Sequential Entity Group Topic Model for Getting Topic Flows of Entity Groups within One Document

Young-Seob Jeong and Ho-Jin Choi

Department of Computer Science, KAIST
373-1 Guseong-dong, Yuseong-gu, Daejeon 305-701, Korea (South)
{pinode,hojinc}@kaist.ac.kr

Abstract. Topic mining is regarded as a powerful method to analyze documents, and topic models are used to annotate relationships or to get a topic flow. The research aim in this paper is to get topic flows of entities and entity groups within one document. We propose two topic models: Entity Group Topic Model (EGTM) and Sequential Entity Group Topic Model (S-EGTM). These models provide two contributions. First, topic distributions of entities and entity groups can be analyzed. Second, the topic flow of each entity or each entity group can be captured, through segments in one document. We develop collapsed gibbs sampling methods for performing approximate inference of the models. By experiments, we demonstrate the models by showing the analysis of topics, prediction performance, and the topic flows over segments in one document.

Keywords: Sequential topic model, Poisson-Dirichlet process, entity group.

1 Introduction

Analyzing documents on the Web is difficult due to the fast growing number of documents. Most of documents are not annotated, leading us to prefer unsupervised methods for analyzing document, and topic mining is one such method. This method is basically a probabilistic way to capture latent semantics, or topics, among documents. Since techniques like Probabilistic Latent Semantic Indexing (PLSI) [1] and Latent Dirichlet Allocation (LDA) [2] were first introduced, many studies have been derived from them: for example, to get relationships among entities in corpora [3, 4], to discover topic flows of documents in time dimension [5], or topic flows of segments in one document [6, 7], and so on. Capturing topic flows in one document (i.e., a fiction or a history) has special characteristics. For instance, adjacent segments in one document would influence each other because the full set of segments (i.e., the document) as a whole has some story. Moreover, the readers probably want to see the story in a perspective of each entity or each relationship. Although existing topic models tried to get topics of entity groups, no model has been proposed to obtain the topic flow of each entity or each relationship in one document. The topic flow in one document should also be useful for the readers to grasp the story easily.

P.-N. Tan et al. (Eds.): PAKDD 2012, Part I, LNAI 7301, pp. 366–378, 2012.

In this paper, we propose two topic models, Entity Group Topic Model (EGTM) and Sequential Entity Group Topic Model (S-EGTM), claiming two contributions. First, topic distribution of each entity and of each entity group can be analyzed. Second, the topic flow of each entity and each relationship through segments in one document can be captured. To realize our proposal, we adopt collapsed gibbs sampling methods [8] to infer the parameters of the models.

The rest of the paper is organized as follows. In the following subsection, we preview the terminology to set out the basic concepts. Section 2 discusses related works. Section 3 describes out approach and algorithms in detail. Section 4 presents experiments and results. Finally, Section 5 concludes.

1.1 Terminology

In this subsection, we summarize the terminology used in this paper to clarify the basic concepts.

- **Entity:** Something which the user want to get information about it. It can be a name, an object, or even a concept such as love and pain.
- **Empty group (empty set):** A group having no entity.
- **Entity group:** A group having one or more entities.
- **Entity group size:** The number of entities in the entity group.
- **Entity pair:** A pair of two entities.
- **Topic (word topic):** A multinomial word distribution.
- **Entity topic:** A multinomial entity distribution of CorrLDA2.
- **Segment:** A part of a document. It can be a paragraph, or even a sentence.
- **Topic flow:** A sequence of topic distribution through segments of a document.
- **Relationship of entities:** A topic distribution of the entity group.

2 Related Work

In this section, we describe related studies with respect to *entity topic mining* and *sequential topic mining*.

The goal of *entity topic mining* is to capture the topic of each entity, or of each relationship of entities. Author Topic Model (ATM) [9] is a model for getting a topic distribution of each author. Although the model does not consider entities, it can be used for getting topics of entities by just considering an entity as an author. However, it does not involve a process of writing entities in the document. There are several studies about a model involving the process. The recent proposed model, named as Nubbi [4], tried to capture two kinds of topics, which are the word distributions of each entity and of each entity pair. However, since it takes two kinds of topics separately, the topics of entities will be different from that of entity pairs. Several topic models for analyzing entities were introduced in [3]. Especially, CorrLDA2 showed its best prediction performance. The model captures not only topics, but also entity topics. The entity topic is basically a list of entities, thus each entity topic plays a role as an entity group. This implies that it has a lack of capability of getting relationship of a certain entity group.

As for *sequential topic mining*, there are works which tried to get topic flows in different dimensions. Dynamic Topic Model (DTM) [5] aimed to capture topic flows of documents in time dimension. Probabilistic way to capture the topic patterns on weblogs, in both of space dimension and time dimension, was introduced in [10]. Multi-grain LDA (MG-LDA) [11] used topic distribution of each window in a document to get the ratable aspects. Although it utilizes sequent topic distributions to deal with multi-grained topics, the objective of the model is not getting a topic flow of the document. STM and Sequential LDA tried to get a topic flow within a document. The both studies are based on a nested extension of the two-parameter Poisson-Dirichlet Process (PDP). The STM assumes that each segment is influenced by the document, while the Sequential LDA assumes that each segment is influenced by its previous segment except for the first segment.

3 Sequential Entity Group Topic Model

Existing works on *entity topic mining* and *sequential topic mining*, however, cannot be used to obtain topic flow of each entity and each relationship within one document. The topic flow of each entity or each relationship should also be useful for the readers to grasp the story more easily. This section introduces two topic models, Entity Group Topic Model (EGTM) and Sequential Entity Group Topic Model (S-EGTM).

3.1 Entity Group Topic Model

A graphical model of EGTM is shown in Figure 1(a). The meaning of notations is described in Table 1. We suggest an assumption that a *relationship* of entities must influence the topic distribution of every corresponding *entity* and *entity group*. To apply the assumption into our model, we employ a power-set. For example, if an entity group, having two entities *A* and *B*, have a relationship, then the relationship influences the topics of its power set such as entity *A*, entity *B*, and *empty set*(*empty group*). Thus, a topic distribution of the empty set will be very similar to that of the document, because it associates with every sentence. Formally, the generative process is represented in Figure 2.

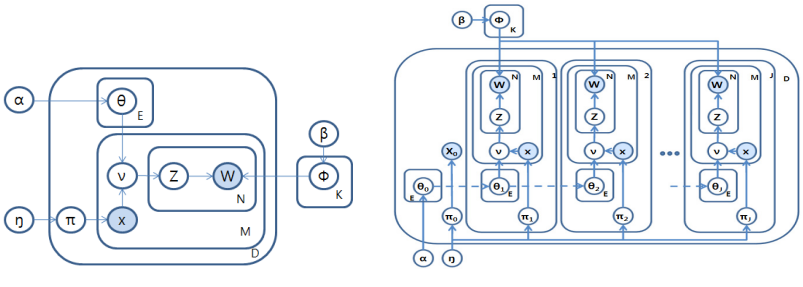


Fig. 1. (a) Graphical model of EGTM. The colored circles represent the variables observable from the documents. (b) Graphical model of S-EGTM.

1. Draw a word distribution Φ from Dirichlet(β)
2. For each document d,
   (1) For each entity group e,
   draw a topic distribution θ_{de} from Dirichlet(α
   (2) Draw an entity group dominance distribution π_d from Dirichlet(η)
   (3) For each sentence s,
      a. Choose an entity group x_{ds} from Multinomial(π_d)
      b. Given entity group x_{ds}, derive v_{ds} by multiplying θ_{de} which are members of a power-set of the x_{ds}
      c. For each word w,
         (a) Choose a topic z from Multinomial(v_{ds})
         (b) Given the topic z, generate a word w from Multinomial(Φ_z)

Fig. 2. The formal generative process of EGTM

As a sentence has only one entity group or an entity, the size of power-set does not grow exponentially. If there is no observed entity in a sentence, then the sentence has an *empty group*. We developed a collapsed gibbs sampling. At each step of the Markov chain, the topic of the ith word is chosen using a conditional probability

$$P(z_i = k \mid \mathbf{z}', \mathbf{w}, \mathbf{x}) \propto \frac{\beta_w + C_{kw}^{TW}}{\sum_v (\beta_v + C_{kv}^{TW})} \prod_{e \sqsubset P(d_s \mathbf{x})} \frac{\alpha_k + C_{dek}^{DET}}{\sum_z (\alpha_z + C_{dez}^{DET})}. \tag{1}$$

The notations are described in Table 1, with a minor exceptional use of notation that C_{kw}^{TW} and C_{dek}^{DET} in this expression exclude the ith word. Three parameters are obtained as follows:

$$\Phi_{kw} = \frac{\beta_w + C_{kw}^{TW}}{\sum_v (\beta_v + C_{kv}^{TW})}, \tag{2}$$

$$\theta_{dek} = \frac{\alpha_k + C_{dek}^{DET}}{\sum_z (\alpha_z + C_{dez}^{DET})}, \tag{3}$$

$$\pi_{de} = \frac{\eta_e + C_{de}^{DE}}{\sum_e (C_{de}^{DE} + \eta_e)}. \tag{4}$$

3.2 Sequential Entity Group Topic Model

A graphical model of S-EGTM is Figure 1(b). Formally, the generative process is represented in Figure 3. As S-EGTM gets a topic flow in a document, the D must be 1. A topic distribution of each segment is affected by that of previous segment, except that the first segment is affected by the document's topic distribution. To model this, we adopted Poisson-Dirichlet Process (PDP), as [7] does. If we use Chinese Restaurant Process (CRP) notations, then a word is a customer. The topics are dishes

Table 1. Meaning of the notations. The upper part contains variables for graphical models. The bottom part contains variables for representing the conditional probabilities.

Notations	Meaning of the notation
D	the number of documents
M	the number of sentences
N	the number of words
J	the number of segments
E	the number of unique entity groups
K	the number of topics
w	observed word
z	topic
ν	multiplying of multiple θ
x	observed entity group
θ	multinomial distribution over topics
Φ	multinomial distribution over words
π	multinomial distribution over entity groups
α	Dirichlet prior vector for θ
β	Dirichlet prior vector for Φ
ŋ	Dirichlet prior vector for π
a	a discount parameter for PDP
b	a strength parameter for PDP
z_i	the topic of *i*th word
′ (quote)	the situation that *i*th word is excepted
z	the topic assignments for all words
e	an entity group
w	a sequence of words of the document
t	in document d, the sequence of vectors which have table counts for each topic
$\mathbf{T}_{dje}$	in segment *j* of document *d*, the number of tables associated with entity group *e*
$\mathbf{N}_{dje}$	in segment *j* of document *d*, the number of words associated with entity group *e*
t_{djez}	in segment *j* of document *d*, the number of tables of entity group *e*, which are assigned the topic *z*
n_{djez}	in segment *j* of document *d*, the number of words of entity group *e*, which are assigned the topic *z*
C_{kw}^{TW}	the number of words that are assigned the topic *k*
C_{dek}^{DET}	in the document *d*, the number of topics that appear in the sentence which the entity group *e* associates
C_{de}^{DE}	a frequency of the entity group *e* in the document *d*
$P(^{ds}\boldsymbol{x})$	the power-set of entity group of the sentence *s* in the document *d*
$S_{T,a}^{N}$	the generalized Stirling number. Intuitively, this is the number of cases that N customers seat on T tables in different sequence
$(b \mid a)_C$	the Pochhammer symbol with increment C
$(b)_C$	The Pochhammer symbol same as $(b \mid 1)_C$
u_{dek}	An index of the first segment which has t_{duek}=0

and the segments are restaurants. The table count t is the number of tables occupied by customers. The customers sitting around a table share a dish. Especially, in nested PDP, the number of tables of next restaurant is a customer of current restaurant.

1. Draw a word distribution Φ from Dirichlet(β)
2. For each document d,
   (1) For each entity group e, draw a topic distribution θ_{d0e} from Dirichlet(α)
   (2) Draw an entity group dominance distribution π_{d0} from Dirichlet(η)
   (3) Choose an entity group x_{d0} from Multinomial(π_{d0})
   (4) For each segment j,
      a. Draw an entity group dominance distribution π_{dj} from Dirichlet(η)
      b. For each e, draw a topic distribution θ_{dje} from PDP(a, b, $\theta_{d(j-1)e}$)
      c. For each sentence s,
         (a) Choose an entity group x_{djs} from Multinomial(π_{dj})
         (b) Given entity group x_{djs}, derive v_{djs} by multiplying θ_{dje} which are members of a power-set of the x_{djs}
         (c) For each word w,
            i. Choose a topic z from Multinomial(v_{djs})
            ii. Given the topic z, generate a word w from Multinomial(Φ_z)

Fig. 3. The formal generative process of S-EGTM

When we do a collapsed gibbs sampling for topics, removing ith topic z_{dgi}=k affects the table counts and topic distributions of entity group e in the segment g. Therefore, we need to consider three cases of conditional probabilities in terms of u_{dek}, as following.

First, when u_{dek}=1,

$$P(z_{dgi}=k \mid \mathbf{z}', \mathbf{w}, \mathbf{x}, \mathbf{t}) \propto \frac{\alpha_k + t'_{d1ek}}{\sum_z (\alpha_z + t'_{d1ez})} (b + aT'_{d1e}) \left(\prod_{j=2}^{g} \frac{b + aT'_{dje}}{b + N_{d(j-1)e} + T'_{dje}} \right) \frac{\beta_w + C_{kw}^{TW}}{\sum_v (\beta_v + C_{kv}^{TW})} \cdot \tag{5}$$

Second, when $1 < u_{dek} \leq g$,

$$P(z_{dgi}=k \mid \mathbf{z}', \mathbf{w}, \mathbf{x}, \mathbf{t}) \propto \left(\prod_{j=u_{dek}}^{g} \frac{b + aT'_{dje}}{b + N_{d(j-1)e} + T'_{dje}} \right) \left(\frac{S^{n_{d(u_{dek}-1)ek}+1}_{t_{d(u_{dek}-1)ek},a}}{S^{n_{d(u_{dek}-1)ek}}_{t_{d(u_{dek}-1)ek},a}} \right) \frac{\beta_w + C_{kw}^{TW}}{\sum_v (\beta_v + C_{kv}^{TW})} \cdot \tag{6}$$

Third, when $g < u_{dek}$,

$$P(z_{dgn}=k \mid \mathbf{z}', \mathbf{w}, \mathbf{x}, \mathbf{t}) \propto \left(\frac{S^{n'_{dgek}+1+t_{d(g+1)ek}}_{t'_{dgek},a}}{S^{n'_{dgek}+t_{d(g+1)ek}}_{t'_{dgek},a}} \right) \frac{\beta_w + C_{kw}^{TW}}{\sum_v (\beta_v + C_{kv}^{TW})} \cdot \tag{7}$$

The notations are described in Table 1, with a minor exceptional use of notation that in this C_k^{TW} expression exclude the ith word. At each step, we also sample a table count because a table count is affected by the number of words having the table's topic and

vice versa. If we assume that we remove a table count t_{dgek}, then new table count is sampled as follows:

$$P(t_{dgek} \mid \mathbf{z}', \mathbf{w}, \mathbf{x}, \mathbf{t}') \propto \left(\frac{\Gamma(\alpha_k + t_{d1ek})}{\Gamma(\sum_z \alpha_z + t_{d1ez})}\right)^{g=1} \left(\frac{S^{n_{d(g-1)ek}+t_{dgek}}_{t_{d(g-1)ek},a}}{(b)_{N_{d(g-1)e}+T'_{dge}}}\right)^{1-(g=1)} \left((b \mid a)_{T'_{dge}} S^{n_{dgek}+t_{d(g+1)ek}}_{t_{dgek},a}\right) \quad (8)$$

The notation g=1 means the first term is active only if it is the first segment, 1-(g=1) means the second term is active only if it is not the first segment. Γ is a gamma function. As considering every candidates of table count is intractable, we have to determine the window size of table count to consider. Among four parameters, we describe the approximate probabilities of two parameters as follows:

$$\theta_{d0ek} = \frac{\alpha_k + t_{d1ek}}{\sum_z (\alpha_z + t_{d1ez})}, \quad (9)$$

$$\theta_{djek} = \frac{n_{djek} - a \cdot t_{djek} + t_{d(j+1)ek} + \theta_{d(j-1)ek}(a \cdot T_{dje} + b)}{b + N_{dje} + T_{d(j+1)e}}. \quad (10)$$

4 Experiments

We used two data sets: the Bible and the fiction 'Alice'. We removed stop-words and did stemming by Porter stemmer. The sentences were recognized by '.', '?', '!', and "newline". After the deleting stop-words, the Bible has 295,884 words and the fiction 'Alice' has 11,605 words. As S-EGTM gets a topic flow in a document, it regards the Bible as a document consisting of 66 segments, and the fiction 'Alice' as a document consisting of 12 segments. In contrast, to compare EGTM with other models, we divided each document into separated files as segments. For every experiment, we set α=0.1, β=0.01, η=1, a=0.5, b=10, and the window size was 1.

4.1 The Size of Power-Set of Entity Groups

When we input a list of entities to consider, then a preprocessing will make a power-set hierarchy of existing entity groups in a document. Since a sentence is restricted to have an entity group, each entity group usually does not have more than three entities. Thus, as shown in Figure 4, the size of power-set does not grow exponentially. The used data is the Bible.

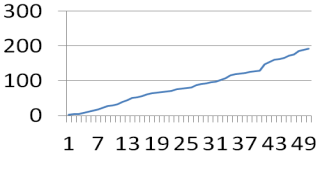


Fig. 4. The number of unique entity groups. The horizontal axis is the number of entities and the vertical axis represents the number of unique entity groups.

4.2 Topic Discovery

We used the Bible as a data and set the number of topics to be 20. We performed inference with 2,000 iterations. In Table 2 and Table 3, five topics of two models, LDA and EGTM, are shown. The obtained topics of the models are similar to each other because the empty group of EGTM associates with every sentence. The topics are coherent and specific to understand. EGTM additionally gives entity lists and relationship lists about each topic. With the lists, we can understand what are the topics that each entity or entity group associates with. For example, the topic *Mission work*, which is about missionary acts of apostles, is mostly handled in the *Act* written by *Paul* who lived in different era with *Abraham*. Nevertheless, the relationship {*God,Abraham*} has the topic *Mission work* the most, in the *Act*. This is caused by *Paul*'s writing about the covenant between *God* and *Abraham*. Since the covenant is that '*through your offspring all peoples on earth will be blessed*', the relationship {*God,Abraham*} has the topic *Mission work* in *Act*. Thus, EGTM helps us to grasp the documents in perspective of an entity or relationship.

Table 2. Topics obtained from LDA. The topic names are manually labeled. The listed chapters have a big proportion of the corresponding topic.

Topics	Gospel	Journey of Jesus & disciples	Mission work	Kingdom of Israel	Field life & Sanctuary
Top words	Christ	disciple	Jew	king	Egypt
	faith	father	Jerusalem	Israel	gold
	love	son	spirit	Judah	curtain
	sin	crowd	holy	son	Israelite
	law	reply	sail	temple	cubit
	spirit	ask	Antioch	reign	blue
	gospel	heaven	prison	Jerusalem	altar
	grace	truth	apostle	father	mountain
	church	answer	gentile	priest	ring
	truth	kingdom	Ship	prophet	acacia
	hope	Pharisee	Asia	Samaria	pole
	power	teacher	travel	altar	ephod
	dead	law	province	servant	tent
Chapters	Romans~ Jude	Matthew~ John	Acts	Kings	Exodus

4.3 Entity Prediction

We compared the EGTM with CorrLDA2 by entity prediction performance. We also made a model, named as *Entity-LDA*, which is a baseline. The Entity-LDA just counts the number of topics in sentences which have each entity, after LDA estimation. We used the Bible as data and varied the number of topics from 10 to 90. We used an entity list consisting of 16 entities: *God, Jesus, Petro, Judas, Paul, Mary, David, John, Abraham, Sarai, Solomon, Moses, Joshua, Aaron, Jeremiah,* and *Jonah*. For fair comparison, we made CorrLDA2 to use the entity list, rather than automatic Named Entity Recognition methods. For CorrLDA2, we set the number of entity topics same as the number of word topics, because we observed that the prediction results are similar with different numbers

of entity topics. We did 10-fold cross validation for the comparison, and got the prediction results using the process in Figure 5.

Table 3. Topics obtained from EGTM. The topic names are manually labeled.

Topics	Gospel	Journey of Jesus & disciples	Mission work	Kingdom of Israel	Field life & Sanctuary
Top words	Christ faith love law sin grace gospel world spirit hope church life boast	disciple father son crowd reply truth ask Pharisee kingdom teacher world heaven answer	Jew Jerusalem holy spirit sail ship gentile speak disciple believe Christ Antioch prison	king Israel Judah temple Jerusalem son reign Samaria prophet father priest altar servant	land Egypt curtain Israelite cubit gold mountain altar ring frame blue tent pole
Chapters	Romans ~ Jude	Matthew ~ John	Acts	Kings	Exodus
Entities	God, Jesus, Paul, John	God, Jesus, Mary, Judas, David, Abraham, Joshua, Moses	God, Jesus, Paul, Judas, John, David, Abraham, Moses	God, David, Abraham, Solomon, Moses	God, Abraham, Moses, Joshua, Aaron
Relation-ships	{God,Jesus}, {God,Paul}, {God,John}, {Jesus,Paul}	{God,Jesus}, {Abraham,Jesus}, {Jesus,David}, {Abraham,Joshua, David}	{God,Jesus}, {Paul,Jesus}, {Paul,Judas}, {John,Jesus}, {David,Judas}, {Paul,John}, {God,Abraham}	{God,David}, {Solomon, David}, {God, Solomon}, {David,God, Solomon}	{God,Abraham}, {God,Moses, Abraham}, {Aaron,God}, {Moses,Joshua}, {Moses,Abraham }

1. Train the Entity-LDA, EGTM and CorrLDA2 with same training data.
2. For each sentence of test data, three models are supposed to choose one of 16 entities using the following predictive distributions:

   Entity-LDA : $P(e|s) \propto \prod_{w} \sum_{t} P(w|t)P(t|e)$,

   CorrLDA2 : $P(e|s) \propto \prod_{w} \sum_{t} P(w|t)P(t|d)$ as in [3],

   EGTM : $P(e|s) \propto \prod_{w} \sum_{t} P(w|t)P(t|e)$

   where w is a word of a sentence s, d represents each training document, t is a topic, and e is the entity. Especially, $P(t|d)$ is obtained by resampling with $P(w|t)$.
3. The accuracy is the number of correct choices divided by the number of total choices.

Fig. 5. The process of the entity prediction

The test data consists of sentences which have at least one entity. If a sentence has multiple entities, then choosing one of them is regarded as a correct choice. As depicted in Figure 6(a), the CorrLDA2 shows fixed performance because the resampling makes *P(t|d)* to be fixed. EGTM outperforms other models because the topics of Entity-LDA have nothing to do with entities and CorrLDA2 does not get the topic distribution of each entity. The performance of EGTM grows as the number of topics grows because the Bible covers various topics. EGTM shows better performances than CorrLDA2 because of two reasons. First, CorrLDA2 does not directly get the topic distribution of each entity and it disperses the topic distribution of each entity into multiple entity topics. Second, CorrLDA2 takes data exclusively. To be specific, the data already used for entity topics will not be used for word topics.

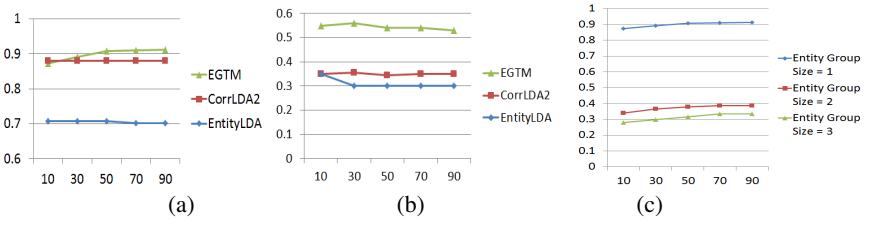


Fig. 6. (a) The entity prediction performances of three models. The horizontal axis is the number of topics. The vertical axis means a prediction rate. (b) The entity pair prediction performances. (c) The entity group prediction performances.

4.4 Entity Pair Prediction

We compared the entity pair prediction performance between EGTM and CorrLDA2. For fair comparison, we used *entity-entity affinity* of [3]. The entity-entity affinity, defined as $P(e_i|e_j)/2+P(e_j|e_i)/2$, is to rank *true pairs* and *false pairs*. The true pairs exist in only unseen document, while the false pairs do not exist. The prediction performance is the number of true pairs in half of high ranked pairs, divided by the number of total pairs. We prepared 50 true pairs and 50 false pairs. The models have different methods to get $P(e_i|e_j)$ which is obtainable from $\sum_t P(e_i|t)P(t|e_j)$. Entity-LDA just counts the number of each topic. CorrLDA2 uses entity topic distributions. For examp$P(t|e_j) = \sum_{et} P(t|et)P(et|e_j)$where *et* means each entity topic. Figure 6(b) describes the prediction performance. Because the most entities of the Bible old testament usually do not appear in the Bible new testament, the overall prediction performances is low. EGTM outperforms CorrLDA2 and Entity-LDA, because EGTM directly takes a topic distribution of each entity.

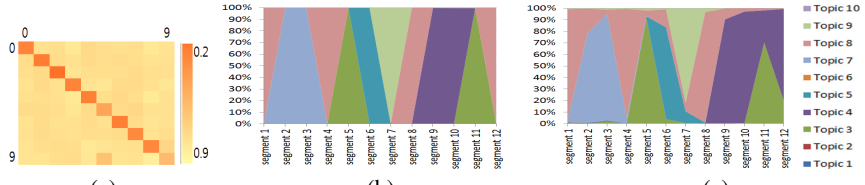
4.5 Entity Group Prediction

We do not compare the prediction performance with other models because the other models lack ability to get topic distributions of entity groups. Instead, we demonstrate

prediction performance with different entity group sizes. The predictive distribution is $P(eg|s) \propto \sum_d P(eg_d) \prod_w \sum_t P(w|t)P(t|eg_d)$, where eg represents the entity group, and d represents each training document. Figure 6(c) shows the prediction performance. The accuracy is the number of correct predictions divided by the number of total predictions. The prediction performance of smaller entity group is better than that of larger entity group, because it is harder to predict more entities.

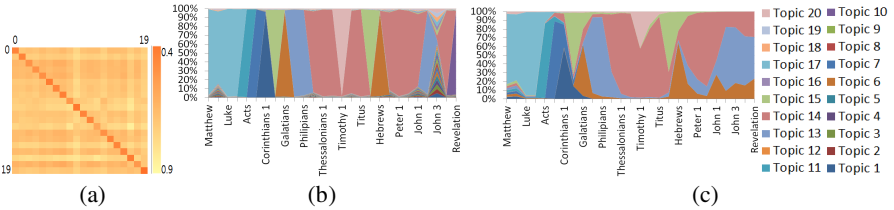
4.6 Topic Flow

We compare the topic flow of S-EGTM with the topic distributions of EGTM. To show the topic consistency between the two models, we trained S-EGTM boosted from the trained EGTM with 2,000 iterations. The Bible new testament and the fiction 'Alice' are used as data. We analyze the entity *Alice* with 10 topics, and analyze a relationship {*Jesus*, *God*} with 20 topics. Figure 7 and Figure 8 show the topic flows of the entity *Alice* and the relationship {*Jesus*, *God*}, respectively.



(a) (b) (c)

Fig. 7. (a) The confusion matrix by Hellinger distance, with the fiction 'Alice' as a data, where S-EGTM topics run along the Y-axis. (b) Topic flow of entity *Alice* by EGTM. (c) Topic flow of entity *Alice* by S-EGTM.



(a) (b) (c)

Fig. 8. (a) The confusion matrix by Hellinger distance, with the Bible new testament as a data, where S-EGTM topics run along the Y-axis. (b) Topic flow of relationship {*Jesus*, *Paul*} by EGTM. (c) Topic flow of relationship {*Jesus*, *Paul*} by S-EGTM.

Figure 7(a) and Figure 8(a) show the confusion matrices of the topic distributions generated by EGTM and S-EGTM. The diagonal cells are darker than others, meaning that the corresponding topics have low Hellinger distance. Thus, the topics of two models are consistent. Other than the Figure 7(a) and Figure 8(a), the horizontal axis means each segment, while the vertical axis represents topic proportion. Clearly, in Figure 7(b), each topic appears in totally different segments, which gives no idea about a topic flow through the segments. In contrast, in Figure

7(c), we can see the pattern that the topic 8(pink color) flows through every segment. As the topic 8 is about *Alice's tracking the rabbit*, its flow through every segment is coherent with the story. Consider the case of the relationship {*Jesus*, *God*} in more detail. In Figure 8(b), the topic *Gospel* (topic 14) is dominant in four separated parts, meaning that the relationship {*Jesus*, *God*} associates with the topic *Gospel* in only those separated four parts. This is caused by that the relationship has sparse topic distribution because it reflects only the sentences having the relationship. The separated appearance of the topic is not coherent with the Bible, because a purpose of the Bible new testament associates with the topic *Gospel* which is strongly about the news of the relationship {*Jesus*, *God*}. In contrast, in Figure 8(c), the topic *Gospel* appears like a flow from *Acts* to *Revelation*. This means the relationship {*Jesus*, *God*} associates with the topic *Gospel* without any cutting, through the segments. This is more coherent with the Bible. Thus, S-EGTM helps us to grasp the topic flow of an entity or a relationship by smoothing the sparse topic distribution of EGTM.

5 Conclusion

In this paper, we proposed two new generative models, Entity Group Topic Model (EGTM) and the Sequential Entity Group Topic Model (S-EGTM). S-EGTM reflects the sequential structure of a document in the hierarchical modeling. We developed collapsed gibbs sampling algorithms for the models. EGTM employs a power-set structure to get topics of entities or entity groups. S-EGTM is a sequential version of the EGTM, and employs nested two-parameter Poisson-Dirichlet process (PDP) to capture a topic flow over the sequence of segments in one document. We have analyzed the topics obtained from EGTM, and showed that topic flows generated by S-EGTM are coherent with the original document. Moreover, the experimental results show that the prediction performance of EGTM is better than that of CorrLDA2. Thus, we believed that the intended mechanisms of the EGTM and S-EGTM models work.

Acknowledgments. This work was supported by the National Research Foundation (NRF) grant (No. 2011-0018264) of Ministry of Education, Science and Technology (MEST) of Korea.

References

1. Hofmann, T.: Probabilistic Latent Semantic Indexing. In: SIGIR, pp. 50–57 (1999)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. In: NIPS, pp. 601–608 (2001)
3. Newman, D., Chemudugunta, C., Smyth, P.: Statistical entity-topic models. In: KDD, pp. 680–686 (2006)
4. Chang, J., Boyd-Graber, J.L., Blei, D.M.: Connections between the lines: augmenting social networks with text. In: KDD, pp. 169–178 (2009)
5. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: ICML, pp. 113–120 (2006)

6. Du, L., Buntine, W.L., Jin, H.: A segmented topic model based on the two-parameter Poisson-Dirichlet process. Machine Learning, 5–19 (2010)
7. Du, L., Buntine, W.L., Jin, H.: Sequential Latent Dirichlet Allocation: Discover Underlying Topic Structures within a Document. In: ICDM, pp. 148–157 (2010)
8. Griffiths, T.L., Steyvers, M.: Finding Scientific Topics. National Academy of Sciences, 5228–5235 (2004)
9. Rosen-Zvi, M., Griffiths, T.L., Steyvers, M., Smyth, P.: The Author-Topic Model for Authors and Documents. In: UAI, pp. 487–494 (2004)
10. Mei, Q., Liu, C., Su, H., Zhai, C.: A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In: WWW, pp. 533–542 (2006)
11. Titov, I., McDonald, R.T.: Modeling Online Reviews with Multi-grain Topic Models. CoRR (2008)