latent topic space shared by different languages. Each document $\mathbf{w}$ is generated depending on a topic proportion vector $\boldsymbol{\theta}$. Documents associated with the same topic proportion vector are considered as matched.

matching and topic modeling. We formulate the proposed model for unsupervised cluster matching in Section 3, and describe its inference procedure based on a Bayesian framework in Section 4. In Section 5, we demonstrate the effectiveness of the proposed model with experiments using document data sets including multilingual corpora of Wikipedia and product reviews. Finally, we present concluding remarks and a discussion of future work in Section 6.

2 Related Work

2.1 Unsupervised Object Matching

There have been proposed a number of methods for unsupervised object matching, which is also called cross-domain object matching, such as kernelized sorting [11], and its convex extension [17], least square object matching [18], matching canonical correlation analysis [12], and Bayesian object matching [19], [20]. These methods find matching by sorting objects so as to maximize dependence, or minimize independence. For example, kernelized sorting uses Hilbert-Schmidt Independence Criterion (HSIC) [21] as a measurement of independence, and sorts objects by minimizing HSIC between objects in two domains using the Hungarian algorithm [22]. There are three limitations in these methods. First, they find only one-to-one matching. Second, the number of domains need to be two. Third, the number of objects in each domain should be the same across all domains. On the other hand, the proposed model does not have these limitations; it finds cluster matching from data with more than two domains, and each domain can contain different numbers of objects.

Recently, [23] proposed an unsupervised many-to-many matching method based on probabilistic latent variable models (MMLVM). However, since it assumes Gaussian noise, it is inappropriate for discrete data such as document collections as shown by [3] and our experiments in Section 5. Another advantage of the proposed model over MMLVM is its efficiency for analyzing documents. The computational complexity of MMLVM is linear to the vocabulary size, which is usually very large. In contest, the complexity of the proposed model is linear to the number of words occurred in a document, which is much smaller than the total vocabulary size. ReMatch [24] is an unsupervised cluster matching method for network data, but not for bag-of-words data, which is our focus. ReMatch finds cluster matching using a probabilistic model, where connectivity between clusters is shared among different domains, and binary edges in networks are assumed to be generated from a Bernoulli distribution. On the other hand, the proposed model assumes words are generated from topic-specific multinomial distributions, and therefore it is able to utilize word frequency information in bag-of-words data.

TABLE 1
Notation.

Symbol	Description
M	number of languages
K	number of topics
D_m	number of documents in language m
V_m	vocabulary size of language m
N_{md}	number of words in document d in language m
w_{mdn}	nth word of document d in language m, $w_{mdn} \in \{1, \cdots, V_m\}$
z_{mdn}	topic for the nth word of document d in language m, $z_{mdn} \in \{1, \cdots, K\}$
s_{md}	index of the topic proportion vector used for document d in language m, $s_{md} \in \{1, \cdots, \infty\}$
$\theta_{\ell k}$	probability of topic k for the ℓth topic proportion vector
ϕ_{mkv}	probability of the vth word in topic k with language m
π_ℓ	probability of cluster ℓ

2.2 Topic Modeling

A number of topic models for modeling documents in multiple languages have been proposed. For example, polylingual topic models [13] find shared topics from aligned documents, and are used for cross lingual entity linking [25]. Probabilistic cross-lingual latent semantic analysis [14] and Joint latent Dirichlet allocation [15] discover shared latent topics by incorporating a bilingual dictionary into topic models. Unsupervised multilingual topic models (MuTo) can analyze unaligned text documents by matching vocabulary terms between different languages based on topic models [16]. MuTo finds correspondences of vocabulary terms between different languages using Hungarian algorithm by maximizing a posterior. When MuTo used prior knowledge such as dictionaries or morphological features, it found matching of semantically similar words in different languages, and discovered coherent topics across different languages. However, when MuTo did not use the prior knowledge, it could not find matching in their experiments.

3 Proposed Model

Although we assume that the given data are text documents with multiple languages in this paper, where each language corresponds to a domain, the proposed model is applicable to a wide range of discrete data, such as image data, where

each image is represented by visual words [26], and purchase log data, where each user is represented by a set of items the user purchased.

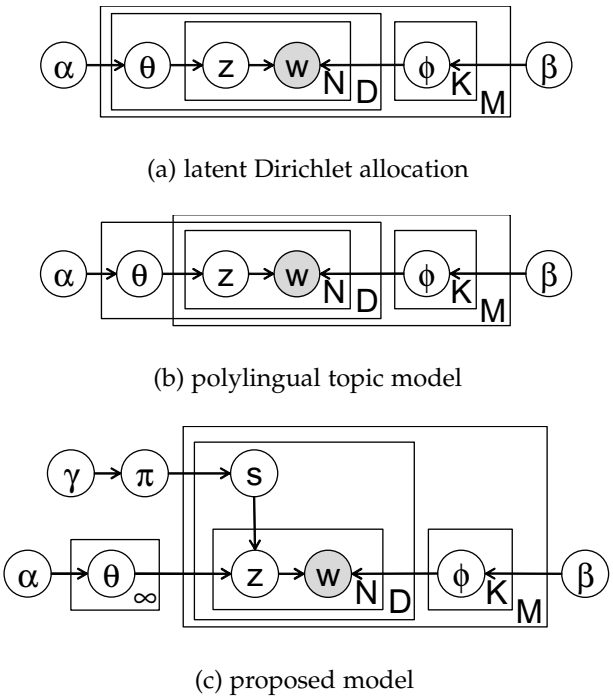
Suppose that we are given documents in M languages $\mathbf{W} = (\mathbf{W}_m)_{m=1}^M$, where $\mathbf{W}_m = (\mathbf{w}_{md})_{d=1}^{D_m}$ is a set of text documents in language m, and $\mathbf{w}_{md} = (w_{mdn})_{n=1}^{N_{md}}$ is a set of words in document d of language m. Our notation is summarized in Table 1. Note that correspondences between documents in different languages and correspondences between vocabulary terms in different languages are not given. The number of documents N_m and the vocabulary size V_m for each language can be different from those of other languages. The task is to find matching clusters of documents across multiple languages in an unsupervised manner.

The proposed model is assumed to have potentially infinite number of topic proportion vectors $\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_\infty$ in a latent topic space shared by all languages. Here, $\boldsymbol{\theta}_\ell$ is a K-dimensional vector, $\theta_{\ell k}$ represents the probability of generating topic k for the ℓth topic proportion vector, $\theta_{\ell k} \geq 0$ and $\sum_{k=1}^K \theta_{\ell k} = 1$. We use a stick-breaking process, which is a way of constructing Dirichlet processes [27]. Let $s_{md} \in \{1, \cdots, \infty\}$ be the latent index of a topic proportion vector for document d in language m. It means that the document d is generated using topic proportion vector $\boldsymbol{\theta}_{s_{md}}$. A topic proportion vector can be used by different documents in different languages. The generative process is the same with that of latent Dirichlet allocation (LDA) given the topic proportions. For each of the N_{md} words in the document, a topic z_{mdn} is chosen according to the topic proportions $\boldsymbol{\theta}_{s_{md}}$. Then word w_{mdn} is generated from a language- and topic-specific multinomial distribution over words $\boldsymbol{\phi}_{mz_{mdn}}$. Here, $\boldsymbol{\phi}_{mk}$ is a V_m-dimensional vector, ϕ_{mkv} represents the probability of generating word v in topic k, $\phi_{mkv} \geq 0$, and $\sum_{v=1}^{V_m} \phi_{mkv} = 1$.

Polylingual topic model (PTM) is a topic model for analyzing documents in multiple languages, such as multilingual corpora [13]. With PTM, documents need to be aligned across different languages. PTM assumes that aligned documents have the same topic proportion vector $\boldsymbol{\theta}_{1d} = \cdots = \boldsymbol{\theta}_{Md}$. Shared topics can be found with PTM by using a common topic proportion vector for aligned documents. However, alignment information is unavailable in our task, and therefore PTM is not applicable. On the other hand, we consider that topic proportion vector for each document is latent, which is indicated by s_{md}, and thus the proposed model can handle unaligned documents.

In summary, the proposed model generates documents in multiple languages $\mathbf{W}$ according to the following process,

1) Draw cluster proportions $\boldsymbol{\pi} \sim \text{Stick}(\gamma)$
2) For each cluster: $\ell = 1, \ldots, \infty$
   a) Draw a topic proportion vector $\boldsymbol{\theta}_\ell \sim \text{Dirichlet}(\alpha)$
3) For each language: $m = 1, \ldots, M$
   a) For each topic: $k = 1, \ldots, K$
      i) Draw a word distribution $\boldsymbol{\phi}_{mk} \sim \text{Dirichlet}(\beta)$
   b) For each document: $d = 1, \ldots, D_m$
      i) Draw a cluster assignment $s_{md} \sim \text{Discrete}(\boldsymbol{\pi})$
      ii) For each word: $n = 1, \ldots, N_{md}$
         A) Draw a topic $z_{mdn} \sim \text{Discrete}(\boldsymbol{\theta}_{s_{md}})$
         B) Draw a word $w_{mdn} \sim \text{Discrete}(\boldsymbol{\phi}_{mz_{mdn}})$



(a) latent Dirichlet allocation

(b) polylingual topic model

(c) proposed model

Fig. 2. Graphical model representation of latent Dirichlet allocation, the polylingual topic model, and the proposed model.

Here, Stick(γ) is the stick-breaking process that generates mixture weights for a Dirichlet process with concentration parameter γ. $\boldsymbol{\pi} = (\pi_1, \pi_2, \cdots)$ is a cluster proportion vector, where π_ℓ represents the probability of selecting cluster ℓ, or the probability of using the ℓth topic proportion vector, $\pi_\ell \geq 0$ and $\sum_{\ell=1}^\infty \pi_\ell = 1$. Dirichlet($\cdot$) represents the Dirichlet distribution, and α and β are Dirichlet parameters. Figure 2 shows graphical model representations of latent Dirichlet allocation (LDA), the polylingual topic model (PTM), and the proposed model, where shaded and unshaded nodes indicate observed and latent variables, respectively. With LDA, each document has a topic proportion vector $\boldsymbol{\theta}$, which is not shared across different languages. With PTM, a topic proportion vector $\boldsymbol{\theta}$ is shared across corresponded documents in different languages by using correspondence information, and align topics over different languages. With the proposed model, a topic proportion vector $\boldsymbol{\theta}$ is shared across documents assigned to the same cluster, which enables us to align topics without correspondence information.

The joint likelihood of words $\mathbf{W}$, latent topic assignments $\mathbf{Z} = (((z_{mdn})_{n=1}^{N_{md}})_{d=1}^{D_m})_{m=1}^M$ and latent clusters $\mathbf{S} = ((s_{md})_{d=1}^{D_m})_{m=1}^M$ is given by

$$p(\mathbf{W}, \mathbf{Z}, \mathbf{S} | \alpha, \beta, \gamma) = p(\mathbf{S}|\gamma) p(\mathbf{Z}|\mathbf{S}, \alpha) p(\mathbf{W}|\mathbf{Z}, \beta). \quad (1)$$

By analytically integrating out cluster proportions $\boldsymbol{\pi}$, the first factor is calculated by

$$p(\mathbf{S}|\gamma) = \frac{\gamma^L \prod_{\ell=1}^{L}(D_\ell - 1)!}{\gamma(\gamma+1)\cdots(\gamma+D-1)}, \tag{2}$$

where L is the number of clusters that contain more than a document, D_ℓ is the number of documents assigned to cluster ℓ, and $D = \sum_{m=1}^{M} D_m$ is the total number of documents. Since we use conjugate Dirichlet priors for multinomial parameters, $\boldsymbol{\Theta} = (\boldsymbol{\theta}_\ell)_{\ell=1}^{L}$ and $\boldsymbol{\Phi} = ((\boldsymbol{\phi}_{mk})_{k=1}^{K})_{m=1}^{M}$, we can integrate out $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ analytically in a similar way with latent Dirichlet allocation [2]. Then, the second factor of (1) is given by

$$p(\mathbf{Z}|\mathbf{S},\alpha) = \frac{\Gamma(\alpha K)^L}{\Gamma(\alpha)^{KL}} \prod_{\ell=1}^{L} \frac{\prod_{k=1}^{K}\Gamma(N_{\ell k}+\alpha)}{\Gamma(N_\ell + \alpha K)}, \tag{3}$$

where $N_{\ell k}$ is the number of words assigned to topic k in documents of cluster ℓ, $N_\ell = \sum_{k=1}^{K} N_{\ell k}$ and $\Gamma(\cdot)$ represents the Gamma function. The third factor of (1) is given by

$$p(\mathbf{W}|\mathbf{Z},\beta) = \prod_{m=1}^{M} \frac{\Gamma(\beta V_m)^K}{\Gamma(\beta)^{V_m K}} \prod_{k=1}^{K} \frac{\prod_{v=1}^{V_m}\Gamma(N_{mkv}+\beta)}{\Gamma(N_{mk}+\beta V_m)}, \tag{4}$$

where N_{mkv} is the number of times word v has been assigned to topic k in language m and $N_{mk} = \sum_{v=1}^{V_m} N_{mkv}$. See Appendix A for the derivation of (3) and (4).

Algorithm 1 Inference procedures for the proposed model.

Input: multiple language data sets $\boldsymbol{X}$, initial number of clusters L, number of topics K, hyperparameters α, β, γ, number of iterations T
Output: cluster assignments $\boldsymbol{S}$, topic assignments $\boldsymbol{Z}$

```
initialize S and Z
for t = 1, ..., T do
    //sampling topic assignments
    for m = 1, ..., M do
        for d = 1, ..., D_m do
            for n = 1, ..., N_md do
                sample z_mdn using probability (5) from {1, ..., K}
            end for
        end for
    end for
    //sampling cluster assignments
    for m = 1, ..., M do
        for d = 1, ..., D_m do
            sample s_md using (6) from {1, ..., L + 1}
            if s_md = L + 1 then
                update the number of clusters L ← L + 1
            end if
        end for
    end for
end for
```

4 Inference

Since cluster proportions $\boldsymbol{\pi}$, topic proportion vectors $\boldsymbol{\Theta}$ and word distributions $\boldsymbol{\Phi}$ can be analytically integrated out, latent variables that we need to infer are latent topic assignments $\mathbf{Z}$ and latent cluster assignments $\mathbf{S}$. The inference of $\mathbf{Z}$ and $\mathbf{S}$ given multilingual documents $\mathbf{W}$ can be efficiently computed using collapsed Gibbs sampling [2]. Given the current state of all but one variable z_{mdn}, the assignment of a latent topic to the nth word in document d of language m is sampled from the following probability:

$$\begin{aligned} &p(z_{mdn} = k|\mathbf{W},\mathbf{Z}_{\backslash mdn},\mathbf{S},\alpha,\beta,\gamma) \\ &\propto \left(N_{s_{md}k\backslash mdn}+\alpha\right)\cdot\frac{N_{mkw_{mdn}\backslash mdn}+\beta}{N_{mk\backslash mdn}+\beta V_m}, \end{aligned} \tag{5}$$

where $\backslash i$ represents the count or set when excluding example i. This probability is derived using (3) and (4). The sampling probability for the latent cluster s_{md} is as follows:

$$\begin{aligned} &p(s_{md} = \ell|\mathbf{W},\mathbf{Z},\mathbf{S}_{\backslash md},\alpha,\beta,\gamma) \\ &\propto \frac{p(s_{md}=\ell,\mathbf{S}_{\backslash md}|\gamma)}{p(\mathbf{S}_{\backslash md}|\gamma)}\cdot\frac{p(\mathbf{Z}|s_{md}=\ell,\mathbf{S}_{\backslash md},\alpha)}{p(\mathbf{Z}_{\backslash md}|\mathbf{S}_{\backslash md},\alpha)}. \end{aligned} \tag{6}$$

Here, the first factor is given by

$$\frac{p(s_{md}=\ell,\mathbf{S}_{\backslash md}|\gamma)}{p(\mathbf{S}_{\backslash md}|\gamma)} \propto \begin{cases} D_{\ell\backslash md} & \text{if } \ell \le L \\ \gamma & \text{if } \ell = L+1, \end{cases} \tag{7}$$

using (2), and the second factor is given by

$$\begin{aligned} &\frac{p(\mathbf{Z}|s_{md}=\ell,\mathbf{S}_{\backslash md},\alpha)}{p(\mathbf{Z}_{\backslash md}|\mathbf{S}_{\backslash md},\alpha)} \\ &= \frac{\Gamma(N_{\ell\backslash md}+\alpha K)}{\Gamma(N_{\ell\backslash md}+N_{md}+\alpha K)}\prod_{k=1}^{K}\frac{\Gamma(N_{\ell k\backslash md}+N_{mdk}+\alpha)}{\Gamma(N_{\ell k\backslash md}+\alpha)}, \end{aligned} \tag{8}$$

using (3). See Appendix B for the derivation.

Algorithm 1 shows the procedures for inferring the proposed model based on the collapsed Gibbs sampling. Here, T is the number of iterations. For the input, we give the initial number of clusters L. In our experiments, we set the initial number of clusters $L = 1$. The cluster assignments $\boldsymbol{S}$ and topic assignments $\boldsymbol{Z}$ are initialized by randomly selecting an integer from $\{1,\cdots,L\}$ and $\{1,\cdots,K\}$, respectively. By iterating collapsed Gibbs sampling of latent topic assignments z_{mdn} for all words $m = 1,\cdots,M$, $d = 1,\cdots,D_m$, $N = 1,\cdots,N_{md}$ with (5) and the sampling of latent clusters s_{md} for all documents $m = 1,\cdots,M$, $d = 1,\cdots,D_m$ with (6), we infer the latent variables of the proposed model. The documents that are assigned into the same cluster are considered as matched. The point estimate of topic proportion vectors and word distributions are obtained by

$$\hat{\theta}_{mdk} = \frac{N_{s_{md}k}+\alpha}{N_{s_{md}}+\alpha K}, \tag{9}$$

where $N_{s_{md}} = \sum_{k=1}^{K} N_{s_{md}k}$ is the number of words assigned to cluster s_{md}, and

$$\hat{\phi}_{mkv} = \frac{N_{mkv}+\beta}{N_{mk}+\beta V_m}, \tag{10}$$

respectively.

The computational complexity of an iteration of the collapsed Gibbs sampling with the proposed model is linear to the number of languages, the number of documents, the number of topics, the number of clusters, and the document length, but it does not depend on the vocabulary size.

5 EXPERIMENTS

We evaluated the proposed model by using two bilingual document sets. In the next section, we explain the data used. In Subsections 5.2 and 5.2, we show the discovered topics and quantitative results, respectively, when we fix the hyperparameters and data sets. In Subsection 5.4, we analyze the effect of hyperparameter and data settings.

5.1 Data

We used the following two data sets: Wikipedia and Review. The Wikipedia data set consists of Wikipedia documents written in English and German. For each language, we sampled 150 documents that were categorized in 'Nobel laureates in Physics', 'Nobel laureates in Chemistry', 'American basketball players', 'American composers', and 'English footballers'. The Review data set consists of review documents in English and Japanese obtained from Amazon.com and Amazon.co.jp. The reviews are written about products categorized in 'Watch', 'Book', 'Electronics', 'Kitchen' and 'Music'. We sampled 1,000 documents for each category, and 5,000 documents in total. Stop-words were omitted for both of the data sets, where stop word lists were downloaded from https://code.google.com/archive/p/stop-words. We used 1,000 and 3,000 most frequently occurring words as features in Wikipedia and Review data sets, respectively.

5.2 Discovered Shared Topics

Table 2 shows some examples of extracted shared topics with the proposed model from the Wikipedia data set. In our experiments, we used the following hyperparameters: $\alpha = 0.1, \beta = 1, \gamma = 1$. The initial number of clusters was set to one. The proposed model successfully found common topics between English and German without alignment information; Topic1 is about Nobel prize, Topic2 is about music, Topic 3 is about soccer, and Topic4 is about basketball. Two categories 'Nobel laureates in Physics' and 'Nobel laureates inn Chemistry' were joined in the first topic. This is reasonable because the two categories are closely related. Note that there are common or similar words in English and German, such as 'manchester', 'nba' and 'jordan', we do not use any morphological information for the inference.

The proposed model also discovered shared topics from the Review data set as shown in Table 3; Topic1 is about music, Topic2 is about watch/camera, Topic3 is about books, and Topic4 is about kitchen products. When matching morphologically similar languages such as English and German, we can use string similarities between words. However, when we try to match morphilogically different languages such as English and Japanese, we cannot use morphological features. Thus, it is important to develop methods for unsupervised matching.

5.3 Quantitative Results

For the quantitative evaluation, we used precision, recall and F-measure as measurements. The precision is calculated by the rate of correctly matched pairs among pairs that are estimated as matched. Here, a matched pair means that the two documents are assigned into the same cluster or category. The recall is calculated by the rate of correctly matched pairs among truly matched pairs. The F-measure is the harmonic mean of precision and recall. For all of the measurements, a higher value indicates a higher matching performance.

We compared the proposed model with MMLVM, Shared-LDA and Mix-KS. MMLVM is unsupervised many-to-many matching latent variable models [23]. Shared-LDA is a latent Dirichlet allocation model that shares a Dirichlet prior for topic proportions across different languages. The hyper-parameters for the Dirichlet prior were estimated by the fixed point iteration method. After the inference, topic proportion vectors were clustered using the k-means method. Mix-KS is a combination of mixture models and kernelized sorting. Mix-KS finds cluster matching by the following procedure. First, documents are clustered in each language independently using a mixture of multinomial distributions. Then, clusters are matched by convex kernelized sorting [17], which is an unsupervised object matching method, using the mean vector of each cluster as features. As the number of clusters in Shared-LDA and Mix-KS, we used the estimated number of clusters by the proposed model for a fair comparison. We also compared with Random as a baseline method, in which documents are randomly assigned to one of the clusters shared by all languages.

Table 4 shows the precision, recall and F-measures, which are averaged over 30 experiments using different sampled documents for each data set. For the proposed model, MMLVM and Shared-LDA, we set the number of topics at $K = 10$. The proposed model has achieved the highest precision, recall and F-measure for both data sets. Since MMLVM assumes Gaussian observation noise, the performance was low for analyzing text documents. Shared-LDA and Mix-KS find correspondence of clusters with two steps of clustering and matching. Therefore, errors accumulated in clustering cannot be corrected in matching process. On the other hand, since the proposed model performs clustering and matching simultaneously in one probabilistic framework, clusters are estimated so as to be optimal when documents from different languages are matched. Mix-KS might find different cluster structures for different languages because it performs clustering for each language separately and has local optima for each language. In contrast, by sharing clusters among different languages, the proposed model alleviates to be trapped in different local optima. The recall is low compared with the precision. This result indicates that documents in the same category are assigned into different clusters, and modeling a category requires multiple topic proportion vectors in these data sets.

The computational time of the proposed model was 1.5 hours with the Wikipedia data on PC with Xeon 5160 3GHz CPU, and that of MMLVM was 45.5 hours. With the Review data, the proposed model took 4.8 hours, and the MMLVM was not finished.

Table 5 (a) shows a confusion matrix for the Wikipedia data set, where clusters are estimated using the proposed model, and the total number of clusters was 11. The documents categorized in 'English footballers' in English and German are perfectly matched in the third cluster, where other documents are not assigned into the third cluster. This result is obtained because the documents in 'English

TABLE 2
Examples of topics extracted by the proposed model from the Wikipedia data. In each topic, the top row shows probable words in English, and the bottom row shows those in German.

Topic1
EN: nobel physics prize planck marconi laureates bardeen chemistry physicist max
DE: chemie physik otto physiker nobelpreis chemiker hochschullehrer preisverleihung nobelstiftung max
Topic2
EN: piano composer composers orchestra songs score composition lewis compositions string
DE: davis komponist love spiel jazz my album american nba grammy
Topic3
EN: cup robson players goals match manchester manager fa scored squad
DE: fc united saison verein manchester nationalmannschaft west mannschaft englischer trainer
Topic4
EN: nba basketball johnson coach players finals boston jordan draft championship
DE: nba jackson basketballspieler saison team bryant music bird jordan vereinigte

TABLE 3
Examples of topics extracted by the proposed model from Review data. In each topic, the top row shows probable words of the topic in English, the middle row shows those in Japanese, and the bottom row shows English translation of those in Japanese at the middle.

Topic1
EN: have song album do good track music sound great CD band get make love time hear vocal go other
JP: 曲 アルバム ない . 良い 今 人 収録 ライブ (- ファン 歌 CD 音楽 いい 歌詞 映像 好き)
[JP: track album no . good now person record live (- fan song CD music nice lyrics video like)]
Topic2
EN: have watch bag do camera case use look get good great fit time make lens small buy other strap need
JP: 購入 時計 良い ない 使用 - カメラ 機能 デザイン 撮影 いい 商品 大きい レンズ 値段 満足 価格 感じ
[JP: buy watch good no use - camera function design filming nice goods big lens price content feel]
Topic3
EN: have book do read story good character find get author love time other make work go write great know
JP: 文学 本 作品 人 ない 物語 自分 小説 本書 世界 面白い 心 作家 話 著者
[JP: literature book piece person no story self novel this-book world interesting heart storywriter tale author]
Topic4
EN: have do use make get good time great work water go cook machine buy coffee other unit product easy
JP: 購入 ない 機能 使用 良い パン - 製品 (値段 いい 今 商品 体重 / 自分 必要 便利 時間 前
[JP: buy no function use good bread - product (price nice now goods weight / self need useful time front]

TABLE 4
Average precision, recall and F-measure with their standard errors. Values in bold typeface are statistically better at the 5% level from those in normal typeface as indicated by a paired t-test.

(a) Precision

	Proposed	MMLVM	Shared-LDA	Mix-KS	Random
Wikipedia	$\mathbf{0.41} \pm 0.04$	$\mathbf{0.40} \pm 0.03$	0.20 ± 0.02	0.26 ± 0.04	0.20 ± 0.00
Review	$\mathbf{0.33} \pm 0.02$	NA	0.19 ± 0.01	0.19 ± 0.03	0.20 ± 0.00

(b) Recall

	Proposed	MMLVM	Shared-LDA	Mix-KS	Random
Wikipedia	$\mathbf{0.44} \pm 0.05$	0.15 ± 0.01	0.28 ± 0.03	0.16 ± 0.03	0.11 ± 0.01
Review	$\mathbf{0.36} \pm 0.03$	NA	0.29 ± 0.02	0.08 ± 0.02	0.03 ± 0.00

(c) F-measure

	Proposed	MMLVM	Shared-LDA	Mix-KS	Random
Wikipedia	$\mathbf{0.42} \pm 0.05$	0.22 ± 0.02	0.23 ± 0.02	0.19 ± 0.03	0.14 ± 0.00
Review	$\mathbf{0.34} \pm 0.02$	NA	0.23 ± 0.02	0.11 ± 0.02	0.06 ± 0.00

footballers' consist of distinct words from documents in other categories. On the other hand, the documents in 'Nobel laureates in Physics' and those in 'Nobel laureates in Chemistry' are clustered into the same cluster, because they are similar categories. The proposed model failed to distinguish 'Nobel laureates in Physics' from 'Nobel laureates in Chemistry'. However, documents from other categories were not assigned into the second cluster, and the proposed model succeeded to distinguish 'Nobel laureates' categories from the other categories.

Table 5 (b) shows a confusion matrix for the Review data set, where the total number of clusters was 25. For documents categorized in 'Books' and 'Music', the proposed model found matching between English and Japanese. However, many documents in 'Watch', 'Electronics' and 'Kitchen' are assigned to the same cluster because they are similar; 'Watch' can be seen as a part of 'Electronics', and some products in 'Kitchen' are electronics.

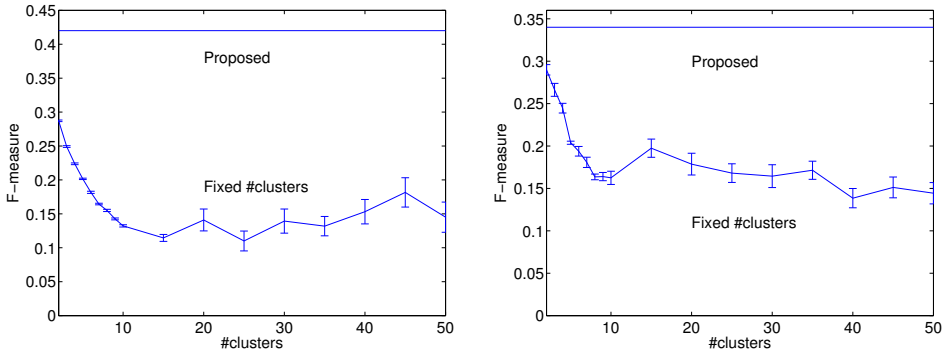
TABLE 5
Confusion matrices for (a) Wikipedia and (b) Review data sets. Each row represents categories of the given data, and each column represents cluster indices inferred by the proposed model. Thus, (c, ℓ) element shows the number of documents that are categorized in the cth category and assigned to the ℓth cluster. The left and right values in each element show the value for languages 1 and 2, respectively. We show clusters that contain more than 5% of total documents.

(a) Wikipedia

	1		2		3		4		5	
category \ cluster	EN	DE	EN	DE	EN	DE	EN	DE	EN	DE
Nobel laureates in Physics	0	4	30	25	0	0	0	0	0	0
Nobel laureates in Chemistry	17	3	11	25	0	0	0	0	0	2
American basketball players	0	0	0	0	0	0	0	3	27	27
American composers	0	14	0	1	0	0	18	5	0	10
English footballers	0	0	0	0	30	30	0	0	0	0

(b) Review

	1		2		3		4	
category \ cluster	EN	JP	EN	JP	EN	JP	EN	JP
Watch	943	951	4	2	1	2	2	7
Books	13	10	1	11	899	875	1	10
Electronics	986	918	0	3	1	1	0	31
Kitchen	427	787	2	3	5	1	463	127
Music	4	15	691	750	15	10	2	12



(a) Wikipedia (b) Review

Fig. 3. F-measure with different fixed numbers of clusters.

We evaluated the effectiveness of the Dirichlet process in the proposed model that enable us to determine the number of clusters in the inference. Figure 3 shows the F-measures with the proposed model and a method with the fixed number of clusters. The comparing method is the same with the proposed model except for that it does not use Dirichlet processes but fixes the number of clusters. The proposed model achieved the better F-measure than the comparing method with any number of fixed clusters. This is because the proposed model can adaptively change the number of clusters in the inference, which alleviates local optima, by using the Dirichlet processes. On the other hand, the comparing method is likely to be trapped into poor local optima since it cannot change the number of clusters. This result indicates that the Dirichlet process is useful for improving matching performance.

5.4 Sensitivity Analysis

Figure 4 shows the 1) F-measure, 2) computational time and 3) estimated number of clusters with (a) different numbers of categories, (b) different numbers of documents, (c) different vocabulary size, (d) different numbers of topics K, and (e) different concentration parameters γ. For (a), we used Wikipedia documents in English and German, where we kept the total number of documents at 400 while changing the number of categories. Here, we used 5,000 most frequently occurring words. The F-measure decreased as the number of categories increased (a-1) since matching many categories was difficult. The proposed model achieved higher F-measure than Shared-LDA and Random with different numbers of categories. The estimated number of clusters did not increase as the number of categories increased. This would be because categories in the given data does not correspond to clusters in our model. Some documents in the same category are clustered into multiple different clusters, which is also seen in the experiment shown in Table 5. Another reason would be that Dirichlet process mixtures are inconsistent for the number of clusters [28].

For (b)–(e), we used Amazon review documents in English and Japanese, where 1,000 documents were sampled from each of the five product categories, and 3,000 most frequently occurring words, $K = 10$ and $\gamma = 1$ were used as the basic settings. The F-measure decreased as the number of documents increased (b-1). It would be because the proposed model found too many clusters (b-3), which were finer than categories in the given data, when many documents were given. The computational time and the estimated number of clusters were increased as the number of documents increased (b-2, 3).

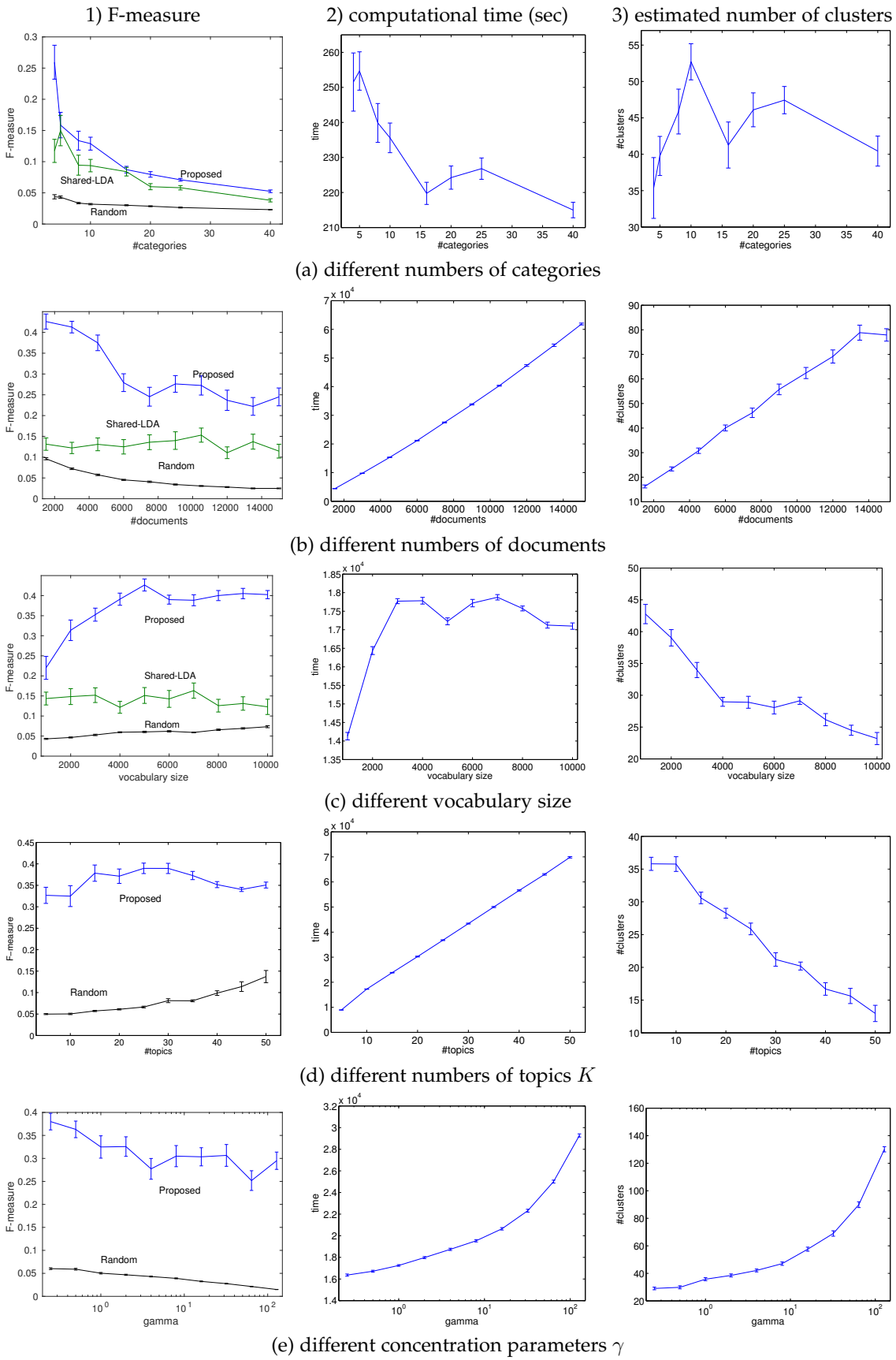


Fig. 4. 1) F-measure, 2) computational time, and 3) estimated number of clusters when we vary (a) the number of categories, (b) the number of documents, (c) vocabulary size, (d) the number of topics, and (e) the concentration parameter.

The F-measure of the proposed model increased as the vocabulary size increased (c-1) since we were able to utilize more information. The computational time was small when the vocabulary size was small $V = 1000$ (c-2), because the average document length was short. The computational time did not change after the vocabulary size was larger than 2,000 (c-2) since the decline of the estimated number of clusters (c-3) canceled out the effect of the gain of the average document length. For all data sets, the proposed model achieved higher F-measure than Shared-LDA and Random (a-1, b-1, c-1).

The F-measure was highest when the number of topics was $K = 30$ (d-1). The computational time linearly increased with the number of topics K (d-2). On the other hand, the computational time of MMLVM cubically increases with K since it calculates the inverse of a $(K \times K)$ matrix, which prohibits us to employ MMLVM with a large number of topics. When the number of topics was high, the estimated number of clusters was low (d-3). It would be because the given data were modeled with a small number of clusters by utilizing many topics.

The F-measure decreased as the concentration parameter increased (e-1) since many irrelevant clusters were generated with a high concentration parameter (e-3). When the concentration parameter γ increased, the estimated number of clusters increased (e-3), and therefore, the computational time increased (e-2).

6 Conclusion

We proposed a topic model to find cluster matching without alignment information for discrete data with multiple domains. The proposed model has a set of topic proportion vectors shared among different languages. By assigning a topic proportion vector for each document, documents in all languages are clustered in a common space. The documents assigned into the same cluster are considered as matched. In the experiments, we confirmed that the proposed model could perform better than a combination of clustering and unsupervised object matching. We also showed that the proposed model could extract shared topics from real multilingual text data sets without dictionaries and parallel documents.

For future work, we will extend the proposed model for a semi-supervised setting, where a small number of correspondence information is available. With the proposed model, the number of topics and concentration parameter are hyperparameters to be set by users. The number of topics can be inferred by using hierarchical Dirichlet processes [29] or nested Dirichlet processes [30]. The concentration parameter can be inferred by using Markov chain Monte Carlo methods assuming a gamma prior [29].

Appendix A
Derivation of (3) and (4)

The derivation of the probability of latent topics $\boldsymbol{Z}$ given latent clusters $\boldsymbol{S}$ and Dirichlet parameter α (3) is given by integrating out topic proportion vectors $\boldsymbol{\Phi}$ as follows,

$$\begin{aligned}
&p(\boldsymbol{Z}|\boldsymbol{S},\alpha)\\
&= \int \prod_{\ell=1}^{L} p(\boldsymbol{\theta}_\ell|\alpha) \prod_{m=1}^{M}\prod_{d=1}^{D_m}\prod_{n=1}^{N_{md}} p(z_{mdn}|\boldsymbol{\theta}_{s_{md}}) d\boldsymbol{\Theta}\\
&= \int \prod_{\ell=1}^{L} \frac{\Gamma(\alpha K)}{\Gamma(\alpha)^K} \prod_{k=1}^{K} \theta_{\ell k}^{\alpha-1} \prod_{m=1}^{M}\prod_{d=1}^{D_m}\prod_{n=1}^{N_{md}} \theta_{s_{md},z_{md}} d\boldsymbol{\Theta}\\
&= \frac{\Gamma(\alpha K)^L}{\Gamma(\alpha)^{KL}} \prod_{\ell=1}^{L}\prod_{k=1}^{K} \theta_{\ell k}^{N_{\ell k}+\alpha-1} d\boldsymbol{\Theta}\\
&= \frac{\Gamma(\alpha K)^L}{\Gamma(\alpha)^{KL}} \prod_{\ell=1}^{L} \frac{\prod_{k=1}^{K}\Gamma(N_{\ell k}+\alpha)}{\Gamma(N_\ell+\alpha K)}.
\end{aligned} \tag{11}$$

Here, in the second equation, we used the fact that $p(\boldsymbol{\theta}_\ell|\alpha)$ is a Dirichlet distribution and $p(z_{mdn}|\boldsymbol{\theta}_{s_{md}}) = \theta_{s_{md},z_{md}}$. In the fourth equation, we used $\int \prod_{k=1}^{K} \theta_k^{\alpha_k-1} d\boldsymbol{\theta} = \frac{\prod_{k=1}^{K}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K}\alpha_k)}$ which is the normalizing constant of the Dirichlet distribution. In a similar way, we can derive the probability of words $\boldsymbol{W}$ given latent topics $\boldsymbol{Z}$ and Dirichlet parameter β (4).

Similarly, the derivation of the probability of words $\boldsymbol{W}$ given latent topics $\boldsymbol{Z}$ and Dirichlet parameter β (4) is given as follows,

$$\begin{aligned}
&p(\boldsymbol{W}|\boldsymbol{Z},\beta)\\
&= \prod_{m=1}^{M} \int \prod_{k=1}^{K} p(\boldsymbol{\phi}_{mk}|\beta) \prod_{d=1}^{D_m}\prod_{n=1}^{N_{md}} p(w_{mdn}|\boldsymbol{\phi}_{md}) d\boldsymbol{\Phi}_m\\
&= \prod_{m=1}^{M} \int \prod_{k=1}^{K} \frac{\Gamma(\beta V_m)}{\Gamma(\beta)^{V_m}} \prod_{v=1}^{V_m} \phi_{mkv}^{\beta-1} \prod_{d=1}^{D_m}\prod_{n=1}^{N_{md}} \phi_{mdw_{mdn}} d\boldsymbol{\Phi}_m\\
&= \prod_{m=1}^{M} \frac{\Gamma(\beta V_m)^K}{\Gamma(\beta)^{V_m K}} \int \prod_{k=1}^{K}\prod_{v=1}^{V_m} \phi_{mkv}^{N_{mkv}+\beta-1} d\boldsymbol{\Phi}_m\\
&= \prod_{m=1}^{M} \frac{\Gamma(\beta V_m)^K}{\Gamma(\beta)^{V_m K}} \prod_{k=1}^{K} \frac{\prod_{v=1}^{V_m}\Gamma(N_{mkv}+\beta)}{\Gamma(N_{mk}+\beta V_m)}.
\end{aligned} \tag{12}$$

Appendix B
Derivation of (8)

The derivation of (8) is given as follows,

$$\begin{aligned}
&p(\boldsymbol{Z}|s_{md}=\ell,\boldsymbol{S}_{\backslash md},\alpha)\\
&= \frac{\Gamma(\alpha K)^L}{\Gamma(\alpha)^{KL}} \prod_{\ell'=1}^{L} \frac{\prod_{k=1}^{K}\Gamma(N_{\ell' k\backslash md}+\alpha)}{\Gamma(N_{\ell'\backslash md}+\alpha K)}\\
&\times \frac{\Gamma(N_{\ell\backslash md}+\alpha K)}{\prod_{k=1}^{K}\Gamma(N_{\ell k\backslash md}+\alpha)} \frac{\prod_{k=1}^{K}\Gamma(N_{\ell k\backslash md}+N_{mdk}+\alpha)}{\Gamma(N_{\ell\backslash md}+N_{md}+\alpha K)}\\
&= p(\boldsymbol{Z}_{\backslash md}|\boldsymbol{S}_{\backslash md},\alpha)\\
&\times \frac{\Gamma(N_{\ell\backslash md}+\alpha K)}{\Gamma(N_{\ell\backslash md}+N_{md}+\alpha K)} \prod_{k=1}^{K} \frac{\Gamma(N_{\ell k\backslash md}+N_{mdk}+\alpha)}{\Gamma(N_{\ell k\backslash md}+\alpha)}.
\end{aligned} \tag{13}$$

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[2] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101 Suppl 1, pp. 5228–5235, 2004.

[3] T. Hofmann, "Probabilistic latent semantic analysis," in *Proceedings of Conference on Uncertainty in Artificial Intelligence*, 1999, pp. 289–296.

[4] ——, "Collaborative filtering via Gaussian probabilistic latent semantic analysis," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003, pp. 259–266.

[5] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003, pp. 127–134.

[6] L. Cao and L. Fei-Fei, "Spatially coherent latent topic model for concurrent object segmentation and classification." in *Proceedings of IEEE Intern. Conf. in Computer Vision (ICCV)*, 2007.

[7] A. Tripathi, A. Klami, and S. Virpioja, "Bilingual sentence matching using kernel CCA," in *MLSP '10: Proceedings of the 2010 IEEE International Workshop on Machine Learning for Signal Processing*, 2010, pp. 130–135.

[8] R. Socher and L. Fei-Fei, "Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR, 2010, pp. 966–973.

[9] B. Li, Q. Yang, and X. Xue, "Transfer learning for collaborative filtering via a rating-matrix generative model," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09, 2009, pp. 617–624.

[10] I. P. Fellegi and A. B. Sunter, "A theory for record linkage," *Journal of the American Statistical Association*, vol. 64, no. 328, pp. 1183–1210, 1969.

[11] N. Quadrianto, A. J. Smola, L. Song, and T. Tuytelaars, "Kernelized sorting," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, no. 10, pp. 1809–1821, 2010.

[12] A. Haghighi, P. Liang, T. Berg-Kirkpatrick, and D. Klein, "Learning bilingual lexicons from monolingual corpora," in *Proceedings of ACL-08: HLT*, 2008, pp. 771–779.

[13] D. Mimno, H. M. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum, "Polylingual topic models," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 2009, pp. 880–889.

[14] D. Zhang, Q. Mei, and C. Zhai, "Cross-lingual latent topic extraction," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 2010, pp. 1128–1137.

[15] J. Jagarlamudi and H. Daumé III, "Extracting multilingual topics from unaligned comparable corpora," in *Advances in Information Retrieval*. Springer, 2010, pp. 444–456.

[16] J. Boyd-Graber and D. Blei, "Multilingual topic models for unaligned text," in *Proceedings of Conference on Uncertainty in Artificial Intelligence*, 2009, pp. 75–82.

[17] N. Djuric, M. Grbovic, and S. Vucetic, "Convex kernelized sorting," in *Proceedings of AAAI Conference on Artificial Intelligence*, 2012.

[18] M. Yamada and M. Sugiyama, "Cross-domain object matching with model selection," in *Proceedings of International Conference on Articial Intelligence and Statistics*, 2011, pp. 807–815.

[19] A. Klami, "Variational Bayesian matching," in *Proceedings of Asian Conference on Machine Learning*, 2012, pp. 205–220.

[20] ——, "Bayesian object matching," *Machine learning*, vol. 92, pp. 225–250, 2013.

[21] A. Smola, A. Gretton, L. Song, and B. Schölkopf, "A Hilbert space embedding for distributions," in *Algorithmic Learning Theory*, 2007, pp. 13–31.

[22] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[23] T. Iwata, T. Hirao, and N. Ueda, "Unsupervised cluster matching via probabilistic latent variable models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2013.

[24] T. Iwata, J. Lloyd, and Z. Ghahramani, "Unsupervised many-to-many object matching for relational data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 3, pp. 607–619, 2016.

[25] T. Zhang, K. Liu, and J. Zhao, "Cross lingual entity linking with bilingual topic model," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2013, pp. 2218–2224.

[26] G. Csurka, C. Dance, J. Willamowski, L. Fan, and C. Bray, "Visual categorization with bags of keypoints," in *ECCV International Workshop on Statistical Learning in Computer Vision*, 2004.

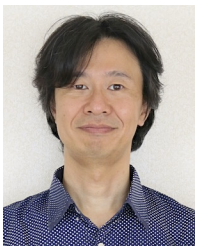
[27] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

[28] J. W. Miller and M. T. Harrison, "A simple example of dirichlet process mixture inconsistency for the number of components," in *Advances in Neural Information Processing Systems*, 2013, pp. 199–206.

[29] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[30] A. Rodriguez, D. B. Dunson, and A. E. Gelfand, "The nested Dirichlet process," *Journal of the American Statistical Association*, vol. 103, no. 483, 2008.

Tomoharu Iwata received the B.S. degree in environmental information from Keio University in 2001, the M.S. degree in arts and sciences from the University of Tokyo in 2003, and the Ph.D. degree in informatics from Kyoto University in 2008. In 2003, he joined NTT Communication Science Laboratories, Japan. From 2012 to 2013, he was a visiting researcher at Machine Learning Group, Department of Engineering, University of Cambridge, UK. He is currently a senior research scientist (distinguished researcher) at Ueda Research Laboratory of NTT Communication Science Laboratories, Kyoto, Japan. His research interests include data mining, machine learning, information visualization, and recommender systems.



Tsutomu Hirao received the B.E. from Kansai University in 1995, M.E. and Ph.D. in Engineering from Nara Institute of Science and Technology in 1997 and 2002, respectively. He is currently with NTT Communication Science Laboratories. His current research interests include Natural Language Processing and Machine Learning.

Naonori Ueda received the B.S., M.S., and Ph D degrees in Communication Engineering from Osaka University, Osaka, Japan, in 1982, 1984, and 1992, respectively. In 1984, he joined the Electrical Communication Laboratories, NTT, Japan, where he was engaged in research on image processing, pattern recognition, and computer vision. In 1991, he joined the NTT Communication Science Laboratories. From 1993 to 1994, he was a visiting scholar at Purdue University, West Lafayette, USA. He was a director of NTT Communication Science Laboratories(April,2010-March,2013). Currently, he is a head of Ueda Research Laboratory (NTT Fellow), and is a director of Machine Learning and Data Science Center. He also serves as a Deputy Director, RIKEN Center for Advanced Intelligence Project, newly established in April, 2016. He is a visiting professor, Graduate School of Informatics, Kyoto University, and National Institute of Informatics (NII). He is a member of the Information Processing Society of Japan(IPSJ), a fellow of the Institute of Electronics, Information, and Communication Engineers in Japan(IEICE), and a senior member of IEEE.