PAPER Special Section on Knowledge, Information and Creativity Support System

Entity Network Prediction Using Multitype Topic Models

Hitohiro SHIOZAKI[†], *Nonmember*, **Koji EGUCHI**[†a)], *and* **Takenao OHKAWA**[†], *Members*

SUMMARY Conveying information about *who*, *what*, *when* and *where* is a primary purpose of some genres of documents, typically news articles. Statistical models that capture dependencies between named entities and topics can play an important role. Although some relationships between *who* and *where* should be mentioned in such a document, no statistical topic models explicitly address in handling such information the textual interactions between a *who*-entity and a *where*-entity. This paper presents a statistical model that directly captures the dependencies between an arbitrary number of word types, such as *who*-entities, *where*-entities and topics, mentioned in each document. We show that this multitype topic model performs better at making predictions on entity networks, in which each vertex represents an entity and each edge weight represents how a pair of entities at the incident vertices is closely related, through our experiments on predictions of *who*-entities and links between them. We also demonstrate the scale-free property in the weighted networks of entities extracted from written mentions.

key words: *statistical topic models, multitype topic models, link prediction, entity networks*

1. Introduction

The primary purpose of documents such as news articles that report factual events is to convey information on *who*, *what*, *when* and *where*. Statistical entity-topic models [1] capture the dependencies between the named entities, in such documents, that usually represent information on *who* or *where* and latent topics that often convey information on *what*. In spite of the fact that each entity type has different characteristics and so it has a different distribution, these models represent all types of entities as a single class. This paper attempts to directly capture dependencies between multiple types of entities, especially (1) *who*-entities, such as persons, organizations and nationalities, (2) *where*-entities, such as locations, geographical/social/political entities and facilities, and (3) other general words.

In this paper, we review a couple of statistical topic models: one of which is called Latent Dirichlet Allocation (LDA) [2], and the other is its variant, SwitchLDA [1] that explicitly models entities mentioned in text. We then develop a multitype topic model that can explicitly capture dependencies between an arbitrary number of word types, such as *who*-entity type, *where*-entity type and general word type. As in [1] we take advantage of recent developments in named entity recognition to identify entities mentioned in articles. We demonstrate that our model can predict *who*-entities more effectively, comparing with two other different topic models. We also exhibit that links between entities can be effectively predicted using our model. We further demonstrate the scale-free property in edge-weighted networks of entities extracted from textual data.

2. Related Work

Statistical topic models (e.g., [2]–[6]) are based on the idea that documents are mixtures of topics, where a topic is a probability distribution over words. Blei et al. [2] proposed one of the topic models called Latent Dirichlet Allocation (LDA), introducing Dirichlet priors on a multinomial distribution over topics for each document and that over words for each topic. More recently, Newman et al. [1] proposed several statistical entity-topic models, extending the LDA model. *SwitchLDA* is one of them. Those models attempted to capture dependencies between entities and topics, where the entities are *mentioned* in text; however, the models did not distinguish specific types of entities, such as *who*-entities and *where*-entities. Therefore, those models are hardly sufficient to represent factual events, each of which consists of multiple types of entities. On the other hand, our goal is to model the events that are mentioned in text. As a step towards this goal, this paper develops a multitype topic model by extending the models mentioned above to represent dependencies between an arbitrary number of word types, such as *who*-entity type, *where*-entity type and general word type.

One of the objectives of this paper is to explicitly generalize multitype topic models with an arbitrary number of word types, as an extension of Newman et al.'s SwitchLDA [1]. This direction was mentioned previously [1]; but it was never actually investigated. We demonstrated through a couple of different experiments that this generalization is especially crucial to adequately model factual events, such as where person names and location names play important roles. We also analyzed edge-weighted entity networks that are constructed using the multitype topic models.

Topic allocations over words are usually unobserved in a document collection, and so we need to infer the unknown distributions from the documents. To estimate the LDA model or its variants, Variational Bayesian inference [2] or Collapsed Variational Bayesian inference [7] can be used. Gibbs sampling method is an alternative approach to esti-

Manuscript received March 31, 2008.

Manuscript revised July 4, 2008.

[†]The authors are with Kobe University, Kobe-shi, 657–8501 Japan.

a) E-mail: eguchi@port.kobe-u.ac.jp

DOI: 10.1093/ietisy/e91–d.11.2589

mate the LDA model [5]. From a viewpoint of perplexity of the estimated models, the Gibbs sampling method works better than the others above when a sufficient number of iterations are performed [7]. Model estimation is not the main focus of this paper, and so we used the Gibbs sampling approach in this paper.

3. Models

In this section we describe three graphical models. We start with LDA, followed by SwitchLDA and GESwitchLDA. The LDA is a popular model that can automatically infer a set of topics from a collection of documents [2]. The SwitchLDA was modeled by extending the LDA to capture dependencies between entities and topics, and its prediction performance was shown to be stable over different corpora [1]. The third model, GESwitchLDA is our model that aims to better fit textual data with multiple types of expressions, such as of *who*-entities, *where*-entities and general words, by generalizing the SwitchLDA model. We use the LDA [2] as a baseline model for comparing with our GESwitchLDA in the experiments in Sect. 4. We also use the SwitchLDA as another baseline model.

Here we introduce the notation used in graphical models, generative processes and Gibbs sampling equations in the rest of this paper: D is the number of documents, T is the number of topics, and N_d is the total number of words in document d. θ indicates a per-document topic distribution, ϕ a per-topic word distribution, and ψ a per-topic word type distribution. α and β are hyperparameters of Dirichlet priors, and γ is a hyperparameter of Beta or Dirichlet prior. In the case of the SwitchLDA, a tilde mark is used to denote the entity version of a variable. In the case of the GESwitchLDA, a tilde mark and a hat mark are used to denote the *who*-entity version and *where*-entity version, respectively.

3.1 LDA

To explain the differences between the three graphical models, let us start with the LDA model shown in Fig. 1. The LDA's generative process is:

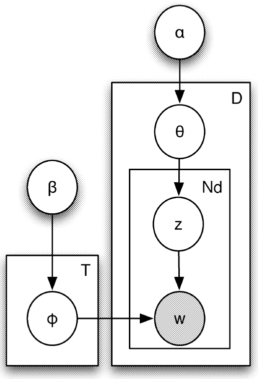


Fig. 1 LDA.

1. For all d documents sample $\theta_d \sim Dirichlet(\alpha)$
2. For all t topics sample $\phi_t \sim Dirichlet(\beta)$
3. For each of the N_d words w_i in document d:
   a. Sample a topic $z_i \sim Multinomial(\theta_d)$
   b. Sample a word $w_i \sim Multinomial(\phi_{z_i})$

Some estimation algorithms were applied to the LDA [2], [5], [7]. Following [5], we use the Gibbs sampling to estimate the LDA model. Note that the LDA does not distinguish specific types of words, and so this distinction was made at post-processing stage (i.e., outside of the model) when we made predictions about *who*-entities in Sect. 4.

3.2 SwitchLDA

SwitchLDA model shown in Fig. 2 was introduced in [1], extending the LDA model. In this model, an additional binomial distribution ψ (with a Beta prior of γ) was incorporated to control the fraction of entities in topics. The generative process of the SwitchLDA is:

1. For all d documents sample $\theta_d \sim Dirichlet(\alpha)$
2. For all t topics sample $\phi_t \sim Dirichlet(\beta)$, $\tilde{\phi}_t \sim Dirichlet(\tilde{\beta})$ and $\psi_t \sim Beta(\gamma)$
3. For each of the N_d words w_i in document d:
   a. Sample a topic $z_i \sim Multinomial(\theta_d)$
   b. Sample a flag $x_i \sim Binomial(\psi_{z_i})$
   c. If $(x_i = 0)$ sample a word $w_i \sim Multinomial(\phi_{z_i})$
   d. If $(x_i = 1)$ sample an entity $w_i \sim Multinomial(\tilde{\phi}_{z_i})$

where x_i is a binary indicator of whether word w_i is an entity or not. The estimation algorithm for the SwitchLDA followed the Gibbs sampling approach, as described in [1]. Note that the SwitchLDA does not distinguish more specific types of entities, and so this distinction was made at post-processing stage (i.e., outside of the model) when we made predictions about *who*-entities in Sect. 4.

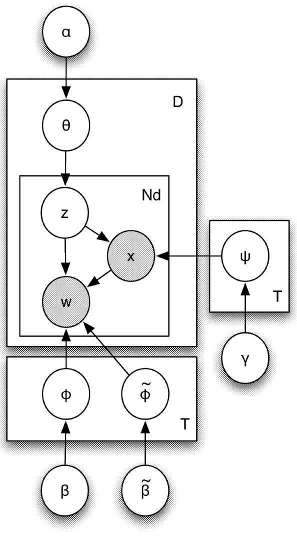


Fig. 2 SwitchLDA.

3.3 GESwitchLDA

In our GESwitchLDA model shown in Fig. 3, we generalize the SwitchLDA to handle an arbitrary number (M) of word types, in order to achieve more flexible modeling of latent topics for type-annotated documents, in which all the word types are labeled. Therefore, instead of using binomial distribution ψ with Beta prior distribution specified by γ that were used in the SwitchLDA model, we redefine ψ as multinomial distribution over M word types with Dirichlet prior specified by γ. The multinomial distribution ψ gives the fraction of word types on a given topic, or the probability that a randomly chosen word on a given topic falls in a specific word type. The generative process of the GESwitchLDA is:

1. For all d documents sample $\theta_d \sim \mathit{Dirichlet}(\alpha)$
2. For all t topics:
   a. Sample $\psi_t \sim \mathit{Dirichlet}(\gamma)$
   b. For each word type $y \in \{0, \cdots, M-1\}$, sample $\phi_t^y \sim \mathit{Dirichlet}(\beta^y)$
3. For each of the N_d words w_i in document d:
   a. Sample a topic $z_i \sim \mathit{Multinomial}(\theta_d)$
   b. Sample a word type $x_i \sim \mathit{Multinomial}(\psi_{z_i})$
   c. For each word type $y \in \{0, \cdots, M-1\}$:
      - If $(x_i = y)$ sample a type-y word $w_i \sim \mathit{Multinomial}(\phi_{z_i}^y)$

where x_i indicates the word type with which w_i is labeled.

The word w and word type x are observed variables, as you can see in the graphical model representation of Fig. 3; however, since the topic z is a latent variable, the following have to be inferred statistically:

- θ_d : distribution over topics given document d,
- ψ_z : distribution over word types given topic z,
- ϕ_z^x : distribution over words given word type x and topic z.

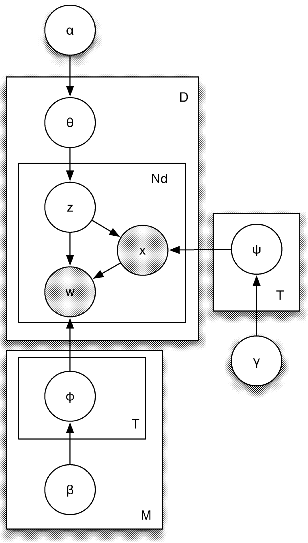


Fig. 3 GESwitchLDA.

We estimated the unknown distributions above using Gibbs sampling in an unsupervised manner, as briefly described in Appendix.

In the experiments in Sect. 4, we divided entities into two classes, *who*-entity and *where*-entity, and thus the number of word types $M = 3$ in this case. The GESwitchLDA's generative process when $M = 3$ is:

1. For all d documents sample $\theta_d \sim \mathit{Dirichlet}(\alpha)$
2. For all t topics sample $\phi_t \sim \mathit{Dirichlet}(\beta)$, $\tilde{\phi}_t \sim \mathit{Dirichlet}(\tilde{\beta})$, $\hat{\phi}_t \sim \mathit{Dirichlet}(\hat{\beta})$ and $\psi_t \sim \mathit{Dirichlet}(\gamma)$
3. For each of the N_d words w_i in document d:
   a. Sample a topic $z_i \sim \mathit{Multinomial}(\theta_d)$
   b. Sample a word type $x_i \sim \mathit{Multinomial}(\psi_{z_i})$
   c. If $(x_i = 0)$ sample a word $w_i \sim \mathit{Multinomial}(\phi_{z_i})$
   d. If $(x_i = 1)$ sample a *who*-entity $w_i \sim \mathit{Multinomial}(\tilde{\phi}_{z_i})$
   e. If $(x_i = 2)$ sample a *where*-entity $w_i \sim \mathit{Multinomial}(\hat{\phi}_{z_i})$

4. Experiments

4.1 Data Sets

Our focus is unsupervised topic modeling over type-annotated documents, in which all the word types are totally labeled, and so we assume here that the named entity tagging is already performed. For our experiments we used the TDT2 and TDT3 collections [8], in which named entities were tagged by the BBN Identifinder [9]. They originally contained a mix of broadcast news and newswire stories. We used only the English stories in these collections, not the stories in other languages or the metadata such as pre-defined topics and categories. We used the TDT2 for training and the TDT3 for testing. Statistics for the data sets are summarized in Table 1. We removed the 418 stopwords included in the stop list used in *InQuery* system [10], and also removed words and entities that occurred in less than 10 documents.

4.2 *Who*-Entity Prediction

Who-entity prediction task is to fill in blanks with words

Table 1 Statistics for data sets.

	TDT2	TDT3
Documents	45,260	26,770
Unique Words	27,685	21,954
Unique *Who*-entities	7,300	4,591
Unique *Where*-entities	1,637	1,121
Total Words	7,634,722	4,583,162
Total *Who*-entities	600,638	378,725
Total *Where*-entities	343,432	199,760

Table 2 *Who*-entity prediction results example. The top row shows an excerpt from an article, with redacted *who*-entities indicated by XXXXX. Middle row shows the list of relevant *who*-entities. The bottom row shows the predicted *who*-entity list ordered by likelihood.

The XXXXX accord and XXXXX Camry are the most popular for buyers and auto thieves. More on that from XXXXX . The latest XXXXX figures show that auto thefts were down overall in , 1997 . By 4% , in fact. But that is little solace for the owners of the cars that topped the national insurance crime bureau's list of most stolen automobiles in the United States. The XXXXX accord and XXXXX Camry occupy the number one and two spots on the list.
actual who-entity list: Honda, Toyota, Charles Feldman, FBI, CNN
predicted who-entity list: Italian, U.N., General Motors, Pakistani, GM, Chrysler , Americans , Indian , American , Ford , Supreme Court , Smith , U.S. , VOA , Congress , Annan , United Nations , Japanese , *FBI, *CNN, Volkswagen , *Honda, European , BMW , Security Council , *Toyota

in each test document, in which all the blanks are known to be *who*-entity type and all the other words are known to be either *where*-entity or general words, as described in Sect. 4.2.1. GESwitchLDA directly captures dependencies between word types, as described in Sect. 3.3. On the other hand, SwitchLDA and LDA indirectly do so as post-processing, as noted in Sects. 3.1 and 3.2. Therefore, this task reveals how well these models involve the word-type dependencies. For each test document, the predicted *who*-entities are ranked in order of likelihood, and so information retrieval evaluation metrics can be used for the evaluation for the *who*-entity prediction task, as described in Sect. 4.2.2.

4.2.1 Estimation and Prediction

We illustrate the process of the *who*-entity prediction in Table 2 using an example from the TDT data. The first row shows an excerpt from an article of the TDT3, with *who*-entities indicated by XXXXX. Middle row shows the list of actual *who*-entities. The bottom row shows the predicted *who*-entity list ordered by the likelihood computed using both words and *where*-entities (or using only words).

For the *who*-entity prediction task, the three models: the LDA, the SwitchLDA and the GESwitchLDA are first trained on words, *who*-entities, and *where*-entities using the TDT2 collection. The models then make predictions about *who*-entities over the TDT3 collection in the following two ways:

1. using words and *where*-entities ("w+o").
2. using only words ("w").

We need to set the number of topics and hyperparameters for the LDA, as well as for the SwitchLDA, and the GESwitchLDA. For all of the experiments, we set the number of topics $T = 100$, 200, and 300 for each of the three models. We fixed Dirichlet prior hyperparameters $\alpha = 50/T$ and $\beta = 0.01$, which were reported to be appropriate for various collections [6]. The other hyperparameters were empirically determined using the training data TDT2. Some examples of the topics captured by the GESwitchLDA are shown in Table 3.

The likelihood of a *who*-entity in each test document is calculated by $P(e|d) = \sum_t P(e|t)P(t|d)$, where $P(e|t)$ is estimated during training via Gibbs sampling, and the topic mixture in the test document $P(t|d)$ is estimated by resampling both all words and all *where*-entities (or by resampling only all words) using learned word distribution $P(w|t)$ and *where*-entity distribution $P(o|t)$.

4.2.2 Evaluation Metrics

After the model estimation, the models computed the likelihood of every possible *who*-entity, and then listed the *who*-entities in order of the likelihood. We computed MAP (mean average precision) [11], and GMAP (geometric mean average precision) [12], as well as average best rank and average median rank [1].

The MAP measure is given by the following equation:

$$\frac{1}{|\mathbf{d}|}\sum_{d\in\mathbf{d}} AvgPrec(d) \tag{1}$$

where **d** is a set of test documents, $AvgPrec(d)$ is given by the average precision, as below, of the predicted *who*-entities in each test document d.

$$AvgPrec = \frac{1}{|\mathbf{r}|}\sum_{r\in\mathbf{s}} Prec(r) \tag{2}$$

where **s** is a set of ranks of predicted *who*-entities for a test document, **r** is a set of all relevant *who*-entities that actually appear in the test document, and $Prec(r)$ gives the precision at a given cut-off rank r. MAP is a very well accepted evaluation criterion in information retrieval. It is also known to be stable and understandable.

The GMAP measure is the geometric mean version of the MAP measure, which is given by:

$$\exp\left(\frac{1}{|\mathbf{d}|}\sum_{q\in\mathbf{d}} \log AvgPrec(d)\right) \tag{3}$$

GMAP prefers robustness of the prediction.

The average best rank is defined as the average of the best rank of relevant *who*-entities, and the average median rank is the average rank of *who*-entities at median of relevant *who*-entity ranked list.

Table 3 Examples of topics captured by GESwitchLDA. In each topic, we list most likely words and their probability at the top, *who*-entities at the middle, and *where*-entities at the bottom.

oil	0.1746	internet	0.0820	game	0.0408
prices	0.0669	web	0.0568	team	0.0379
production	0.0381	information	0.0457	coach	0.0308
price	0.0325	site	0.0454	basketball	0.0273
gas	0.0315	mail	0.0266	tournament	0.0229
crude	0.0215	sites	0.0261	national	0.0212
barrels	0.0182	online	0.0187	play	0.0179
cut	0.0176	computer	0.0166	college	0.0179
world	0.0146	service	0.0132	season	0.0166
silver	0.0142	users	0.0122	points	0.0157
gasoline	0.0133	world	0.0105	final	0.0133
barrel	0.0132	data	0.0096	win	0.0131
pipeline	0.0128	electronic	0.0092	championship	0.0122
natural	0.0123	wide	0.0092	point	0.0122
cents	0.0120	line	0.0091	four	0.0120
OPEC	0.4140	America_Online	0.1271	Duke	0.0528
Texaco	0.1185	Reuters	0.1197	John	0.0518
Berkshire	0.0893	Bloomberg	0.0540	Stanford	0.0391
Shell	0.0536	Yahoo	0.0488	Kentucky	0.0332
Exxon	0.0503	AOL	0.0443	NCAA	0.0321
crisco	0.0503	NYT	0.0392	Rutgers	0.0307
Pertamina	0.0455	Excite	0.0325	Huskies	0.0303
Buffett	0.0422	Amazon.com	0.0310	Big_East	0.0217
Caspian	0.0422	Online	0.0281	UConn	0.0184
Chevron	0.0406	Holmes	0.0229	Wildcats	0.0184
Turkmenistan	0.0972	Cambridge	0.0942	North_Carolina	0.1261
Saudi_Arabia	0.0938	Va.	0.0779	St.	0.1258
Caspian	0.0914	Honolulu	0.0747	Connecticut	0.0810
Azerbaijan	0.0868	Fla.	0.0714	Kentucky	0.0661
Olean	0.0845	Bridge	0.0649	Utah	0.0619
Venezuela	0.0752	Amazon	0.0617	Michigan	0.0474
Mexico	0.0590	San_Francisco	0.0552	Princeton	0.0455
Caspian_Sea	0.0579	Calif.	0.0487	Rhode_Island	0.0436
Ecuador	0.0556	Dayton	0.0455	Arizona	0.0409
Baku	0.0498	Mass.	0.0422	Tennessee	0.0229

Table 4 Best results of *who*-entity prediction (without name identification).

model	MAP	GMAP	avg best rank	avg median rank
LDA (w+o, T=300)	0.1998	0.0818	118.10	482.93
SwitchLDA (w+o, T=300)	0.2036	0.0816	119.78	484.38
GESwitchLDA (w+o, T=300)	0.2048	0.0833	119.08	480.64
LDA (w, T=200)	0.1565	0.0558	135.13	549.86
SwitchLDA (w, T=300)	0.1603	0.0568	136.98	565.48
GESwitchLDA (w, T=300)	0.1595	0.0569	135.55	560.18

4.2.3 Results

The best results for the LDA, SwitchLDA and GESwitchLDA models are shown in Table 4. To obtain the best results, we determined through experiments that $T = 300$ was the best parameter for all the three models, except the case of the LDA using only words. We determined that $T = 200$ was the best parameter for the LDA using only words. We determined the best parameters $\tilde{\beta} = \hat{\beta} = 0.01$ for both the SwitchLDA and the GESwitchLDA, $\gamma = 5.0$ for the SwitchLDA, and $\gamma = 4.0$ for the GESwitchLDA.

Given the best parameters in our experiments, our GESwitchLDA model gave the best results, in terms of both MAP and GMAP, over the other two models in the case of using both words and *where*-entities for prediction. In terms of MAP, the GESwitchLDA gave 2.5% improvement in this case [†], comparing with the best results of the LDA model under the same condition. We further performed the Wilcoxon signed-rank test (two-tailed) to the pair of GESwitchLDA - LDA and the pair of GESwitchLDA - SwitchLDA. In terms of MAP, the resulting p-values of these pairs were less than 0.01 in the case of using both words and *where*-entities. It means the the performance improvement of the GESwitchLDA over both the SwitchLDA and the LDA was statistically significant, in this case. As for the case of using only words, the improvement of the GESwitchLDA over the LDA was also statistically significant at 0.01 level; however, that over the SwitchLDA was not. In terms of average best rank and average median rank, we observed that few very bad results made performance values unfairly poor. In contrast, MAP was observed to be more stable in this sense.

We also calculated likelihood of *who*-entities in the manner of not using resampling. In detail, we calculated

[†] Although this value is apparently small, it is statistically significant, as described later. One possible reason for this small value is that the evaluation values were averaged all over a large number of test documents, as shown in Table 1; also, the predicted entities that did not appear in a document were deemed to be irrelevant even if some were closely related to the document content.

Table 5 Best results of *who*-entity prediction without resampling (without name identification).

model	MAP	GMAP	avg best rank	avg median rank
GESwitchLDA (w+o, T=300)	0.1970	0.0784	110.17	461.82
GESwitchLDA (w, T=300)	0.1554	0.0613	120.72	516.01

Table 6 Best results of *who*-entity prediction with name identification of GESwitchLDA.

model	MAP	GMAP	avg best rank	avg median rank
GESwitchLDA (w+o, T=300)	0.2141	0.0893	114.21	439.21
GESwitchLDA (w, T=300)	0.1611	0.0605	128.28	505.01

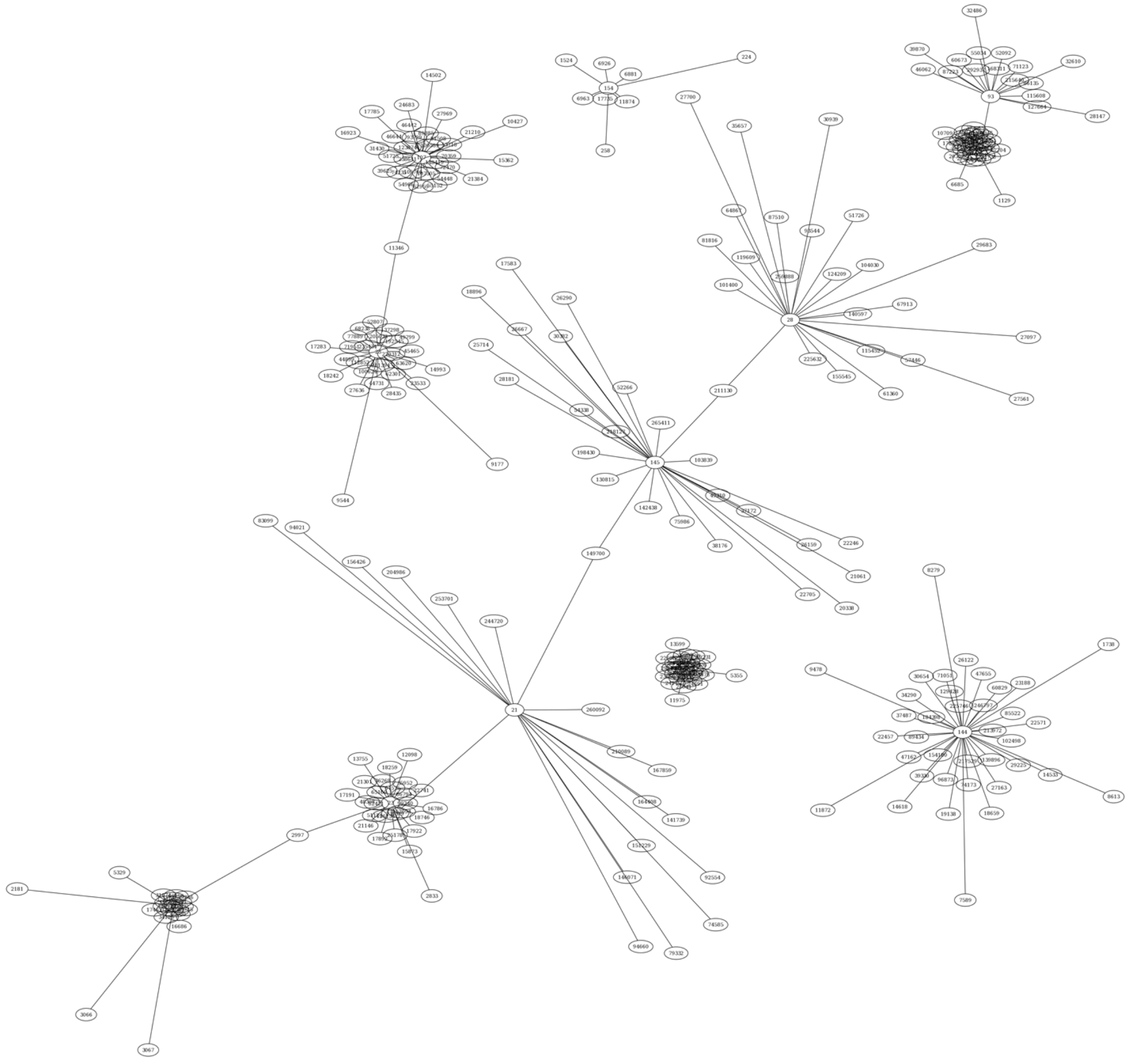
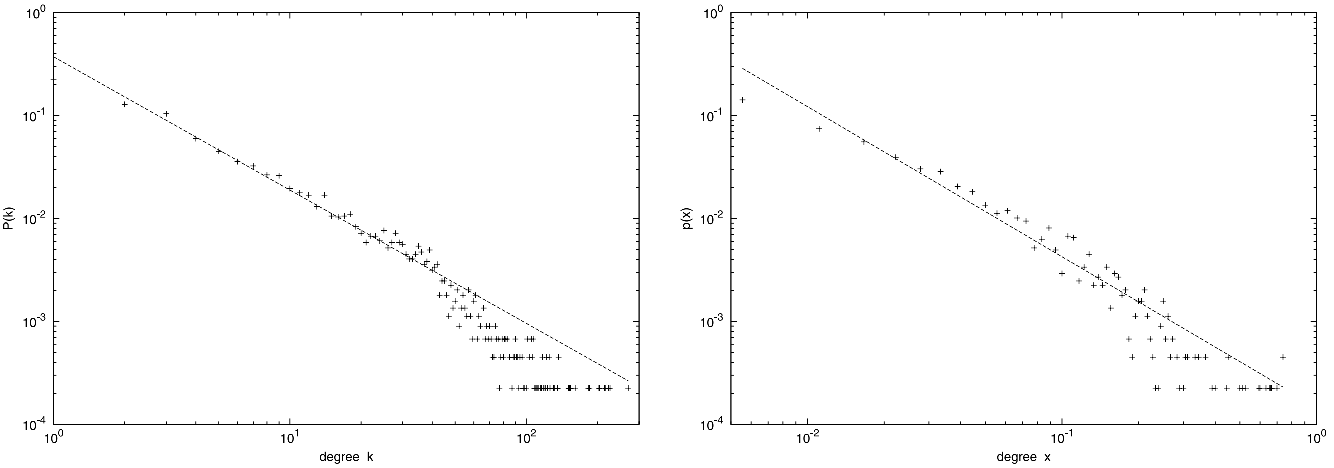


Fig. 4 Overview of a constructed entity network.

the likelihood of an *who*-entity in each test document by $P(e|d) = \sum_t P(e|t)P(t|d)$, where $P(t|d) = \sum_w P(t|w)P(w|d) + \sum_o P(t|o)P(o|d)$, instead of resampling from the test document in Sect. 4.2.1. In the equation above, w and o indicate a word and a *where*-entity, respectively. In this manner we can predict *who*-entities incrementally for a given document. The results using the GESwitchLDA are shown in Table 5. The results show that the model can predict *who*-entities even for incoming streams of documents, keeping fairly good prediction performance. Furthermore, we also applied some heuristics for name identification at preprocessing stage, such as, when only the first name of a person appears in a document, replacing it with his/her full name found by searching backward in the document. The results of the GESwitchLDA are shown in Table 6, where the performance was improved by applying the name identification processing.



(a) Discrete degree distribution

(b) Continuous degree distribution

Fig. 5 Degree distributions of an entity network with 7,300 vertices. (a) Discrete degree distribution with the average degree $\langle k\rangle = 12.24$ and the dashed line with slope $\xi = -1.295$. (b) Continuous degree distribution with $\langle x\rangle = 0.02525$ and $\xi = -1.458$.

4.3 Entity Link Prediction

4.3.1 Network Analysis

We first computed affinity of a pair of *who*-entities e_i and e_j by either of the following measures:

$$\textit{affinity1}: \; P(e_i|e_j)/2 + P(e_j|e_i)/2$$
$$\textit{affinity2}: \; P(e_i|e_j)P(e_j|e_i)$$

where $P(e_i|e_j) = \sum_t P(e_i|t)P(t|e_j)$ is estimated during training over the TDT2 collection using the GESwitchLDA model in the same manner in Sect. 4.2.1. We then listed entity pairs in order of the affinity. The *affinity1* was used in [14], and the *affinity2* indicates joint probability of $P(e_i|e_j)$ and $P(e_j|e_i)$ assuming that these are independent of each other.

Figure 4 shows an overview of an entity network constructed from the TDT2 collection, on the basis of the *affinity1* of *who*-entities that was mentioned above. In the entity network, each vertex represents a *who*-entity and each edge length represents strength of affinity between a pair of entities at the incident vertices. We then analyze the properties of such networks. For this analysis, we use the *affinity1* as an inter-entity affinity measure, but the *affinity2* can be used alternatively. We counted how many vertices there are in the entity network for each degree when a (discrete) degree k is defined as the number of the edges that are connected to a vertex, supposing every edge is assigned equal weight one, under the condition that the corresponding inter-entity affinity $P(e_i|e_j)/2 + P(e_j|e_i)/2 \geq 0.001$. The resulting degree distribution $P(k)$ is shown in Fig. 5 (a). We also computed degree distribution in another way, keeping edge weights that were obtained by the affinity of entities, and supposing that a (continuous) degree x is defined as the sum of the weights of the connected edges to a vertex, that is, the degree of entity e_i is obtained by $x(e_i) = \sum_j P(e_i|e_j)/2 + P(e_j|e_i)/2$.

Table 7 Results of *who*-entity link prediction with name identification.

affinity metric	model	MAP	accuracy
affinity1	LDA (T=100)	0.6062	0.5394
	SwitchLDA (T=100)	0.6235	0.5552
	GESwitchLDA (T=100)	0.6258	0.5564
affinity2	LDA (T=100)	0.6083	0.5401
	SwitchLDA (T=100)	0.6310	0.5588
	GESwitchLDA (T=100)	0.6328	0.5587

In order to draw a density curve of the continuous degree distribution $p(x)$, we set the number of classes to 200 and the class interval as $\Delta = \max_i x(e_i)/200$. As in the discrete degree distribution, we ignored the cases when inter-entity affinity $P(e_i|e_j)/2 + P(e_j|e_i)/2$ was less than 0.001. The resulting degree distribution is shown in Fig. 5 (b). We can observe that each of the degree distributions plotted in Fig. 5 conforms quite well to a power-law curve (i.e., straight line on a double logarithmic scale). Therefore, it can be said that the scale-free property [13] that are often seen in real-world complex networks like social networks can be observed even from the weighted relationship between *who*-entities extracted from written mentions.

4.3.2 Link Prediction

We further carried out experiments in order to investigate the predictive power of our GESwitchLDA model for *unknown* entity links, comparing with the LDA and the SwitchLDA models. Following [1], we generated two sets of *who*-entity pairs: (1) the true pairs that contain pairs that were never seen in any training document but were seen in test documents; and (2) false pairs that contain pairs that were never seen in any training or test document. The number of true pairs N_t and false pairs N_f were 104,721 and 98,977, respectively. We computed the inter-entity affinity using either the *affinity1* or the *affinity2*, as defined in Sect. 4.3.1, over all the true pairs and false pairs, and listed the entity

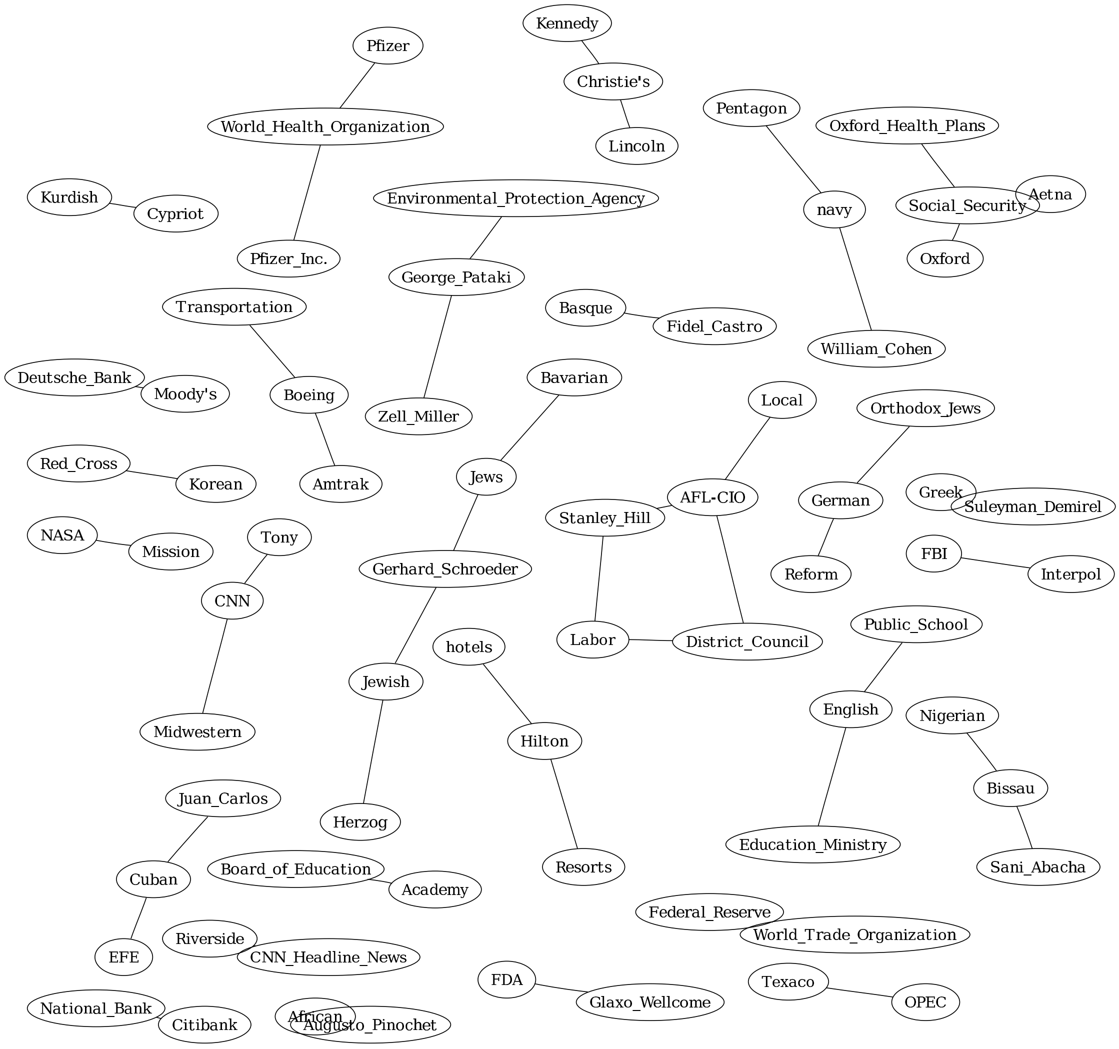


Fig. 6 Examples of predicted entity networks.

pairs in order of the inter-entity affinity. The evaluation results can be seen in Table 7. We used a couple of evaluation metrics: mean average precision (MAP) at the list of entity pairs in order of the inter-entity affinity, and accuracy at the top-ranked N_t predicted result. Our GESwitchLDA modestly outperformed the other two models: the LDA and the SwitchLDA, in terms of both MAP and accuracy. The *affinity2* works slightly better than the *affinity1*. The maximum improvement given by GESwitchLDA was 4.03% over LDA in terms of MAP in the case using the *affinity2*. Some examples of the predicted entity networks are shown in Fig. 6, where each vertex represents a *who*-entity and each edge length represents strength of affinity between a pair of entities at the incident vertices.

Although the networks of *who*-entities were discussed above, more specific social networks (i.e., person-entity networks) or mixed networks of *who*-entities and *where*-entities can also be predicted in the same manner.

5. Conclusions

We developed a multitype topic model, GESwitchLDA, by generalizing for an arbitrary number of word types such as words, *who*-entities (i.e., persons, organizations, or nationalities) and *where*-entities (i.e., locations, geographical/social/political entities, or facilities), in order to enable to capture dependencies between them. We compared this model with two other models on *who*-entity prediction task and entity link prediction task, using real data of news articles. We showed that the GESwitchLDA achieved significant improvement over the previous models in terms of some measures that are well-accepted in information retrieval research area, by distinguishing multiple types of entities: in this case, *who* and *where*.

Using this multitype topic model, entity networks can be effectively constructed from textual information. The entity networks are similar to social networks, where each vertex represents a person name; however, in the entity networks, not only person names but also organization names or *where*-entities can be involved, if necessary. Moreover, the social networks are usually constructed from explicit links between persons, such as from collaborations of film actors, from coauthorships, or via a social networking service [14]. On the other hand, our entity networks are extracted from *written mentions* and each edge is assigned a weight that represents inter-entity affinity computed via topic modeling. Even in the weighted networks of entities, we demonstrated the scale-free property that is often seen in social networks.

The multitype topic model can also be applied to other multiple types of words. For example, this model can be applied to documents that are manually or socially tagged, such as in Wikipedia. This model can also be applied to capture multiple types of entities in bio-medical articles, such as protein names, gene names and chemical compound names, even if more than two entity types are involved. In another direction of future work, we plan to extend the model to incorporate a temporal aspect of events. For entity network analysis, applying other distributional similarity metrics are left for the future work.

Acknowledgements

We thank Giridhar Kumaran, University of Massachusetts Amherst, for providing the annotation data. We also thank Atsuhiro Takasu of the National Institute of Informatics and the anonymous reviewers for their helpful comments. This work was supported in part by the Grant-in-Aid for Scientific Research on Priority Areas "Info-plosion" (#19024055), Scientific Research (B) (#20300038) and Exploratory Research (#18650057) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

[1] D. Newman, C. Chemudugunta, P. Smyth, and M. Steyvers, "Statistical entity-topic models," Proc. 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.680–686, ACM Press, New York, NY, USA, 2006.

[2] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent dirichlet allocation," J. Machine Learning Research, vol.3, pp.993–1022, 2003.

[3] T. Hofmann, "Probabilistic latent semantic indexing," Proc. 22nd International Conference on Research and Development in Information Retrieval, pp.50–57, Berkeley, California, USA, 1999.

[4] N. Ueda and K. Saito, "Parametric mixture models for multi-labeled text," Advances in Neural Information Processing Systems, vol.15, pp.721–728, MIT Press, Cambridge, Massachusetts, USA, 2003.

[5] T.L. Griffiths and M. Steyvers, "Finding scientific topics," Proc. National Academy of Sciences of the United States of America, vol.101, pp.5228–5235, 2004.

[6] M. Steyvers and T. Griffiths, Handbook of Latent Semantic Analysis, chapter 21, Lawrence Erbaum Associates, Mahwah, New Jersey, London, 2007.

[7] Y.W. Teh, D. Newman, and M. Welling, "A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation," Advances in Neural Information Processing Systems, vol.19, pp.1353–1360, MIT Press, Cambridge, Massachusetts, USA, 2006.

[8] J. Allan, Topic Detection and Tracking : Event-based Information Organization, chapter 1, Kluwer Academic Publishers, 2002.

[9] D.M. Bikel, R.L. Schwartz, and R.M. Weischedel, "An algorithm that learns what's in a name," Mach. Learn., vol.34, pp.211–231, 1999.

[10] J.P. Callan, W.B. Croft, and S.M. Harding, "The INQUERY retrieval system," Proc. 3rd International Conference on Database and Expert Systems Applications, pp.78–83, Valencia, Spain, 1992.

[11] R. Baeza-Yates and B. Ribeiro-Neto, eds., Modern Information Retrieval, chapter 3, pp.73–97, Addison-Wesley, 1999.

[12] S. Robertson, "On gmap: And other transformations," CIKM '06: Proc. 15th ACM International Conference on Information and Knowledge Management, pp.78–83, ACS, New York, NY, USA, 2006.

[13] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," Science, vol.286, no.5439, pp.509–512, 1999.

[14] M.E.J. Newman, "The structure and function of complex networks," SIAM Review, vol.45, no.2, pp.167–256, 2003.

Appendix

The following equations are used for Gibbs Sampling to estimate the GESwitchLDA model [1], [5].

$$P(z_i = t|w_i = v, x = y, z_{-i}, x_{-i}, w_{-i}, \alpha, \beta, \gamma) \propto$$

$$\frac{C^{TD}_{td,-i} + \alpha}{\sum_t C^{TD}_{td,-i} + T\alpha} \frac{n^{y}_{t,-i} + \gamma}{n^{all}_{t,-i} + M\gamma} \frac{C^{W_yT}_{w_yt,-i} + \beta^y}{\sum_w C^{W_yT}_{w_yt,-i} + W\beta^y}$$

$$\text{where} \quad n^y_t = \sum_{w_y} C^{W_yT}_{w_yt}, \quad n^{all}_t = \sum_y n^y_t.$$

In the equations, α and β are Dirichlet priors, and γ is another Dirichlet prior. β^y corresponds to Dirichlet prior for type-y words. T, D and W_y indicate the number of topics, the number of documents, and the number of vocabulary words of a specific word type y in the entire document collection, respectively. $C^{TD}_{td,-i}$ indicates a count that a topic t is assigned to a document d, but not including the current assignment of z_i. Similarly, $C^{W_yT}_{w_yt,-i}$ indicates a count that a type-y word w_y is assigned to a topic t, but not including the current assignment of z_i.

Hitohiro Shiozaki received the B.E. and M.E. degrees in Computer Science and Systems Engineering from Kobe University, Japan in 2006 and 2008, respectively. He is currently with Kawasaki Heavy Industries, Ltd.

Koji Eguchi is an Associate Professor at the Department of Computer Science and Systems Engineering, Kobe University, Japan and a Visiting Associate Professor at National Institute of Informatics (NII), Japan. His research interests include information retrieval, web computing and data mining.

Takenao Ohkawa is a Professor at the Department of Computer Science and Systems Engineering, Kobe University. His research interests include intelligent software and bioinformatics.